inf_phi2

March 1, 2024

[1]:
```
# Data Exploration Imports
import matplotlib.pyplot as plt
import seaborn as sns

import numpy as np
import pandas as pd
from tqdm import tqdm
import torch
import torch.nn as nn
import transformers
from peft import LoraConfig, PeftConfig
from transformers import (AutoModelForCausalLM,
                          AutoTokenizer,
                          BitsAndBytesConfig,
                          TrainingArguments,
                          pipeline,
                          logging)
from sklearn.metrics import (accuracy_score,
                             classification_report,
                             confusion_matrix)
```

[2]:
```
!nvidia-smi
```

```
Wed Feb 21 21:34:30 2024
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 441.37       Driver Version: 441.37       CUDA Version: 10.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name            TCC/WDDM | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  GeForce GTX 166… WDDM  | 00000000:01:00.0  On |                  N/A |
| N/A   69C    P8     7W /  N/A |    729MiB /  6144MiB |      2%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
```

```
|    0      4140    C+G   …osoft.LockApp_cw5n1h2txyewy\LockApp.exe N/A      |
|    0      5860    C+G   …4__8wekyb3d8bbwe\XboxGameBarWidgets.exe N/A      |
|    0      6260    C+G   …__8wekyb3d8bbwe\PhoneExperienceHost.exe N/A      |
|    0      6552    C+G   Insufficient Permissions                   N/A      |
|    0      7428    C+G   …es\Google\Chrome\Application\chrome.exe N/A      |
|    0      7524    C+G   Insufficient Permissions                   N/A      |
|    0      9000    C+G   …5n1h2txyewy\StartMenuExperienceHost.exe N/A      |
|    0     10004    C+G   C:\Windows\explorer.exe                    N/A      |
|    0     10008    C+G   …ent.CBS_cw5n1h2txyewy\TextInputHost.exe N/A      |
|    0     11060    C+G   …t_cw5n1h2txyewy\ShellExperienceHost.exe N/A      |
|    0     11224    C+G   …ation\121.0.2277.128\msedgewebview2.exe N/A      |
|    0     11628    C+G   …_x64__8wekyb3d8bbwe\Microsoft.Notes.exe N/A      |
|    0     12368    C+G   …Client.CBS_cw5n1h2txyewy\SearchHost.exe N/A      |
|    0     14332    C+G   …4__zpdnekdrzrea0\XboxGameBarSpotify.exe N/A      |
|    0     14672    C+G   …mmersiveControlPanel\SystemSettings.exe N/A      |
|    0     14776    C+G   …ChxApp_cw5n1h2txyewy\CHXSmartScreen.exe N/A      |
|    0     15776    C+G   …_x64__8wekyb3d8bbwe\WindowsTerminal.exe N/A      |
+-----------------------------------------------------------------------------+
```

[43]:
```
test_df = pd.read_csv('../data/test.csv').drop("Unnamed: 0", axis=1)
test_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7858 entries, 0 to 7857
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   question         7858 non-null   object
 1   context          7858 non-null   object
 2   answer           7858 non-null   object
 3   table_count      7858 non-null   int64
 4   sub_query_count  7858 non-null   int64
 5   joins_count      7858 non-null   int64
 6   where_count      7858 non-null   int64
 7   group_by_count   7858 non-null   int64
 8   columns_count    7858 non-null   int64
 9   complexity       7858 non-null   int64
 10  difficulty       7858 non-null   object
dtypes: int64(7), object(4)
memory usage: 675.4+ KB
```

[102]:
```
from sklearn.model_selection import train_test_split
df_a, df_b = train_test_split(test_df, test_size=0.2,␣
 ↪stratify=test_df["difficulty"])
df_a, df_c = train_test_split(df_a, test_size=0.2, stratify=df_a["difficulty"])
```

[103]:
```
df_a['difficulty'].value_counts(), df_b['difficulty'].value_counts(),␣
 ↪df_c['difficulty'].value_counts()
```

[103]:
```
(difficulty
 easy      3484
 medium    1519
 hard        25
 Name: count, dtype: int64,
 difficulty
 easy      1089
 medium     475
 hard         8
 Name: count, dtype: int64,
 difficulty
 easy      872
 medium    380
 hard        6
 Name: count, dtype: int64)
```

[6]:
```
def generate_test_prompt(data_point):
    return f"""### Task
Generate a SQL query to answer the following question:
`{data_point['question']}`

### Database Schema
The query will run on a database with the following schema:
{data_point['context']}

### Answer
Given the database schema, here is the SQL query that answers␣
 ↪`{data_point['question']}`:
```sql""".strip()
```

[9]:
```
from llama_cpp import Llama

phi2 = Llama(model_path="../phi2_sqlcoder_f16.gguf")
```

```
llama_model_loader: loaded meta data with 19 key-value pairs and 453 tensors
from ../phi2_sqlcoder_f16.gguf (version GGUF V3 (latest))
llama_model_loader: Dumping metadata keys/values. Note: KV overrides do not
apply in this output.
llama_model_loader: - kv   0:                       general.architecture str
= phi2
llama_model_loader: - kv   1:                               general.name str
= Phi2
llama_model_loader: - kv   2:                      phi2.context_length u32
= 2048
llama_model_loader: - kv   3:                    phi2.embedding_length u32
= 2560
llama_model_loader: - kv   4:                 phi2.feed_forward_length u32
= 10240
```
```

```
llama_model_loader: - kv   5:                          phi2.block_count u32
= 32
llama_model_loader: - kv   6:                 phi2.attention.head_count u32
= 32
llama_model_loader: - kv   7:              phi2.attention.head_count_kv u32
= 32
llama_model_loader: - kv   8:         phi2.attention.layer_norm_epsilon f32
= 0.000010
llama_model_loader: - kv   9:                 phi2.rope.dimension_count u32
= 32
llama_model_loader: - kv  10:                         general.file_type u32
= 1
llama_model_loader: - kv  11:              tokenizer.ggml.add_bos_token bool
= false
llama_model_loader: - kv  12:                       tokenizer.ggml.model str
= gpt2
llama_model_loader: - kv  13:                      tokenizer.ggml.tokens
arr[str,51200]   = ["!", "\"", "#", "$", "%", "&", "'", …
llama_model_loader: - kv  14:                  tokenizer.ggml.token_type
arr[i32,51200]   = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
llama_model_loader: - kv  15:                      tokenizer.ggml.merges
arr[str,50000]   = ["Ġ t", "Ġ a", "h e", "i n", "r e",…
llama_model_loader: - kv  16:                tokenizer.ggml.bos_token_id u32
= 50256
llama_model_loader: - kv  17:                tokenizer.ggml.eos_token_id u32
= 50256
llama_model_loader: - kv  18:            tokenizer.ggml.unknown_token_id u32
= 50256
llama_model_loader: - type  f32:  259 tensors
llama_model_loader: - type  f16:  194 tensors
llm_load_vocab: mismatch in special tokens definition ( 910/51200 vs 944/51200
).
llm_load_print_meta: format           = GGUF V3 (latest)
llm_load_print_meta: arch             = phi2
llm_load_print_meta: vocab type       = BPE
llm_load_print_meta: n_vocab          = 51200
llm_load_print_meta: n_merges         = 50000
llm_load_print_meta: n_ctx_train      = 2048
llm_load_print_meta: n_embd           = 2560
llm_load_print_meta: n_head           = 32
llm_load_print_meta: n_head_kv        = 32
llm_load_print_meta: n_layer          = 32
llm_load_print_meta: n_rot            = 32
llm_load_print_meta: n_embd_head_k    = 80
llm_load_print_meta: n_embd_head_v    = 80
llm_load_print_meta: n_gqa            = 1
llm_load_print_meta: n_embd_k_gqa     = 2560
llm_load_print_meta: n_embd_v_gqa     = 2560
```

```
llm_load_print_meta: f_norm_eps       = 1.0e-05
llm_load_print_meta: f_norm_rms_eps   = 0.0e+00
llm_load_print_meta: f_clamp_kqv      = 0.0e+00
llm_load_print_meta: f_max_alibi_bias = 0.0e+00
llm_load_print_meta: n_ff             = 10240
llm_load_print_meta: n_expert         = 0
llm_load_print_meta: n_expert_used    = 0
llm_load_print_meta: rope scaling     = linear
llm_load_print_meta: freq_base_train  = 10000.0
llm_load_print_meta: freq_scale_train = 1
llm_load_print_meta: n_yarn_orig_ctx  = 2048
llm_load_print_meta: rope_finetuned   = unknown
llm_load_print_meta: model type       = 3B
llm_load_print_meta: model ftype      = F16
llm_load_print_meta: model params     = 2.78 B
llm_load_print_meta: model size       = 5.18 GiB (16.01 BPW)
llm_load_print_meta: general.name     = Phi2
llm_load_print_meta: BOS token        = 50256 '<|endoftext|>'
llm_load_print_meta: EOS token        = 50256 '<|endoftext|>'
llm_load_print_meta: UNK token        = 50256 '<|endoftext|>'
llm_load_print_meta: LF token         = 30 '?'
llm_load_tensors: ggml ctx size =    0.17 MiB
llm_load_tensors:        CPU buffer size =  5303.65 MiB
…
…
llama_new_context_with_model: n_ctx      = 512
llama_new_context_with_model: freq_base  = 10000.0
llama_new_context_with_model: freq_scale = 1
llama_kv_cache_init:        CPU KV buffer size =   160.00 MiB
llama_new_context_with_model: KV self size  =  160.00 MiB, K (f16):   80.00 MiB,
V (f16):   80.00 MiB
llama_new_context_with_model:        CPU input buffer size   =     7.01 MiB
llama_new_context_with_model:        CPU compute buffer size =   105.00 MiB
llama_new_context_with_model: graph splits (measure): 1
AVX = 1 | AVX_VNNI = 0 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI =
0 | FMA = 1 | NEON = 0 | ARM_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 |
BLAS = 0 | SSE3 = 1 | SSSE3 = 0 | VSX = 0 | MATMUL_INT8 = 0 |
Model metadata: {'general.architecture': 'phi2', 'phi2.context_length': '2048',
'general.name': 'Phi2', 'phi2.attention.head_count_kv': '32',
'phi2.embedding_length': '2560', 'tokenizer.ggml.add_bos_token': 'false',
'phi2.feed_forward_length': '10240', 'tokenizer.ggml.bos_token_id': '50256',
'phi2.block_count': '32', 'phi2.attention.head_count': '32',
'phi2.attention.layer_norm_epsilon': '0.000010', 'phi2.rope.dimension_count':
'32', 'tokenizer.ggml.eos_token_id': '50256', 'general.file_type': '1',
'tokenizer.ggml.model': 'gpt2', 'tokenizer.ggml.unknown_token_id': '50256'}
```

[81]:
```
from datetime import datetime

def predict(df, llm, out):
    for point in tqdm(df.iloc, desc="No. of rows", total=df.shape[0]):
        start = datetime.now()
        prompt = generate_test_prompt(point)
        out['prompt'].append(prompt)
        out['actu'].append(point['answer'])
        result = llm(prompt=prompt,
            max_tokens = 50,
            temperature = 0.2,
            stop = ['```'])
        answer = result
        end = datetime.now()
        out['inf_time'].append((end - start).total_seconds())
        out['pred'].append(answer['choices'][0]['text'].strip())
        out['temperature'].append(0.2)
        out['difficulty'].append(point['difficulty'])
        out['token_in'].append(result['usage']['prompt_tokens'])
        out['token_out'].append(result['usage']['completion_tokens']+1)
        out['tokens_per_sec'].append(result['usage']['completion_tokens']/((end␣
 ↪- start).total_seconds()))
    return out
```

[104]:
```
out = {"prompt": [], "pred": [], "actu": [], "inf_time": [], "temperature": [],␣
 ↪"difficulty": [], "token_in": [], "token_out": [], "tokens_per_sec": []}

out = predict(df_c, phi2, out)
json.dump(out, open("phi2_eval_df_c.json", "w"))
```

```
No. of rows:   0%|          | 0/1258 [00:00<?, ?it/s]Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.65 ms /    21 runs   (    0.32
ms per token,  3156.95 tokens per second)
llama_print_timings: prompt eval time =    9239.41 ms /    97 tokens (   95.25
ms per token,    10.50 tokens per second)
llama_print_timings:        eval time =    4370.24 ms /    20 runs   (  218.51
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   13702.39 ms /   117 tokens
No. of rows:   0%| | 1/1258 [00:13<4:47:14, 13.71sLlama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       5.77 ms /    18 runs   (    0.32
ms per token,  3118.50 tokens per second)
llama_print_timings: prompt eval time =    7916.35 ms /    83 tokens (   95.38
```

```
ms per token,    10.48 tokens per second)
llama_print_timings:        eval time =    3641.99 ms /    17 runs   (  214.23
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   11640.30 ms /   100 tokens
No. of rows:   0%| | 2/1258 [00:25<4:21:34, 12.50sLlama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.61 ms /    33 runs   (    0.35
ms per token,  2842.62 tokens per second)
llama_print_timings: prompt eval time =   10414.86 ms /   112 tokens (   92.99
ms per token,    10.75 tokens per second)
llama_print_timings:        eval time =    7359.29 ms /    32 runs   (  229.98
ms per token,     4.35 tokens per second)
llama_print_timings:       total time =   17936.83 ms /   144 tokens
No. of rows:   0%| | 3/1258 [00:43<5:13:23, 14.98sLlama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.47 ms /    22 runs   (    0.34
ms per token,  2946.69 tokens per second)
llama_print_timings: prompt eval time =    8817.81 ms /    95 tokens (   92.82
ms per token,    10.77 tokens per second)
llama_print_timings:        eval time =    4522.27 ms /    21 runs   (  215.35
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   13448.17 ms /   116 tokens
No. of rows:   0%| | 4/1258 [00:56<5:00:29, 14.38sLlama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.39 ms /    50 runs   (    0.37
ms per token,  2718.28 tokens per second)
llama_print_timings: prompt eval time =   12926.23 ms /   130 tokens (   99.43
ms per token,    10.06 tokens per second)
llama_print_timings:        eval time =   10699.64 ms /    49 runs   (  218.36
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   23886.58 ms /   179 tokens
No. of rows:   0%| | 5/1258 [01:20<6:11:59, 17.81sLlama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.01 ms /    20 runs   (    0.35
ms per token,  2853.47 tokens per second)
llama_print_timings: prompt eval time =    9054.91 ms /    89 tokens (  101.74
ms per token,     9.83 tokens per second)
llama_print_timings:        eval time =    4009.81 ms /    19 runs   (  211.04
ms per token,     4.74 tokens per second)
llama_print_timings:       total time =   13162.92 ms /   108 tokens
```

```
No. of rows:   0%| | 6/1258 [01:33<5:38:46, 16.23sLlama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.88 ms /    20 runs   (    0.34
ms per token,  2906.55 tokens per second)
llama_print_timings: prompt eval time =   10950.48 ms /   106 tokens (  103.31
ms per token,     9.68 tokens per second)
llama_print_timings:        eval time =    4023.02 ms /    19 runs   (  211.74
ms per token,     4.72 tokens per second)
llama_print_timings:       total time =   15072.98 ms /   125 tokens
No. of rows:   1%| | 7/1258 [01:48<5:30:35, 15.86sLlama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.70 ms /    30 runs   (    0.36
ms per token,  2804.52 tokens per second)
llama_print_timings: prompt eval time =   12035.81 ms /   111 tokens (  108.43
ms per token,     9.22 tokens per second)
llama_print_timings:        eval time =    6137.50 ms /    29 runs   (  211.64
ms per token,     4.73 tokens per second)
llama_print_timings:       total time =   18324.20 ms /   140 tokens
No. of rows:   1%| | 8/1258 [02:07<5:46:45, 16.64sLlama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.89 ms /    28 runs   (    0.39
ms per token,  2570.69 tokens per second)
llama_print_timings: prompt eval time =   10249.48 ms /   102 tokens (  100.49
ms per token,     9.95 tokens per second)
llama_print_timings:        eval time =    5846.03 ms /    27 runs   (  216.52
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   16243.26 ms /   129 tokens
No. of rows:   1%| | 9/1258 [02:23<5:43:53, 16.52sLlama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.33 ms /    50 runs   (    0.39
ms per token,  2586.39 tokens per second)
llama_print_timings: prompt eval time =   21834.19 ms /   213 tokens (  102.51
ms per token,     9.76 tokens per second)
llama_print_timings:        eval time =   10669.69 ms /    49 runs   (  217.75
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   32768.75 ms /   262 tokens
No. of rows:   1%| | 10/1258 [02:56<7:28:01, 21.54Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       7.50 ms /    20 runs   (    0.37
ms per token,  2668.45 tokens per second)
llama_print_timings: prompt eval time =   11448.56 ms /    96 tokens (  119.26
ms per token,     8.39 tokens per second)
llama_print_timings:        eval time =    4138.90 ms /    19 runs   (  217.84
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   15693.99 ms /   115 tokens
No. of rows:   1%| | 11/1258 [03:11<6:50:29, 19.75Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.07 ms /    40 runs   (    0.40
ms per token,  2489.57 tokens per second)
llama_print_timings: prompt eval time =   15402.81 ms /   135 tokens (  114.09
ms per token,     8.76 tokens per second)
llama_print_timings:        eval time =    8586.20 ms /    39 runs   (  220.16
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   24207.82 ms /   174 tokens
No. of rows:   1%| | 12/1258 [03:36<7:18:23, 21.11Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.47 ms /    22 runs   (    0.39
ms per token,  2597.40 tokens per second)
llama_print_timings: prompt eval time =   10514.10 ms /    83 tokens (  126.68
ms per token,     7.89 tokens per second)
llama_print_timings:        eval time =    4992.82 ms /    21 runs   (  237.75
ms per token,     4.21 tokens per second)
llama_print_timings:       total time =   15628.08 ms /   104 tokens
No. of rows:   1%| | 13/1258 [03:51<6:43:35, 19.45Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.21 ms /    24 runs   (    0.38
ms per token,  2605.58 tokens per second)
llama_print_timings: prompt eval time =    8810.33 ms /    82 tokens (  107.44
ms per token,     9.31 tokens per second)
llama_print_timings:        eval time =    5185.77 ms /    23 runs   (  225.47
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   14127.87 ms /   105 tokens
No. of rows:   1%| | 14/1258 [04:05<6:10:00, 17.85Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.69 ms /    36 runs   (    0.38
ms per token,  2629.66 tokens per second)
llama_print_timings: prompt eval time =   10816.48 ms /   103 tokens (  105.01
ms per token,     9.52 tokens per second)
```

```
llama_print_timings:        eval time =    7637.51 ms /    35 runs   (  218.21
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   18643.35 ms /   138 tokens
No. of rows:   1%| | 15/1258 [04:24<6:14:41, 18.09Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.68 ms /    29 runs   (    0.40
ms per token,  2482.88 tokens per second)
llama_print_timings: prompt eval time =   10877.65 ms /    87 tokens (  125.03
ms per token,     8.00 tokens per second)
llama_print_timings:        eval time =    6259.71 ms /    28 runs   (  223.56
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   17294.55 ms /   115 tokens
No. of rows:   1%| | 16/1258 [04:41<6:09:28, 17.85Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.61 ms /    26 runs   (    0.37
ms per token,  2704.39 tokens per second)
llama_print_timings: prompt eval time =    9543.65 ms /    91 tokens (  104.88
ms per token,     9.54 tokens per second)
llama_print_timings:        eval time =    7164.83 ms /    25 runs   (  286.59
ms per token,     3.49 tokens per second)
llama_print_timings:       total time =   16845.07 ms /   116 tokens
No. of rows:   1%| | 17/1258 [04:58<6:02:58, 17.55Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.28 ms /    26 runs   (    0.40
ms per token,  2530.17 tokens per second)
llama_print_timings: prompt eval time =   10106.73 ms /    96 tokens (  105.28
ms per token,     9.50 tokens per second)
llama_print_timings:        eval time =    7234.61 ms /    25 runs   (  289.38
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   17481.91 ms /   121 tokens
No. of rows:   1%| | 18/1258 [05:16<6:02:19, 17.53Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.99 ms /    21 runs   (    0.38
ms per token,  2627.96 tokens per second)
llama_print_timings: prompt eval time =    8177.19 ms /    78 tokens (  104.84
ms per token,     9.54 tokens per second)
llama_print_timings:        eval time =    4745.23 ms /    20 runs   (  237.26
ms per token,     4.21 tokens per second)
llama_print_timings:       total time =   13036.70 ms /    98 tokens
No. of rows:   2%| | 19/1258 [05:29<5:34:09, 16.18Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.36 ms /    18 runs   (    0.41
ms per token,  2446.32 tokens per second)
llama_print_timings: prompt eval time =    8316.96 ms /    79 tokens (  105.28
ms per token,     9.50 tokens per second)
llama_print_timings:        eval time =    3795.26 ms /    17 runs   (  223.25
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   12213.45 ms /    96 tokens
No. of rows:   2%| | 20/1258 [05:41<5:09:25, 15.00Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.90 ms /    50 runs   (    0.38
ms per token,  2645.36 tokens per second)
llama_print_timings: prompt eval time =   18303.03 ms /   165 tokens (  110.93
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =   10810.45 ms /    49 runs   (  220.62
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   29390.91 ms /   214 tokens
No. of rows:   2%| | 21/1258 [06:10<6:38:16, 19.32Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.30 ms /    18 runs   (    0.41
ms per token,  2465.42 tokens per second)
llama_print_timings: prompt eval time =    9559.56 ms /    89 tokens (  107.41
ms per token,     9.31 tokens per second)
llama_print_timings:        eval time =    3736.62 ms /    17 runs   (  219.80
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   13403.19 ms /   106 tokens
No. of rows:   2%| | 22/1258 [06:24<6:01:23, 17.54Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.39 ms /    29 runs   (    0.39
ms per token,  2546.99 tokens per second)
llama_print_timings: prompt eval time =   10096.37 ms /    93 tokens (  108.56
ms per token,     9.21 tokens per second)
llama_print_timings:        eval time =    6135.00 ms /    28 runs   (  219.11
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   16389.93 ms /   121 tokens
No. of rows:   2%| | 23/1258 [06:40<5:54:05, 17.20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.63 ms /    23 runs   (    0.42
```

```
ms per token,  2387.87 tokens per second)
llama_print_timings: prompt eval time =   10873.56 ms /    88 tokens (  123.56
ms per token,     8.09 tokens per second)
llama_print_timings:        eval time =    4932.51 ms /    22 runs   (  224.21
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   15941.79 ms /   110 tokens
No. of rows:   2%| | 24/1258 [06:56<5:46:04, 16.83Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.30 ms /    34 runs   (    0.42
ms per token,  2377.12 tokens per second)
llama_print_timings: prompt eval time =   12692.73 ms /   105 tokens (  120.88
ms per token,     8.27 tokens per second)
llama_print_timings:        eval time =    9020.20 ms /    33 runs   (  273.34
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   21907.76 ms /   138 tokens
No. of rows:   2%| | 25/1258 [07:18<6:17:05, 18.35Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.34 ms /    25 runs   (    0.37
ms per token,  2677.81 tokens per second)
llama_print_timings: prompt eval time =   10023.31 ms /    92 tokens (  108.95
ms per token,     9.18 tokens per second)
llama_print_timings:        eval time =    6899.54 ms /    24 runs   (  287.48
ms per token,     3.48 tokens per second)
llama_print_timings:       total time =   17058.96 ms /   116 tokens
No. of rows:   2%| | 26/1258 [07:35<6:08:56, 17.97Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.50 ms /    21 runs   (    0.40
ms per token,  2469.43 tokens per second)
llama_print_timings: prompt eval time =    8622.01 ms /    79 tokens (  109.14
ms per token,     9.16 tokens per second)
llama_print_timings:        eval time =    4686.16 ms /    20 runs   (  234.31
ms per token,     4.27 tokens per second)
llama_print_timings:       total time =   13427.86 ms /    99 tokens
No. of rows:   2%| | 27/1258 [07:49<5:40:41, 16.61Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.90 ms /    23 runs   (    0.39
ms per token,  2585.43 tokens per second)
llama_print_timings: prompt eval time =   11180.42 ms /   104 tokens (  107.50
ms per token,     9.30 tokens per second)
llama_print_timings:        eval time =    4820.77 ms /    22 runs   (  219.13
```

```
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   16130.22 ms /   126 tokens
No. of rows:   2%| | 28/1258 [08:05<5:37:34, 16.47Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.58 ms /    31 runs   (    0.37
ms per token,  2676.57 tokens per second)
llama_print_timings: prompt eval time =   13031.49 ms /   120 tokens (  108.60
ms per token,     9.21 tokens per second)
llama_print_timings:        eval time =    7960.60 ms /    30 runs   (  265.35
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   21161.78 ms /   150 tokens
No. of rows:   2%| | 29/1258 [08:26<6:06:13, 17.88Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.60 ms /    20 runs   (    0.38
ms per token,  2631.93 tokens per second)
llama_print_timings: prompt eval time =    9480.77 ms /    88 tokens (  107.74
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    4155.89 ms /    19 runs   (  218.73
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   13746.29 ms /   107 tokens
No. of rows:   2%| | 30/1258 [08:40<5:40:31, 16.64Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.08 ms /    18 runs   (    0.39
ms per token,  2544.17 tokens per second)
llama_print_timings: prompt eval time =   11088.35 ms /    87 tokens (  127.45
ms per token,     7.85 tokens per second)
llama_print_timings:        eval time =    3742.91 ms /    17 runs   (  220.17
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   14931.52 ms /   104 tokens
No. of rows:   2%| | 31/1258 [08:55<5:29:51, 16.13Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.56 ms /    50 runs   (    0.39
ms per token,  2555.71 tokens per second)
llama_print_timings: prompt eval time =   12470.36 ms /   115 tokens (  108.44
ms per token,     9.22 tokens per second)
llama_print_timings:        eval time =   12560.06 ms /    49 runs   (  256.33
ms per token,     3.90 tokens per second)
llama_print_timings:       total time =   25313.95 ms /   164 tokens
No. of rows:   3%| | 32/1258 [09:20<6:25:59, 18.89Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.28 ms /    24 runs   (    0.39
ms per token,  2586.49 tokens per second)
llama_print_timings: prompt eval time =    9942.43 ms /    90 tokens (  110.47
ms per token,     9.05 tokens per second)
llama_print_timings:        eval time =    5077.98 ms /    23 runs   (  220.78
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   15154.82 ms /   113 tokens
No. of rows:   3%| | 33/1258 [09:35<6:02:48, 17.77Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      20.00 ms /    50 runs   (    0.40
ms per token,  2499.75 tokens per second)
llama_print_timings: prompt eval time =   12101.68 ms /    95 tokens (  127.39
ms per token,     7.85 tokens per second)
llama_print_timings:        eval time =   11276.23 ms /    49 runs   (  230.13
ms per token,     4.35 tokens per second)
llama_print_timings:       total time =   23667.59 ms /   144 tokens
No. of rows:   3%| | 34/1258 [09:59<6:38:41, 19.54Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.78 ms /    22 runs   (    0.44
ms per token,  2249.49 tokens per second)
llama_print_timings: prompt eval time =    9456.48 ms /    84 tokens (  112.58
ms per token,     8.88 tokens per second)
llama_print_timings:        eval time =    4860.15 ms /    21 runs   (  231.44
ms per token,     4.32 tokens per second)
llama_print_timings:       total time =   14454.37 ms /   105 tokens
No. of rows:   3%| | 35/1258 [10:13<6:07:17, 18.02Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.38 ms /    27 runs   (    0.38
ms per token,  2600.65 tokens per second)
llama_print_timings: prompt eval time =   10924.96 ms /    98 tokens (  111.48
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    7976.78 ms /    26 runs   (  306.80
ms per token,     3.26 tokens per second)
llama_print_timings:       total time =   19053.19 ms /   124 tokens
No. of rows:   3%| | 36/1258 [10:32<6:13:21, 18.33Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.33 ms /    20 runs   (    0.42
ms per token,  2400.38 tokens per second)
```

```
llama_print_timings: prompt eval time =   10066.79 ms /    92 tokens (  109.42
ms per token,     9.14 tokens per second)
llama_print_timings:        eval time =    5605.60 ms /    19 runs   (  295.03
ms per token,     3.39 tokens per second)
llama_print_timings:       total time =   15788.81 ms /   111 tokens
No. of rows:   3%| | 37/1258 [10:48<5:57:32, 17.57Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.67 ms /    25 runs   (    0.39
ms per token,  2585.05 tokens per second)
llama_print_timings: prompt eval time =    9898.21 ms /    89 tokens (  111.22
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    5314.71 ms /    24 runs   (  221.45
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   15352.39 ms /   113 tokens
No. of rows:   3%| | 38/1258 [11:03<5:43:48, 16.91Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.69 ms /    33 runs   (    0.38
ms per token,  2600.06 tokens per second)
llama_print_timings: prompt eval time =   12729.44 ms /   103 tokens (  123.59
ms per token,     8.09 tokens per second)
llama_print_timings:        eval time =    8848.07 ms /    32 runs   (  276.50
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   21762.99 ms /   135 tokens
No. of rows:   3%| | 39/1258 [11:25<6:13:05, 18.36Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.76 ms /    29 runs   (    0.41
ms per token,  2465.15 tokens per second)
llama_print_timings: prompt eval time =   10800.44 ms /    99 tokens (  109.10
ms per token,     9.17 tokens per second)
llama_print_timings:        eval time =    6618.24 ms /    28 runs   (  236.37
ms per token,     4.23 tokens per second)
llama_print_timings:       total time =   17588.35 ms /   127 tokens
No. of rows:   3%| | 40/1258 [11:43<6:08:06, 18.13Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.95 ms /    50 runs   (    0.40
ms per token,  2506.39 tokens per second)
llama_print_timings: prompt eval time =   13563.97 ms /   121 tokens (  112.10
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =   10944.42 ms /    49 runs   (  223.36
ms per token,     4.48 tokens per second)
```

```
llama_print_timings:       total time =   24792.63 ms /   170 tokens
No. of rows:   3%| | 41/1258 [12:08<6:48:27, 20.14Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.51 ms /    24 runs   (    0.40
ms per token,  2523.13 tokens per second)
llama_print_timings: prompt eval time =    8788.94 ms /    78 tokens (  112.68
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    5885.47 ms /    23 runs   (  255.89
ms per token,     3.91 tokens per second)
llama_print_timings:       total time =   14812.35 ms /   101 tokens
No. of rows:   3%| | 42/1258 [12:22<6:15:43, 18.54Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.80 ms /    30 runs   (    0.39
ms per token,  2542.37 tokens per second)
llama_print_timings: prompt eval time =   11174.46 ms /   102 tokens (  109.55
ms per token,     9.13 tokens per second)
llama_print_timings:        eval time =    6482.72 ms /    29 runs   (  223.54
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   17829.35 ms /   131 tokens
No. of rows:   3%| | 43/1258 [12:40<6:11:11, 18.33Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.52 ms /    50 runs   (    0.39
ms per token,  2561.74 tokens per second)
llama_print_timings: prompt eval time =   13673.27 ms /   110 tokens (  124.30
ms per token,     8.04 tokens per second)
llama_print_timings:        eval time =   12663.23 ms /    49 runs   (  258.43
ms per token,     3.87 tokens per second)
llama_print_timings:       total time =   26623.95 ms /   159 tokens
No. of rows:   3%| | 44/1258 [13:07<7:01:18, 20.82Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.11 ms /    28 runs   (    0.40
ms per token,  2519.34 tokens per second)
llama_print_timings: prompt eval time =    9602.70 ms /    86 tokens (  111.66
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =    7518.56 ms /    27 runs   (  278.47
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   17280.13 ms /   113 tokens
No. of rows:   4%| | 45/1258 [13:24<6:39:29, 19.76Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.73 ms /    30 runs   (    0.39
ms per token,  2556.89 tokens per second)
llama_print_timings: prompt eval time =    9878.84 ms /    92 tokens (  107.38
ms per token,     9.31 tokens per second)
llama_print_timings:        eval time =    6397.98 ms /    29 runs   (  220.62
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   16447.40 ms /   121 tokens
No. of rows:   4%| | 46/1258 [13:41<6:19:05, 18.77Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.09 ms /    42 runs   (    0.38
ms per token,  2609.67 tokens per second)
llama_print_timings: prompt eval time =   15434.81 ms /   123 tokens (  125.49
ms per token,     7.97 tokens per second)
llama_print_timings:        eval time =    9288.83 ms /    41 runs   (  226.56
ms per token,     4.41 tokens per second)
llama_print_timings:       total time =   24967.45 ms /   164 tokens
No. of rows:   4%| | 47/1258 [14:06<6:56:25, 20.63Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.37 ms /    23 runs   (    0.41
ms per token,  2453.60 tokens per second)
llama_print_timings: prompt eval time =    9988.00 ms /    88 tokens (  113.50
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    4840.33 ms /    22 runs   (  220.01
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   14959.47 ms /   110 tokens
No. of rows:   4%| | 48/1258 [14:21<6:21:48, 18.93Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.96 ms /    26 runs   (    0.50
ms per token,  2006.48 tokens per second)
llama_print_timings: prompt eval time =   12942.20 ms /   104 tokens (  124.44
ms per token,     8.04 tokens per second)
llama_print_timings:        eval time =    5571.35 ms /    25 runs   (  222.85
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   18667.82 ms /   129 tokens
No. of rows:   4%| | 49/1258 [14:39<6:19:52, 18.85Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.00 ms /    18 runs   (    0.39
ms per token,  2572.53 tokens per second)
llama_print_timings: prompt eval time =    9974.61 ms /    74 tokens (  134.79
```

```
ms per token,     7.42 tokens per second)
llama_print_timings:        eval time =    3763.45 ms /    17 runs   (  221.38
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   13841.57 ms /    91 tokens
No. of rows:   4%| | 50/1258 [14:53<5:49:19, 17.35Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.93 ms /    38 runs   (    0.39
ms per token,  2545.55 tokens per second)
llama_print_timings: prompt eval time =   11949.34 ms /   106 tokens (  112.73
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    9578.31 ms /    37 runs   (  258.87
ms per token,     3.86 tokens per second)
llama_print_timings:       total time =   21748.88 ms /   143 tokens
No. of rows:   4%| | 51/1258 [15:15<6:15:40, 18.67Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.93 ms /    25 runs   (    0.40
ms per token,  2517.12 tokens per second)
llama_print_timings: prompt eval time =   11127.11 ms /   100 tokens (  111.27
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    5328.50 ms /    24 runs   (  222.02
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   16602.79 ms /   124 tokens
No. of rows:   4%| | 52/1258 [15:31<6:02:52, 18.05Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.64 ms /    25 runs   (    0.39
ms per token,  2592.29 tokens per second)
llama_print_timings: prompt eval time =   12906.14 ms /   104 tokens (  124.10
ms per token,     8.06 tokens per second)
llama_print_timings:        eval time =    5343.33 ms /    24 runs   (  222.64
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   18394.97 ms /   128 tokens
No. of rows:   4%| | 53/1258 [15:50<6:04:44, 18.16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.21 ms /    24 runs   (    0.38
ms per token,  2605.01 tokens per second)
llama_print_timings: prompt eval time =   11368.69 ms /    87 tokens (  130.67
ms per token,     7.65 tokens per second)
llama_print_timings:        eval time =    5136.16 ms /    23 runs   (  223.31
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   16640.17 ms /   110 tokens
```

```
No. of rows:   4%| | 54/1258 [16:06<5:55:20, 17.71Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.74 ms /    32 runs   (    0.40
ms per token,  2512.37 tokens per second)
llama_print_timings: prompt eval time =   11036.37 ms /   101 tokens (  109.27
ms per token,     9.15 tokens per second)
llama_print_timings:        eval time =    6895.70 ms /    31 runs   (  222.44
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   18119.22 ms /   132 tokens
No. of rows:   4%| | 55/1258 [16:25<5:57:29, 17.83Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.37 ms /    28 runs   (    0.41
ms per token,  2463.49 tokens per second)
llama_print_timings: prompt eval time =   12060.44 ms /   110 tokens (  109.64
ms per token,     9.12 tokens per second)
llama_print_timings:        eval time =    7685.33 ms /    27 runs   (  284.64
ms per token,     3.51 tokens per second)
llama_print_timings:       total time =   19914.82 ms /   137 tokens
No. of rows:   4%| | 56/1258 [16:45<6:09:49, 18.46Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.86 ms /    32 runs   (    0.40
ms per token,  2489.30 tokens per second)
llama_print_timings: prompt eval time =   12729.88 ms /   113 tokens (  112.65
ms per token,     8.88 tokens per second)
llama_print_timings:        eval time =    7051.46 ms /    31 runs   (  227.47
ms per token,     4.40 tokens per second)
llama_print_timings:       total time =   19967.62 ms /   144 tokens
No. of rows:   5%| | 57/1258 [17:05<6:18:37, 18.92Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.37 ms /    30 runs   (    0.41
ms per token,  2426.01 tokens per second)
llama_print_timings: prompt eval time =   12252.16 ms /   110 tokens (  111.38
ms per token,     8.98 tokens per second)
llama_print_timings:        eval time =    6552.41 ms /    29 runs   (  225.95
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   18983.63 ms /   139 tokens
No. of rows:   5%| | 58/1258 [17:24<6:18:46, 18.94Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =      14.02 ms /    36 runs   (    0.39
ms per token,  2567.76 tokens per second)
llama_print_timings: prompt eval time =   13668.55 ms /   109 tokens (  125.40
ms per token,     7.97 tokens per second)
llama_print_timings:        eval time =    9797.24 ms /    35 runs   (  279.92
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   23673.87 ms /   144 tokens
No. of rows:   5%| | 59/1258 [17:47<6:46:53, 20.36Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.04 ms /    28 runs   (    0.39
ms per token,  2535.54 tokens per second)
llama_print_timings: prompt eval time =   11185.17 ms /   100 tokens (  111.85
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =    6017.99 ms /    27 runs   (  222.89
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   17366.44 ms /   127 tokens
No. of rows:   5%| | 60/1258 [18:05<6:28:34, 19.46Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.28 ms /    20 runs   (    0.41
ms per token,  2414.29 tokens per second)
llama_print_timings: prompt eval time =    9610.96 ms /    86 tokens (  111.76
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    4267.34 ms /    19 runs   (  224.60
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   13996.86 ms /   105 tokens
No. of rows:   5%| | 61/1258 [18:19<5:55:34, 17.82Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.05 ms /    27 runs   (    0.41
ms per token,  2443.88 tokens per second)
llama_print_timings: prompt eval time =   10272.72 ms /    94 tokens (  109.28
ms per token,     9.15 tokens per second)
llama_print_timings:        eval time =    5800.13 ms /    26 runs   (  223.08
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   16228.24 ms /   120 tokens
No. of rows:   5%| | 62/1258 [18:35<5:45:47, 17.35Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.08 ms /    18 runs   (    0.39
ms per token,  2543.81 tokens per second)
llama_print_timings: prompt eval time =    7980.75 ms /    73 tokens (  109.33
ms per token,     9.15 tokens per second)
```

```
llama_print_timings:        eval time =    3822.76 ms /    17 runs   (  224.87
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   11909.75 ms /    90 tokens
No. of rows:   5%| | 63/1258 [18:47<5:13:03, 15.72Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.38 ms /    28 runs   (    0.41
ms per token,  2460.24 tokens per second)
llama_print_timings: prompt eval time =   12274.28 ms /   112 tokens (  109.59
ms per token,     9.12 tokens per second)
llama_print_timings:        eval time =    6101.35 ms /    27 runs   (  225.98
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   18539.74 ms /   139 tokens
No. of rows:   5%| | 64/1258 [19:05<5:29:45, 16.57Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.48 ms /    23 runs   (    0.41
ms per token,  2426.67 tokens per second)
llama_print_timings: prompt eval time =   12632.44 ms /    99 tokens (  127.60
ms per token,     7.84 tokens per second)
llama_print_timings:        eval time =    5013.58 ms /    22 runs   (  227.89
ms per token,     4.39 tokens per second)
llama_print_timings:       total time =   17786.31 ms /   121 tokens
No. of rows:   5%| | 65/1258 [19:23<5:36:41, 16.93Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.92 ms /    26 runs   (    0.42
ms per token,  2380.08 tokens per second)
llama_print_timings: prompt eval time =   10361.35 ms /    95 tokens (  109.07
ms per token,     9.17 tokens per second)
llama_print_timings:        eval time =    5561.02 ms /    25 runs   (  222.44
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   16072.72 ms /   120 tokens
No. of rows:   5%| | 66/1258 [19:39<5:31:18, 16.68Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.10 ms /    20 runs   (    0.46
ms per token,  2197.56 tokens per second)
llama_print_timings: prompt eval time =   10754.63 ms /    82 tokens (  131.15
ms per token,     7.62 tokens per second)
llama_print_timings:        eval time =    4744.28 ms /    19 runs   (  249.70
ms per token,     4.00 tokens per second)
llama_print_timings:       total time =   15623.34 ms /   101 tokens
No. of rows:   5%| | 67/1258 [19:55<5:24:52, 16.37Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.20 ms /    23 runs   (    0.40
ms per token,  2501.09 tokens per second)
llama_print_timings: prompt eval time =    8834.79 ms /    80 tokens (  110.43
ms per token,     9.06 tokens per second)
llama_print_timings:        eval time =    4884.32 ms /    22 runs   (  222.01
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   13853.90 ms /   102 tokens
No. of rows:   5%| | 68/1258 [20:09<5:09:43, 15.62Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.47 ms /    33 runs   (    0.47
ms per token,  2132.89 tokens per second)
llama_print_timings: prompt eval time =   13123.93 ms /   118 tokens (  111.22
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    8917.21 ms /    32 runs   (  278.66
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   22237.56 ms /   150 tokens
No. of rows:   5%| | 69/1258 [20:31<5:48:48, 17.60Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.15 ms /    22 runs   (    0.42
ms per token,  2405.16 tokens per second)
llama_print_timings: prompt eval time =    9806.13 ms /    89 tokens (  110.18
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    5542.77 ms /    21 runs   (  263.94
ms per token,     3.79 tokens per second)
llama_print_timings:       total time =   15483.32 ms /   110 tokens
No. of rows:   6%| | 70/1258 [20:46<5:35:58, 16.97Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.39 ms /    29 runs   (    0.43
ms per token,  2340.03 tokens per second)
llama_print_timings: prompt eval time =   12519.91 ms /   112 tokens (  111.78
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    6247.22 ms /    28 runs   (  223.11
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   18939.52 ms /   140 tokens
No. of rows:   6%| | 71/1258 [21:05<5:47:27, 17.56Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.39 ms /    30 runs   (    0.41
```

```
ms per token,  2421.70 tokens per second)
llama_print_timings: prompt eval time =   12701.33 ms /   100 tokens (  127.01
ms per token,     7.87 tokens per second)
llama_print_timings:        eval time =    6438.67 ms /    29 runs   (  222.02
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   19319.92 ms /   129 tokens
No. of rows:   6%| | 72/1258 [21:25<5:57:38, 18.09Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.35 ms /    22 runs   (    0.43
ms per token,  2352.69 tokens per second)
llama_print_timings: prompt eval time =    9319.04 ms /    85 tokens (  109.64
ms per token,     9.12 tokens per second)
llama_print_timings:        eval time =    6653.16 ms /    21 runs   (  316.82
ms per token,     3.16 tokens per second)
llama_print_timings:       total time =   16108.54 ms /   106 tokens
No. of rows:   6%| | 73/1258 [21:41<5:45:38, 17.50Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.50 ms /    20 runs   (    0.38
ms per token,  2666.31 tokens per second)
llama_print_timings: prompt eval time =    8971.92 ms /    81 tokens (  110.76
ms per token,     9.03 tokens per second)
llama_print_timings:        eval time =    4226.49 ms /    19 runs   (  222.45
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   13314.47 ms /   100 tokens
No. of rows:   6%| | 74/1258 [21:54<5:20:34, 16.25Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.52 ms /    23 runs   (    0.41
ms per token,  2415.46 tokens per second)
llama_print_timings: prompt eval time =   12529.50 ms /   100 tokens (  125.30
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =    4947.08 ms /    22 runs   (  224.87
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   17615.34 ms /   122 tokens
No. of rows:   6%| | 75/1258 [22:12<5:28:28, 16.66Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.68 ms /    25 runs   (    0.39
ms per token,  2581.58 tokens per second)
llama_print_timings: prompt eval time =    8767.91 ms /    78 tokens (  112.41
ms per token,     8.90 tokens per second)
llama_print_timings:        eval time =    5409.64 ms /    24 runs   (  225.40
```

```
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   14327.26 ms /   102 tokens
No. of rows:   6%| | 76/1258 [22:26<5:14:27, 15.96Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.37 ms /    26 runs   (    0.40
ms per token,  2508.20 tokens per second)
llama_print_timings: prompt eval time =   11839.38 ms /    93 tokens (  127.31
ms per token,     7.86 tokens per second)
llama_print_timings:        eval time =    5594.12 ms /    25 runs   (  223.76
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   17582.54 ms /   118 tokens
No. of rows:   6%| | 77/1258 [22:44<5:23:44, 16.45Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.45 ms /    24 runs   (    0.39
ms per token,  2538.88 tokens per second)
llama_print_timings: prompt eval time =   12595.82 ms /    97 tokens (  129.85
ms per token,     7.70 tokens per second)
llama_print_timings:        eval time =    5140.28 ms /    23 runs   (  223.49
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   17876.66 ms /   120 tokens
No. of rows:   6%| | 78/1258 [23:02<5:31:55, 16.88Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.94 ms /    41 runs   (    0.39
ms per token,  2571.50 tokens per second)
llama_print_timings: prompt eval time =   11745.09 ms /    92 tokens (  127.66
ms per token,     7.83 tokens per second)
llama_print_timings:        eval time =    9278.89 ms /    40 runs   (  231.97
ms per token,     4.31 tokens per second)
llama_print_timings:       total time =   21264.57 ms /   132 tokens
No. of rows:   6%| | 79/1258 [23:23<5:57:36, 18.20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.71 ms /    22 runs   (    0.40
ms per token,  2526.70 tokens per second)
llama_print_timings: prompt eval time =    9128.47 ms /    83 tokens (  109.98
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =    4662.07 ms /    21 runs   (  222.00
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   13918.23 ms /   104 tokens
No. of rows:   6%| | 80/1258 [23:37<5:32:08, 16.92Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.72 ms /    21 runs   (    0.42
ms per token,  2408.26 tokens per second)
llama_print_timings: prompt eval time =   10879.86 ms /    84 tokens (  129.52
ms per token,     7.72 tokens per second)
llama_print_timings:        eval time =    4471.95 ms /    20 runs   (  223.60
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   15474.40 ms /   104 tokens
No. of rows:   6%| | 81/1258 [23:52<5:23:24, 16.49Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.15 ms /    24 runs   (    0.38
ms per token,  2622.38 tokens per second)
llama_print_timings: prompt eval time =   11337.84 ms /    86 tokens (  131.84
ms per token,     7.59 tokens per second)
llama_print_timings:        eval time =    5166.59 ms /    23 runs   (  224.63
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   16642.12 ms /   109 tokens
No. of rows:   7%| | 82/1258 [24:09<5:24:08, 16.54Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.40 ms /    18 runs   (    0.41
ms per token,  2432.10 tokens per second)
llama_print_timings: prompt eval time =    8267.22 ms /    74 tokens (  111.72
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    3793.84 ms /    17 runs   (  223.17
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   12176.97 ms /    91 tokens
No. of rows:   7%| | 83/1258 [24:21<4:58:13, 15.23Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.37 ms /    32 runs   (    0.42
ms per token,  2393.06 tokens per second)
llama_print_timings: prompt eval time =   13082.39 ms /   115 tokens (  113.76
ms per token,     8.79 tokens per second)
llama_print_timings:        eval time =    8605.24 ms /    31 runs   (  277.59
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   21877.56 ms /   146 tokens
No. of rows:   7%| | 84/1258 [24:43<5:37:02, 17.23Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.53 ms /    36 runs   (    0.38
ms per token,  2660.95 tokens per second)
```

```
llama_print_timings: prompt eval time =   13296.30 ms /   118 tokens (  112.68
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    9495.23 ms /    35 runs   (  271.29
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   23000.61 ms /   153 tokens
No. of rows:   7%| | 85/1258 [25:06<6:10:41, 18.96Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.92 ms /    34 runs   (    0.44
ms per token,  2279.28 tokens per second)
llama_print_timings: prompt eval time =   11387.03 ms /   104 tokens (  109.49
ms per token,     9.13 tokens per second)
llama_print_timings:        eval time =    9144.38 ms /    33 runs   (  277.10
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   20738.58 ms /   137 tokens
No. of rows:   7%| | 86/1258 [25:27<6:20:52, 19.50Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.44 ms /    34 runs   (    0.42
ms per token,  2354.24 tokens per second)
llama_print_timings: prompt eval time =   13371.72 ms /   120 tokens (  111.43
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    9148.92 ms /    33 runs   (  277.24
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   22725.59 ms /   153 tokens
No. of rows:   7%| | 87/1258 [25:49<6:39:26, 20.47Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.41 ms /    28 runs   (    0.41
ms per token,  2454.63 tokens per second)
llama_print_timings: prompt eval time =   11611.04 ms /   103 tokens (  112.73
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    6101.68 ms /    27 runs   (  225.99
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   17878.56 ms /   130 tokens
No. of rows:   7%| | 88/1258 [26:07<6:24:04, 19.70Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.67 ms /    29 runs   (    0.40
ms per token,  2484.58 tokens per second)
llama_print_timings: prompt eval time =   13337.59 ms /   119 tokens (  112.08
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    7847.49 ms /    28 runs   (  280.27
ms per token,     3.57 tokens per second)
```

```
llama_print_timings:       total time =   21355.03 ms /   147 tokens
No. of rows:   7%| | 89/1258 [26:29<6:33:24, 20.19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.50 ms /    24 runs   (    0.40
ms per token,  2526.85 tokens per second)
llama_print_timings: prompt eval time =    9849.57 ms /    90 tokens (  109.44
ms per token,     9.14 tokens per second)
llama_print_timings:        eval time =    5173.37 ms /    23 runs   (  224.93
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   15166.78 ms /   113 tokens
No. of rows:   7%| | 90/1258 [26:44<6:03:48, 18.69Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.59 ms /    23 runs   (    0.42
ms per token,  2397.83 tokens per second)
llama_print_timings: prompt eval time =   10314.98 ms /    79 tokens (  130.57
ms per token,     7.66 tokens per second)
llama_print_timings:        eval time =    4912.73 ms /    22 runs   (  223.31
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   15365.57 ms /   101 tokens
No. of rows:   7%| | 91/1258 [26:59<5:44:09, 17.69Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.44 ms /    22 runs   (    0.43
ms per token,  2330.51 tokens per second)
llama_print_timings: prompt eval time =   10127.04 ms /    80 tokens (  126.59
ms per token,     7.90 tokens per second)
llama_print_timings:        eval time =    4679.70 ms /    21 runs   (  222.84
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   14938.23 ms /   101 tokens
No. of rows:   7%| | 92/1258 [27:14<5:27:50, 16.87Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.99 ms /    25 runs   (    0.40
ms per token,  2502.50 tokens per second)
llama_print_timings: prompt eval time =   12185.68 ms /    97 tokens (  125.63
ms per token,     7.96 tokens per second)
llama_print_timings:        eval time =    5410.60 ms /    24 runs   (  225.44
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   17750.85 ms /   121 tokens
No. of rows:   7%| | 93/1258 [27:32<5:32:44, 17.14Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.72 ms /    23 runs   (    0.42
ms per token,  2366.99 tokens per second)
llama_print_timings: prompt eval time =   10015.13 ms /    91 tokens (  110.06
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =    6654.99 ms /    22 runs   (  302.50
ms per token,     3.31 tokens per second)
llama_print_timings:       total time =   16809.10 ms /   113 tokens
No. of rows:   7%| | 94/1258 [27:49<5:30:34, 17.04Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.63 ms /    20 runs   (    0.38
ms per token,  2620.55 tokens per second)
llama_print_timings: prompt eval time =   10009.62 ms /    87 tokens (  115.05
ms per token,     8.69 tokens per second)
llama_print_timings:        eval time =    4496.09 ms /    19 runs   (  236.64
ms per token,     4.23 tokens per second)
llama_print_timings:       total time =   14629.42 ms /   106 tokens
No. of rows:   8%| | 95/1258 [28:03<5:16:16, 16.32Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.42 ms /    28 runs   (    0.41
ms per token,  2451.19 tokens per second)
llama_print_timings: prompt eval time =   11119.94 ms /   101 tokens (  110.10
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    6254.56 ms /    27 runs   (  231.65
ms per token,     4.32 tokens per second)
llama_print_timings:       total time =   17542.98 ms /   128 tokens
No. of rows:   8%| | 96/1258 [28:21<5:23:10, 16.69Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.81 ms /    42 runs   (    0.40
ms per token,  2498.36 tokens per second)
llama_print_timings: prompt eval time =   13753.36 ms /   123 tokens (  111.82
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =   11062.63 ms /    41 runs   (  269.82
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   25068.71 ms /   164 tokens
No. of rows:   8%| | 97/1258 [28:46<6:11:39, 19.21Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.18 ms /    26 runs   (    0.39
ms per token,  2554.78 tokens per second)
llama_print_timings: prompt eval time =    9751.03 ms /    88 tokens (  110.81
```

```
ms per token,     9.02 tokens per second)
llama_print_timings:        eval time =    5582.24 ms /    25 runs   (  223.29
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   15485.41 ms /   113 tokens
No. of rows:   8%| | 98/1258 [29:01<5:49:42, 18.09Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.47 ms /    23 runs   (    0.41
ms per token,  2427.44 tokens per second)
llama_print_timings: prompt eval time =   10464.54 ms /    95 tokens (  110.15
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    4929.16 ms /    22 runs   (  224.05
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   15534.99 ms /   117 tokens
No. of rows:   8%| | 99/1258 [29:17<5:34:40, 17.33Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.56 ms /    31 runs   (    0.41
ms per token,  2468.74 tokens per second)
llama_print_timings: prompt eval time =   12793.16 ms /   101 tokens (  126.66
ms per token,     7.89 tokens per second)
llama_print_timings:        eval time =    6680.80 ms /    30 runs   (  222.69
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   19659.10 ms /   131 tokens
No. of rows:   8%| | 100/1258 [29:37<5:47:57, 18.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.19 ms /    24 runs   (    0.38
ms per token,  2611.53 tokens per second)
llama_print_timings: prompt eval time =   10291.10 ms /    79 tokens (  130.27
ms per token,     7.68 tokens per second)
llama_print_timings:        eval time =    5117.45 ms /    23 runs   (  222.50
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   15550.72 ms /   102 tokens
No. of rows:   8%| | 101/1258 [29:52<5:33:23, 17.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.53 ms /    22 runs   (    0.43
ms per token,  2307.53 tokens per second)
llama_print_timings: prompt eval time =   11726.58 ms /    89 tokens (  131.76
ms per token,     7.59 tokens per second)
llama_print_timings:        eval time =    4807.64 ms /    21 runs   (  228.94
ms per token,     4.37 tokens per second)
llama_print_timings:       total time =   16667.93 ms /   110 tokens
```

```
No. of rows:   8%| | 102/1258 [30:09<5:29:32, 17.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.62 ms /    28 runs   (    0.42
ms per token,  2408.81 tokens per second)
llama_print_timings: prompt eval time =   10758.92 ms /    93 tokens (  115.69
ms per token,     8.64 tokens per second)
llama_print_timings:        eval time =    6082.29 ms /    27 runs   (  225.27
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   17017.20 ms /   120 tokens
No. of rows:   8%| | 103/1258 [30:26<5:28:50, 17.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.45 ms /    22 runs   (    0.43
ms per token,  2329.27 tokens per second)
llama_print_timings: prompt eval time =   11934.46 ms /    93 tokens (  128.33
ms per token,     7.79 tokens per second)
llama_print_timings:        eval time =    6157.48 ms /    21 runs   (  293.21
ms per token,     3.41 tokens per second)
llama_print_timings:       total time =   18228.57 ms /   114 tokens
No. of rows:   8%| | 104/1258 [30:44<5:35:11, 17.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.62 ms /    19 runs   (    0.40
ms per token,  2491.80 tokens per second)
llama_print_timings: prompt eval time =    8795.38 ms /    78 tokens (  112.76
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    4382.18 ms /    18 runs   (  243.45
ms per token,     4.11 tokens per second)
llama_print_timings:       total time =   13290.37 ms /    96 tokens
No. of rows:   8%| | 105/1258 [30:57<5:11:05, 16.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.58 ms /    24 runs   (    0.40
ms per token,  2504.70 tokens per second)
llama_print_timings: prompt eval time =   10453.13 ms /    95 tokens (  110.03
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =    5114.07 ms /    23 runs   (  222.35
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   15708.41 ms /   118 tokens
No. of rows:   8%| | 106/1258 [31:13<5:08:04, 16.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       7.76 ms /    19 runs   (    0.41
ms per token,  2447.51 tokens per second)
llama_print_timings: prompt eval time =    9960.70 ms /    76 tokens (  131.06
ms per token,     7.63 tokens per second)
llama_print_timings:        eval time =    4345.45 ms /    18 runs   (  241.41
ms per token,     4.14 tokens per second)
llama_print_timings:       total time =   14417.86 ms /    94 tokens
No. of rows:   9%| | 107/1258 [31:28<4:58:31, 15.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.66 ms /    21 runs   (    0.41
ms per token,  2426.06 tokens per second)
llama_print_timings: prompt eval time =    9266.58 ms /    83 tokens (  111.65
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =    4449.76 ms /    20 runs   (  222.49
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   13844.65 ms /   103 tokens
No. of rows:   9%| | 108/1258 [31:41<4:48:22, 15.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.55 ms /    24 runs   (    0.40
ms per token,  2512.04 tokens per second)
llama_print_timings: prompt eval time =   11529.55 ms /    92 tokens (  125.32
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =    6318.59 ms /    23 runs   (  274.72
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   17995.41 ms /   115 tokens
No. of rows:   9%| | 109/1258 [31:59<5:05:10, 15.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.16 ms /    18 runs   (    0.40
ms per token,  2515.02 tokens per second)
llama_print_timings: prompt eval time =   10549.51 ms /    89 tokens (  118.53
ms per token,     8.44 tokens per second)
llama_print_timings:        eval time =    3910.33 ms /    17 runs   (  230.02
ms per token,     4.35 tokens per second)
llama_print_timings:       total time =   14570.23 ms /   106 tokens
No. of rows:   9%| | 110/1258 [32:14<4:57:05, 15.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.87 ms /    23 runs   (    0.43
ms per token,  2331.47 tokens per second)
llama_print_timings: prompt eval time =   10088.14 ms /    86 tokens (  117.30
ms per token,     8.52 tokens per second)
```

```
llama_print_timings:        eval time =    5307.63 ms /    22 runs   (  241.26
ms per token,     4.14 tokens per second)
llama_print_timings:       total time =   15544.04 ms /   108 tokens
No. of rows:   9%| | 111/1258 [32:30<4:57:00, 15.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.71 ms /    20 runs   (    0.49
ms per token,  2059.94 tokens per second)
llama_print_timings: prompt eval time =   11920.04 ms /    85 tokens (  140.24
ms per token,     7.13 tokens per second)
llama_print_timings:        eval time =    4645.02 ms /    19 runs   (  244.47
ms per token,     4.09 tokens per second)
llama_print_timings:       total time =   16707.83 ms /   104 tokens
No. of rows:   9%| | 112/1258 [32:46<5:03:30, 15.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.77 ms /    33 runs   (    0.42
ms per token,  2395.82 tokens per second)
llama_print_timings: prompt eval time =   15264.47 ms /   109 tokens (  140.04
ms per token,     7.14 tokens per second)
llama_print_timings:        eval time =    9067.57 ms /    32 runs   (  283.36
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   24533.75 ms /   141 tokens
No. of rows:   9%| | 113/1258 [33:11<5:52:42, 18.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.68 ms /    22 runs   (    0.39
ms per token,  2536.02 tokens per second)
llama_print_timings: prompt eval time =    9283.53 ms /    85 tokens (  109.22
ms per token,     9.16 tokens per second)
llama_print_timings:        eval time =    5095.14 ms /    21 runs   (  242.63
ms per token,     4.12 tokens per second)
llama_print_timings:       total time =   14508.26 ms /   106 tokens
No. of rows:   9%| | 114/1258 [33:25<5:29:45, 17.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.59 ms /    34 runs   (    0.40
ms per token,  2501.66 tokens per second)
llama_print_timings: prompt eval time =   12142.46 ms /   110 tokens (  110.39
ms per token,     9.06 tokens per second)
llama_print_timings:        eval time =    8893.03 ms /    33 runs   (  269.49
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   21232.60 ms /   143 tokens
No. of rows:   9%| | 115/1258 [33:47<5:52:01, 18.4Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.85 ms /    26 runs   (    0.53
ms per token,  1877.93 tokens per second)
llama_print_timings: prompt eval time =    9547.11 ms /    87 tokens (  109.74
ms per token,     9.11 tokens per second)
llama_print_timings:        eval time =    6086.98 ms /    25 runs   (  243.48
ms per token,     4.11 tokens per second)
llama_print_timings:       total time =   15797.42 ms /   112 tokens
No. of rows:   9%| | 116/1258 [34:02<5:36:27, 17.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.27 ms /    23 runs   (    0.40
ms per token,  2480.32 tokens per second)
llama_print_timings: prompt eval time =   10490.58 ms /    96 tokens (  109.28
ms per token,     9.15 tokens per second)
llama_print_timings:        eval time =    4996.79 ms /    22 runs   (  227.13
ms per token,     4.40 tokens per second)
llama_print_timings:       total time =   15625.68 ms /   118 tokens
No. of rows:   9%| | 117/1258 [34:18<5:24:30, 17.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.91 ms /    25 runs   (    0.40
ms per token,  2522.45 tokens per second)
llama_print_timings: prompt eval time =   11035.54 ms /   102 tokens (  108.19
ms per token,     9.24 tokens per second)
llama_print_timings:        eval time =    5254.70 ms /    24 runs   (  218.95
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   16432.91 ms /   126 tokens
No. of rows:   9%| | 118/1258 [34:35<5:20:39, 16.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.44 ms /    25 runs   (    0.38
ms per token,  2648.02 tokens per second)
llama_print_timings: prompt eval time =   11560.48 ms /    91 tokens (  127.04
ms per token,     7.87 tokens per second)
llama_print_timings:        eval time =    5238.97 ms /    24 runs   (  218.29
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   16945.44 ms /   115 tokens
No. of rows:   9%| | 119/1258 [34:51<5:20:48, 16.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.63 ms /    37 runs   (    0.42
```

```
ms per token,  2367.85 tokens per second)
llama_print_timings: prompt eval time =   13797.00 ms /   112 tokens (  123.19
ms per token,     8.12 tokens per second)
llama_print_timings:        eval time =    7971.98 ms /    36 runs   (  221.44
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   21983.72 ms /   148 tokens
No. of rows:  10%| | 120/1258 [35:13<5:49:27, 18.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.70 ms /    24 runs   (    0.40
ms per token,  2473.46 tokens per second)
llama_print_timings: prompt eval time =   11650.63 ms /    88 tokens (  132.39
ms per token,     7.55 tokens per second)
llama_print_timings:        eval time =    5162.89 ms /    23 runs   (  224.47
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   16958.24 ms /   111 tokens
No. of rows:  10%| | 121/1258 [35:30<5:40:55, 17.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.86 ms /    30 runs   (    0.43
ms per token,  2333.54 tokens per second)
llama_print_timings: prompt eval time =   11610.13 ms /    91 tokens (  127.58
ms per token,     7.84 tokens per second)
llama_print_timings:        eval time =    8291.42 ms /    29 runs   (  285.91
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   20119.17 ms /   120 tokens
No. of rows:  10%| | 122/1258 [35:51<5:52:43, 18.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.66 ms /    30 runs   (    0.39
ms per token,  2572.02 tokens per second)
llama_print_timings: prompt eval time =   10776.13 ms /    96 tokens (  112.25
ms per token,     8.91 tokens per second)
llama_print_timings:        eval time =    6499.43 ms /    29 runs   (  224.12
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   17453.03 ms /   125 tokens
No. of rows:  10%| | 123/1258 [36:08<5:45:48, 18.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.36 ms /    23 runs   (    0.41
ms per token,  2457.26 tokens per second)
llama_print_timings: prompt eval time =   10018.58 ms /    91 tokens (  110.09
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    5052.73 ms /    22 runs   (  229.67
```

```
ms per token,     4.35 tokens per second)
llama_print_timings:       total time =   15210.81 ms /   113 tokens
No. of rows:  10%| | 124/1258 [36:23<5:28:08, 17.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.69 ms /    29 runs   (    0.44
ms per token,  2284.72 tokens per second)
llama_print_timings: prompt eval time =   12131.91 ms /   110 tokens (  110.29
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    6521.88 ms /    28 runs   (  232.92
ms per token,     4.29 tokens per second)
llama_print_timings:       total time =   18839.16 ms /   138 tokens
No. of rows:  10%| | 125/1258 [36:42<5:36:15, 17.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.20 ms /    28 runs   (    0.40
ms per token,  2499.55 tokens per second)
llama_print_timings: prompt eval time =   12944.78 ms /   102 tokens (  126.91
ms per token,     7.88 tokens per second)
llama_print_timings:        eval time =    6190.03 ms /    27 runs   (  229.26
ms per token,     4.36 tokens per second)
llama_print_timings:       total time =   19309.44 ms /   129 tokens
No. of rows:  10%| | 126/1258 [37:01<5:44:30, 18.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.64 ms /    31 runs   (    0.41
ms per token,  2452.53 tokens per second)
llama_print_timings: prompt eval time =   13098.99 ms /   101 tokens (  129.69
ms per token,     7.71 tokens per second)
llama_print_timings:        eval time =    6749.44 ms /    30 runs   (  224.98
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   20040.26 ms /   131 tokens
No. of rows:  10%| | 127/1258 [37:21<5:54:17, 18.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.53 ms /    32 runs   (    0.39
ms per token,  2554.89 tokens per second)
llama_print_timings: prompt eval time =   14245.78 ms /   114 tokens (  124.96
ms per token,     8.00 tokens per second)
llama_print_timings:        eval time =    8647.53 ms /    31 runs   (  278.95
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   23085.96 ms /   145 tokens
No. of rows:  10%| | 128/1258 [37:45<6:18:15, 20.0Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.80 ms /    19 runs   (    0.41
ms per token,  2436.52 tokens per second)
llama_print_timings: prompt eval time =   10062.93 ms /    91 tokens (  110.58
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =    5071.91 ms /    18 runs   (  281.77
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   15251.51 ms /   109 tokens
No. of rows:  10%| | 129/1258 [38:00<5:50:39, 18.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.56 ms /    21 runs   (    0.41
ms per token,  2452.13 tokens per second)
llama_print_timings: prompt eval time =   10567.53 ms /    94 tokens (  112.42
ms per token,     8.90 tokens per second)
llama_print_timings:        eval time =    4454.39 ms /    20 runs   (  222.72
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   15151.06 ms /   114 tokens
No. of rows:  10%| | 130/1258 [38:15<5:30:45, 17.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.22 ms /    20 runs   (    0.41
ms per token,  2433.98 tokens per second)
llama_print_timings: prompt eval time =    9934.22 ms /    87 tokens (  114.19
ms per token,     8.76 tokens per second)
llama_print_timings:        eval time =    5534.53 ms /    19 runs   (  291.29
ms per token,     3.43 tokens per second)
llama_print_timings:       total time =   15592.14 ms /   106 tokens
No. of rows:  10%| | 131/1258 [38:31<5:19:12, 16.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.20 ms /    23 runs   (    0.40
ms per token,  2500.27 tokens per second)
llama_print_timings: prompt eval time =    8495.92 ms /    78 tokens (  108.92
ms per token,     9.18 tokens per second)
llama_print_timings:        eval time =    4906.29 ms /    22 runs   (  223.01
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   13538.68 ms /   100 tokens
No. of rows:  10%| | 132/1258 [38:44<4:59:32, 15.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.93 ms /    26 runs   (    0.42
ms per token,  2378.56 tokens per second)
```

```
llama_print_timings: prompt eval time =   10550.64 ms /    80 tokens (  131.88
ms per token,     7.58 tokens per second)
llama_print_timings:        eval time =    6035.57 ms /    25 runs   (  241.42
ms per token,     4.14 tokens per second)
llama_print_timings:       total time =   16743.93 ms /   105 tokens
No. of rows:  11%| | 133/1258 [39:01<5:03:43, 16.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.92 ms /    20 runs   (    0.40
ms per token,  2524.61 tokens per second)
llama_print_timings: prompt eval time =    9090.49 ms /    81 tokens (  112.23
ms per token,     8.91 tokens per second)
llama_print_timings:        eval time =    4229.06 ms /    19 runs   (  222.58
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   13437.13 ms /   100 tokens
No. of rows:  11%| | 134/1258 [39:14<4:47:59, 15.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.52 ms /    31 runs   (    0.40
ms per token,  2476.04 tokens per second)
llama_print_timings: prompt eval time =   14797.70 ms /   123 tokens (  120.31
ms per token,     8.31 tokens per second)
llama_print_timings:        eval time =    6692.53 ms /    30 runs   (  223.08
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   21676.86 ms /   153 tokens
No. of rows:  11%| | 135/1258 [39:36<5:23:10, 17.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.49 ms /    22 runs   (    0.39
ms per token,  2592.20 tokens per second)
llama_print_timings: prompt eval time =   11662.27 ms /    92 tokens (  126.76
ms per token,     7.89 tokens per second)
llama_print_timings:        eval time =    6218.10 ms /    21 runs   (  296.10
ms per token,     3.38 tokens per second)
llama_print_timings:       total time =   18010.33 ms /   113 tokens
No. of rows:  11%| | 136/1258 [39:54<5:27:05, 17.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.67 ms /    22 runs   (    0.39
ms per token,  2538.95 tokens per second)
llama_print_timings: prompt eval time =    9993.35 ms /    89 tokens (  112.28
ms per token,     8.91 tokens per second)
llama_print_timings:        eval time =    4833.60 ms /    21 runs   (  230.17
ms per token,     4.34 tokens per second)
```

```
llama_print_timings:       total time =   14956.97 ms /   110 tokens
No. of rows:  11%| | 137/1258 [40:09<5:12:38, 16.7Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.87 ms /    26 runs   (    0.42
ms per token,  2391.90 tokens per second)
llama_print_timings: prompt eval time =    8863.22 ms /    81 tokens (  109.42
ms per token,     9.14 tokens per second)
llama_print_timings:        eval time =    5729.65 ms /    25 runs   (  229.19
ms per token,     4.36 tokens per second)
llama_print_timings:       total time =   14754.71 ms /   106 tokens
No. of rows:  11%| | 138/1258 [40:24<5:01:19, 16.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.09 ms /    23 runs   (    0.40
ms per token,  2531.65 tokens per second)
llama_print_timings: prompt eval time =   10189.07 ms /    91 tokens (  111.97
ms per token,     8.93 tokens per second)
llama_print_timings:        eval time =    4907.12 ms /    22 runs   (  223.05
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   15237.65 ms /   113 tokens
No. of rows:  11%| | 139/1258 [40:39<4:56:01, 15.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.07 ms /    23 runs   (    0.39
ms per token,  2535.83 tokens per second)
llama_print_timings: prompt eval time =   10810.73 ms /    82 tokens (  131.84
ms per token,     7.59 tokens per second)
llama_print_timings:        eval time =    4875.79 ms /    22 runs   (  221.63
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   15822.79 ms /   104 tokens
No. of rows:  11%| | 140/1258 [40:55<4:55:31, 15.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.81 ms /    22 runs   (    0.40
ms per token,  2496.31 tokens per second)
llama_print_timings: prompt eval time =   10862.36 ms /    84 tokens (  129.31
ms per token,     7.73 tokens per second)
llama_print_timings:        eval time =    5669.78 ms /    21 runs   (  269.99
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   16664.91 ms /   105 tokens
No. of rows:  11%| | 141/1258 [41:11<4:59:44, 16.1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.68 ms /    23 runs   (    0.42
ms per token,  2377.26 tokens per second)
llama_print_timings: prompt eval time =    9978.07 ms /    89 tokens (  112.11
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    5191.24 ms /    22 runs   (  235.97
ms per token,     4.24 tokens per second)
llama_print_timings:       total time =   15309.58 ms /   111 tokens
No. of rows:  11%| | 142/1258 [41:27<4:55:05, 15.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.77 ms /    21 runs   (    0.42
ms per token,  2395.89 tokens per second)
llama_print_timings: prompt eval time =    9800.75 ms /    88 tokens (  111.37
ms per token,     8.98 tokens per second)
llama_print_timings:        eval time =    4448.00 ms /    20 runs   (  222.40
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   14372.89 ms /   108 tokens
No. of rows:  11%| | 143/1258 [41:41<4:46:33, 15.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.81 ms /    21 runs   (    0.42
ms per token,  2382.57 tokens per second)
llama_print_timings: prompt eval time =   11665.13 ms /    90 tokens (  129.61
ms per token,     7.72 tokens per second)
llama_print_timings:        eval time =    4770.20 ms /    20 runs   (  238.51
ms per token,     4.19 tokens per second)
llama_print_timings:       total time =   16561.50 ms /   110 tokens
No. of rows:  11%| | 144/1258 [41:58<4:52:42, 15.7Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.05 ms /    29 runs   (    0.45
ms per token,  2222.22 tokens per second)
llama_print_timings: prompt eval time =   11746.31 ms /   104 tokens (  112.95
ms per token,     8.85 tokens per second)
llama_print_timings:        eval time =    6694.16 ms /    28 runs   (  239.08
ms per token,     4.18 tokens per second)
llama_print_timings:       total time =   18633.93 ms /   132 tokens
No. of rows:  12%| | 145/1258 [42:16<5:08:26, 16.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.58 ms /    21 runs   (    0.41
ms per token,  2447.27 tokens per second)
llama_print_timings: prompt eval time =    9363.50 ms /    83 tokens (  112.81
```

```
ms per token,     8.86 tokens per second)
llama_print_timings:        eval time =    4498.01 ms /    20 runs   (  224.90
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   13986.74 ms /   103 tokens
No. of rows:  12%| | 146/1258 [42:30<4:53:33, 15.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.75 ms /    35 runs   (    0.39
ms per token,  2545.45 tokens per second)
llama_print_timings: prompt eval time =   14235.44 ms /   114 tokens (  124.87
ms per token,     8.01 tokens per second)
llama_print_timings:        eval time =    7626.77 ms /    34 runs   (  224.32
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   22075.69 ms /   148 tokens
No. of rows:  12%| | 147/1258 [42:52<5:27:59, 17.7Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.26 ms /    32 runs   (    0.38
ms per token,  2611.18 tokens per second)
llama_print_timings: prompt eval time =   13001.33 ms /   104 tokens (  125.01
ms per token,     8.00 tokens per second)
llama_print_timings:        eval time =    8633.25 ms /    31 runs   (  278.49
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   21825.76 ms /   135 tokens
No. of rows:  12%| | 148/1258 [43:14<5:50:30, 18.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.36 ms /    26 runs   (    0.40
ms per token,  2509.89 tokens per second)
llama_print_timings: prompt eval time =   11399.69 ms /    99 tokens (  115.15
ms per token,     8.68 tokens per second)
llama_print_timings:        eval time =    5627.96 ms /    25 runs   (  225.12
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   17186.47 ms /   124 tokens
No. of rows:  12%| | 149/1258 [43:31<5:40:30, 18.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.53 ms /    29 runs   (    0.40
ms per token,  2514.31 tokens per second)
llama_print_timings: prompt eval time =   11804.71 ms /   106 tokens (  111.37
ms per token,     8.98 tokens per second)
llama_print_timings:        eval time =    7982.61 ms /    28 runs   (  285.09
ms per token,     3.51 tokens per second)
llama_print_timings:       total time =   19959.94 ms /   134 tokens
```

```
No. of rows:  12%| | 150/1258 [43:51<5:48:40, 18.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.74 ms /    29 runs   (    0.40
ms per token,  2471.03 tokens per second)
llama_print_timings: prompt eval time =   11398.37 ms /   100 tokens (  113.98
ms per token,     8.77 tokens per second)
llama_print_timings:        eval time =    6534.10 ms /    28 runs   (  233.36
ms per token,     4.29 tokens per second)
llama_print_timings:       total time =   18109.21 ms /   128 tokens
No. of rows:  12%| | 151/1258 [44:10<5:44:12, 18.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.85 ms /    31 runs   (    0.41
ms per token,  2412.64 tokens per second)
llama_print_timings: prompt eval time =   10380.67 ms /    94 tokens (  110.43
ms per token,     9.06 tokens per second)
llama_print_timings:        eval time =    6776.17 ms /    30 runs   (  225.87
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   17346.33 ms /   124 tokens
No. of rows:  12%| | 152/1258 [44:27<5:36:38, 18.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.91 ms /    24 runs   (    0.41
ms per token,  2421.55 tokens per second)
llama_print_timings: prompt eval time =    8614.20 ms /    78 tokens (  110.44
ms per token,     9.05 tokens per second)
llama_print_timings:        eval time =    5183.81 ms /    23 runs   (  225.38
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   13941.13 ms /   101 tokens
No. of rows:  12%| | 153/1258 [44:41<5:12:33, 16.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.04 ms /    42 runs   (    0.41
ms per token,  2464.64 tokens per second)
llama_print_timings: prompt eval time =   17553.26 ms /   146 tokens (  120.23
ms per token,     8.32 tokens per second)
llama_print_timings:        eval time =    9204.85 ms /    41 runs   (  224.51
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   27020.40 ms /   187 tokens
No. of rows:  12%| | 154/1258 [45:08<6:07:47, 19.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =      19.67 ms /    46 runs   (    0.43
ms per token,  2338.59 tokens per second)
llama_print_timings: prompt eval time =   13457.62 ms /   107 tokens (  125.77
ms per token,     7.95 tokens per second)
llama_print_timings:        eval time =   10322.76 ms /    45 runs   (  229.39
ms per token,     4.36 tokens per second)
llama_print_timings:       total time =   24068.51 ms /   152 tokens
No. of rows:  12%| | 155/1258 [45:32<6:29:58, 21.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.16 ms /    20 runs   (    0.41
ms per token,  2450.38 tokens per second)
llama_print_timings: prompt eval time =    8816.85 ms /    80 tokens (  110.21
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    4217.93 ms /    19 runs   (  222.00
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   13155.18 ms /    99 tokens
No. of rows:  12%| | 156/1258 [45:45<5:45:15, 18.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.54 ms /    26 runs   (    0.41
ms per token,  2466.09 tokens per second)
llama_print_timings: prompt eval time =   11919.45 ms /    94 tokens (  126.80
ms per token,     7.89 tokens per second)
llama_print_timings:        eval time =    7431.84 ms /    25 runs   (  297.27
ms per token,     3.36 tokens per second)
llama_print_timings:       total time =   19508.97 ms /   119 tokens
No. of rows:  12%| | 157/1258 [46:05<5:48:56, 19.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.24 ms /    25 runs   (    0.41
ms per token,  2441.88 tokens per second)
llama_print_timings: prompt eval time =   11573.82 ms /   103 tokens (  112.37
ms per token,     8.90 tokens per second)
llama_print_timings:        eval time =    5407.23 ms /    24 runs   (  225.30
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   17133.56 ms /   127 tokens
No. of rows:  13%| | 158/1258 [46:22<5:38:17, 18.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.70 ms /    36 runs   (    0.41
ms per token,  2449.65 tokens per second)
llama_print_timings: prompt eval time =   11780.73 ms /   105 tokens (  112.20
ms per token,     8.91 tokens per second)
```

```
llama_print_timings:        eval time =    9578.55 ms /    35 runs   (  273.67
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   21574.28 ms /   140 tokens
No. of rows:  13%| | 159/1258 [46:43<5:55:10, 19.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.07 ms /    36 runs   (    0.47
ms per token,  2108.47 tokens per second)
llama_print_timings: prompt eval time =   14166.81 ms /   129 tokens (  109.82
ms per token,     9.11 tokens per second)
llama_print_timings:        eval time =    8408.61 ms /    35 runs   (  240.25
ms per token,     4.16 tokens per second)
llama_print_timings:       total time =   22810.34 ms /   164 tokens
No. of rows:  13%| | 160/1258 [47:06<6:13:40, 20.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.64 ms /    24 runs   (    0.40
ms per token,  2489.37 tokens per second)
llama_print_timings: prompt eval time =   10396.10 ms /    80 tokens (  129.95
ms per token,     7.70 tokens per second)
llama_print_timings:        eval time =    5178.73 ms /    23 runs   (  225.16
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   15725.09 ms /   103 tokens
No. of rows:  13%| | 161/1258 [47:22<5:47:33, 19.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.28 ms /    20 runs   (    0.41
ms per token,  2415.17 tokens per second)
llama_print_timings: prompt eval time =   10750.31 ms /    84 tokens (  127.98
ms per token,     7.81 tokens per second)
llama_print_timings:        eval time =    4231.47 ms /    19 runs   (  222.71
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   15102.07 ms /   103 tokens
No. of rows:  13%| | 162/1258 [47:37<5:25:53, 17.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.83 ms /    34 runs   (    0.41
ms per token,  2457.71 tokens per second)
llama_print_timings: prompt eval time =   13887.95 ms /   127 tokens (  109.35
ms per token,     9.14 tokens per second)
llama_print_timings:        eval time =    9078.28 ms /    33 runs   (  275.10
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   23176.91 ms /   160 tokens
No. of rows:  13%| | 163/1258 [48:00<5:54:50, 19.4Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.49 ms /    22 runs   (    0.39
ms per token,  2591.89 tokens per second)
llama_print_timings: prompt eval time =   10281.04 ms /    89 tokens (  115.52
ms per token,     8.66 tokens per second)
llama_print_timings:        eval time =    4674.72 ms /    21 runs   (  222.61
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   15085.76 ms /   110 tokens
No. of rows:  13%| | 164/1258 [48:15<5:30:46, 18.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.59 ms /    36 runs   (    0.41
ms per token,  2467.78 tokens per second)
llama_print_timings: prompt eval time =   13383.72 ms /   120 tokens (  111.53
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    7820.97 ms /    35 runs   (  223.46
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   21422.10 ms /   155 tokens
No. of rows:  13%| | 165/1258 [48:37<5:48:26, 19.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.62 ms /    24 runs   (    0.40
ms per token,  2494.28 tokens per second)
llama_print_timings: prompt eval time =   10167.42 ms /    76 tokens (  133.78
ms per token,     7.47 tokens per second)
llama_print_timings:        eval time =    5294.42 ms /    23 runs   (  230.19
ms per token,     4.34 tokens per second)
llama_print_timings:       total time =   15605.14 ms /    99 tokens
No. of rows:  13%| | 166/1258 [48:52<5:28:52, 18.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.63 ms /    25 runs   (    0.39
ms per token,  2595.78 tokens per second)
llama_print_timings: prompt eval time =   12165.65 ms /    94 tokens (  129.42
ms per token,     7.73 tokens per second)
llama_print_timings:        eval time =    6923.97 ms /    24 runs   (  288.50
ms per token,     3.47 tokens per second)
llama_print_timings:       total time =   19240.04 ms /   118 tokens
No. of rows:  13%| | 167/1258 [49:12<5:35:02, 18.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.09 ms /    21 runs   (    0.39
```

```
ms per token,  2594.51 tokens per second)
llama_print_timings: prompt eval time =    8915.57 ms /    80 tokens (  111.44
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    4757.09 ms /    20 runs   (  237.85
ms per token,     4.20 tokens per second)
llama_print_timings:       total time =   13795.97 ms /   100 tokens
No. of rows:  13%| | 168/1258 [49:25<5:09:33, 17.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.98 ms /    32 runs   (    0.41
ms per token,  2465.33 tokens per second)
llama_print_timings: prompt eval time =   11739.84 ms /   107 tokens (  109.72
ms per token,     9.11 tokens per second)
llama_print_timings:        eval time =    8544.20 ms /    31 runs   (  275.62
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   20478.43 ms /   138 tokens
No. of rows:  13%| | 169/1258 [49:46<5:27:57, 18.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.52 ms /    23 runs   (    0.41
ms per token,  2415.21 tokens per second)
llama_print_timings: prompt eval time =    8944.40 ms /    81 tokens (  110.42
ms per token,     9.06 tokens per second)
llama_print_timings:        eval time =    4883.61 ms /    22 runs   (  221.98
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   13967.80 ms /   103 tokens
No. of rows:  14%| | 170/1258 [50:00<5:05:23, 16.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       4.45 ms /    11 runs   (    0.40
ms per token,  2470.24 tokens per second)
llama_print_timings: prompt eval time =   11000.47 ms /    83 tokens (  132.54
ms per token,     7.55 tokens per second)
llama_print_timings:        eval time =    2279.03 ms /    10 runs   (  227.90
ms per token,     4.39 tokens per second)
llama_print_timings:       total time =   13346.83 ms /    93 tokens
No. of rows:  14%| | 171/1258 [50:13<4:46:13, 15.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.77 ms /    21 runs   (    0.42
ms per token,  2395.35 tokens per second)
llama_print_timings: prompt eval time =    9614.20 ms /    86 tokens (  111.79
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    4490.34 ms /    20 runs   (  224.52
```

```
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   14239.85 ms /   106 tokens
No. of rows:  14%| | 172/1258 [50:27<4:37:28, 15.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.22 ms /    31 runs   (    0.39
ms per token,  2536.82 tokens per second)
llama_print_timings: prompt eval time =   12562.39 ms /   114 tokens (  110.20
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    7803.24 ms /    30 runs   (  260.11
ms per token,     3.84 tokens per second)
llama_print_timings:       total time =   20554.65 ms /   144 tokens
No. of rows:  14%| | 173/1258 [50:48<5:05:36, 16.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.71 ms /    28 runs   (    0.45
ms per token,  2202.64 tokens per second)
llama_print_timings: prompt eval time =   11669.32 ms /   106 tokens (  110.09
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    6033.64 ms /    27 runs   (  223.47
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   17876.74 ms /   133 tokens
No. of rows:  14%| | 174/1258 [51:06<5:10:42, 17.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.02 ms /    28 runs   (    0.39
ms per token,  2540.83 tokens per second)
llama_print_timings: prompt eval time =   12614.89 ms /   100 tokens (  126.15
ms per token,     7.93 tokens per second)
llama_print_timings:        eval time =    6008.58 ms /    27 runs   (  222.54
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   18789.80 ms /   127 tokens
No. of rows:  14%| | 175/1258 [51:25<5:19:02, 17.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.36 ms /    29 runs   (    0.39
ms per token,  2552.14 tokens per second)
llama_print_timings: prompt eval time =   13174.73 ms /   105 tokens (  125.47
ms per token,     7.97 tokens per second)
llama_print_timings:        eval time =    8219.71 ms /    28 runs   (  293.56
ms per token,     3.41 tokens per second)
llama_print_timings:       total time =   21570.88 ms /   133 tokens
No. of rows:  14%| | 176/1258 [51:46<5:39:50, 18.8Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      21.35 ms /    50 runs   (    0.43
ms per token,  2341.92 tokens per second)
llama_print_timings: prompt eval time =   13450.99 ms /   122 tokens (  110.25
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =   12904.89 ms /    49 runs   (  263.37
ms per token,     3.80 tokens per second)
llama_print_timings:       total time =   26671.44 ms /   171 tokens
No. of rows:  14%| | 177/1258 [52:13<6:21:56, 21.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.70 ms /    29 runs   (    0.40
ms per token,  2479.27 tokens per second)
llama_print_timings: prompt eval time =   13461.05 ms /   119 tokens (  113.12
ms per token,     8.84 tokens per second)
llama_print_timings:        eval time =    6275.70 ms /    28 runs   (  224.13
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   19915.17 ms /   147 tokens
No. of rows:  14%| | 178/1258 [52:33<6:14:39, 20.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.78 ms /    34 runs   (    0.41
ms per token,  2467.88 tokens per second)
llama_print_timings: prompt eval time =   13175.02 ms /   118 tokens (  111.65
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =    9063.73 ms /    33 runs   (  274.66
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   22442.49 ms /   151 tokens
No. of rows:  14%| | 179/1258 [52:55<6:23:07, 21.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.87 ms /    20 runs   (    0.44
ms per token,  2254.54 tokens per second)
llama_print_timings: prompt eval time =    9300.59 ms /    83 tokens (  112.06
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    4280.47 ms /    19 runs   (  225.29
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   13708.36 ms /   102 tokens
No. of rows:  14%| | 180/1258 [53:09<5:41:54, 19.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.92 ms /    27 runs   (    0.40
ms per token,  2471.62 tokens per second)
```

```
llama_print_timings: prompt eval time =   12074.58 ms /    97 tokens (  124.48
ms per token,     8.03 tokens per second)
llama_print_timings:        eval time =    5788.89 ms /    26 runs   (  222.65
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   18030.44 ms /   123 tokens
No. of rows:  14%| | 181/1258 [53:27<5:36:13, 18.7Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.52 ms /    30 runs   (    0.38
ms per token,  2604.85 tokens per second)
llama_print_timings: prompt eval time =   13024.36 ms /   104 tokens (  125.23
ms per token,     7.99 tokens per second)
llama_print_timings:        eval time =    8176.05 ms /    29 runs   (  281.93
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   21378.98 ms /   133 tokens
No. of rows:  14%| | 182/1258 [53:48<5:50:11, 19.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.10 ms /    28 runs   (    0.40
ms per token,  2521.61 tokens per second)
llama_print_timings: prompt eval time =   11033.86 ms /   100 tokens (  110.34
ms per token,     9.06 tokens per second)
llama_print_timings:        eval time =    8043.41 ms /    27 runs   (  297.90
ms per token,     3.36 tokens per second)
llama_print_timings:       total time =   19248.80 ms /   127 tokens
No. of rows:  15%| | 183/1258 [54:08<5:48:25, 19.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.45 ms /    27 runs   (    0.39
ms per token,  2584.23 tokens per second)
llama_print_timings: prompt eval time =   10672.37 ms /    94 tokens (  113.54
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    5864.41 ms /    26 runs   (  225.55
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   16699.93 ms /   120 tokens
No. of rows:  15%| | 184/1258 [54:24<5:33:23, 18.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.81 ms /    25 runs   (    0.51
ms per token,  1952.06 tokens per second)
llama_print_timings: prompt eval time =   10719.24 ms /    97 tokens (  110.51
ms per token,     9.05 tokens per second)
llama_print_timings:        eval time =    7097.97 ms /    24 runs   (  295.75
ms per token,     3.38 tokens per second)
```

```
llama_print_timings:       total time =   17972.85 ms /   121 tokens
No. of rows:  15%| | 185/1258 [54:42<5:29:34, 18.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.33 ms /    28 runs   (    0.40
ms per token,  2471.32 tokens per second)
llama_print_timings: prompt eval time =   11917.37 ms /   106 tokens (  112.43
ms per token,     8.89 tokens per second)
llama_print_timings:        eval time =    7796.79 ms /    27 runs   (  288.77
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   19883.81 ms /   133 tokens
No. of rows:  15%| | 186/1258 [55:02<5:37:08, 18.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.27 ms /    22 runs   (    0.42
ms per token,  2372.48 tokens per second)
llama_print_timings: prompt eval time =   10732.64 ms /    97 tokens (  110.65
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =    6424.29 ms /    21 runs   (  305.92
ms per token,     3.27 tokens per second)
llama_print_timings:       total time =   17292.39 ms /   118 tokens
No. of rows:  15%| | 187/1258 [55:20<5:28:21, 18.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.70 ms /    25 runs   (    0.43
ms per token,  2337.32 tokens per second)
llama_print_timings: prompt eval time =   10603.15 ms /    93 tokens (  114.01
ms per token,     8.77 tokens per second)
llama_print_timings:        eval time =    6011.93 ms /    24 runs   (  250.50
ms per token,     3.99 tokens per second)
llama_print_timings:       total time =   16774.29 ms /   117 tokens
No. of rows:  15%| | 188/1258 [55:36<5:19:23, 17.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.21 ms /    33 runs   (    0.40
ms per token,  2497.73 tokens per second)
llama_print_timings: prompt eval time =   10462.73 ms /    95 tokens (  110.13
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    7097.72 ms /    32 runs   (  221.80
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   17758.66 ms /   127 tokens
No. of rows:  15%| | 189/1258 [55:54<5:18:24, 17.8Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.98 ms /    30 runs   (    0.40
ms per token,  2503.13 tokens per second)
llama_print_timings: prompt eval time =   13330.40 ms /   106 tokens (  125.76
ms per token,     7.95 tokens per second)
llama_print_timings:        eval time =    6466.39 ms /    29 runs   (  222.98
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   19975.86 ms /   135 tokens
No. of rows:  15%| | 190/1258 [56:14<5:29:19, 18.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.97 ms /    50 runs   (    0.40
ms per token,  2504.26 tokens per second)
llama_print_timings: prompt eval time =   12480.16 ms /   112 tokens (  111.43
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =   11074.10 ms /    49 runs   (  226.00
ms per token,     4.42 tokens per second)
llama_print_timings:       total time =   23862.66 ms /   161 tokens
No. of rows:  15%| | 191/1258 [56:38<5:57:43, 20.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.39 ms /    26 runs   (    0.40
ms per token,  2501.20 tokens per second)
llama_print_timings: prompt eval time =    9293.06 ms /    84 tokens (  110.63
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =    5564.51 ms /    25 runs   (  222.58
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   15013.06 ms /   109 tokens
No. of rows:  15%| | 192/1258 [56:53<5:30:12, 18.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.24 ms /    26 runs   (    0.39
ms per token,  2538.07 tokens per second)
llama_print_timings: prompt eval time =   13409.43 ms /   106 tokens (  126.50
ms per token,     7.90 tokens per second)
llama_print_timings:        eval time =    7239.10 ms /    25 runs   (  289.56
ms per token,     3.45 tokens per second)
llama_print_timings:       total time =   20807.48 ms /   131 tokens
No. of rows:  15%| | 193/1258 [57:14<5:41:46, 19.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.54 ms /    23 runs   (    0.41
ms per token,  2411.41 tokens per second)
llama_print_timings: prompt eval time =   10186.50 ms /    89 tokens (  114.46
```

```
ms per token,     8.74 tokens per second)
llama_print_timings:        eval time =    4892.19 ms /    22 runs   (  222.37
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   15219.41 ms /   111 tokens
No. of rows:  15%| | 194/1258 [57:29<5:20:01, 18.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.61 ms /    42 runs   (    0.40
ms per token,  2529.05 tokens per second)
llama_print_timings: prompt eval time =   14735.07 ms /   136 tokens (  108.35
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =    9371.30 ms /    41 runs   (  228.57
ms per token,     4.38 tokens per second)
llama_print_timings:       total time =   24363.09 ms /   177 tokens
No. of rows:  16%| | 195/1258 [57:53<5:53:16, 19.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.94 ms /    33 runs   (    0.39
ms per token,  2549.44 tokens per second)
llama_print_timings: prompt eval time =   13530.40 ms /   107 tokens (  126.45
ms per token,     7.91 tokens per second)
llama_print_timings:        eval time =    7163.24 ms /    32 runs   (  223.85
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   20897.71 ms /   139 tokens
No. of rows:  16%| | 196/1258 [58:14<5:58:04, 20.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.90 ms /    25 runs   (    0.40
ms per token,  2524.23 tokens per second)
llama_print_timings: prompt eval time =   12607.27 ms /   112 tokens (  112.56
ms per token,     8.88 tokens per second)
llama_print_timings:        eval time =    5345.68 ms /    24 runs   (  222.74
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   18104.22 ms /   136 tokens
No. of rows:  16%| | 197/1258 [58:32<5:46:30, 19.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.90 ms /    20 runs   (    0.39
ms per token,  2533.25 tokens per second)
llama_print_timings: prompt eval time =    9750.87 ms /    87 tokens (  112.08
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    4186.12 ms /    19 runs   (  220.32
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   14055.82 ms /   106 tokens
```

```
No. of rows:  16%| | 198/1258 [58:47<5:16:54, 17.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.91 ms /    22 runs   (    0.41
ms per token,  2468.58 tokens per second)
llama_print_timings: prompt eval time =   11051.33 ms /    86 tokens (  128.50
ms per token,     7.78 tokens per second)
llama_print_timings:        eval time =    4631.14 ms /    21 runs   (  220.53
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   15816.93 ms /   107 tokens
No. of rows:  16%| | 199/1258 [59:02<5:05:21, 17.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.43 ms /    23 runs   (    0.41
ms per token,  2439.80 tokens per second)
llama_print_timings: prompt eval time =   12074.01 ms /    93 tokens (  129.83
ms per token,     7.70 tokens per second)
llama_print_timings:        eval time =    4902.10 ms /    22 runs   (  222.82
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   17117.63 ms /   115 tokens
No. of rows:  16%| | 200/1258 [59:19<5:04:11, 17.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.30 ms /    37 runs   (    0.44
ms per token,  2269.52 tokens per second)
llama_print_timings: prompt eval time =   13376.22 ms /   108 tokens (  123.85
ms per token,     8.07 tokens per second)
llama_print_timings:        eval time =    9716.26 ms /    36 runs   (  269.90
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   23319.69 ms /   144 tokens
No. of rows:  16%| | 201/1258 [59:43<5:36:01, 19.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.27 ms /    25 runs   (    0.41
ms per token,  2434.51 tokens per second)
llama_print_timings: prompt eval time =   10555.61 ms /    95 tokens (  111.11
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =    5343.74 ms /    24 runs   (  222.66
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   16050.94 ms /   119 tokens
No. of rows:  16%| | 202/1258 [59:59<5:19:45, 18.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       8.90 ms /    23 runs   (    0.39
ms per token,  2584.27 tokens per second)
llama_print_timings: prompt eval time =   11292.47 ms /    86 tokens (  131.31
ms per token,     7.62 tokens per second)
llama_print_timings:        eval time =    5010.46 ms /    22 runs   (  227.75
ms per token,     4.39 tokens per second)
llama_print_timings:       total time =   16440.16 ms /   108 tokens
No. of rows:  16%| | 203/1258 [1:00:15<5:10:22, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.53 ms /    27 runs   (    0.39
ms per token,  2565.32 tokens per second)
llama_print_timings: prompt eval time =   11813.61 ms /   103 tokens (  114.70
ms per token,     8.72 tokens per second)
llama_print_timings:        eval time =    5782.68 ms /    26 runs   (  222.41
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   17758.46 ms /   129 tokens
No. of rows:  16%| | 204/1258 [1:00:33<5:10:40, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.42 ms /    23 runs   (    0.41
ms per token,  2442.39 tokens per second)
llama_print_timings: prompt eval time =   10755.42 ms /    95 tokens (  113.21
ms per token,     8.83 tokens per second)
llama_print_timings:        eval time =    4879.09 ms /    22 runs   (  221.78
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   15772.48 ms /   117 tokens
No. of rows:  16%| | 205/1258 [1:00:49<5:00:18, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.77 ms /    29 runs   (    0.37
ms per token,  2691.91 tokens per second)
llama_print_timings: prompt eval time =   13563.37 ms /   110 tokens (  123.30
ms per token,     8.11 tokens per second)
llama_print_timings:        eval time =    6271.78 ms /    28 runs   (  223.99
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   20007.38 ms /   138 tokens
No. of rows:  16%| | 206/1258 [1:01:09<5:15:16, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.91 ms /    32 runs   (    0.40
ms per token,  2477.93 tokens per second)
llama_print_timings: prompt eval time =   11867.44 ms /    93 tokens (  127.61
ms per token,     7.84 tokens per second)
```

```
llama_print_timings:        eval time =    6904.34 ms /    31 runs   (  222.72
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   18965.54 ms /   124 tokens
No. of rows:  16%| | 207/1258 [1:01:28<5:20:15, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.06 ms /    26 runs   (    0.39
ms per token,  2583.98 tokens per second)
llama_print_timings: prompt eval time =   10020.66 ms /    90 tokens (  111.34
ms per token,     8.98 tokens per second)
llama_print_timings:        eval time =    5546.58 ms /    25 runs   (  221.86
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   15723.12 ms /   115 tokens
No. of rows:  17%| | 208/1258 [1:01:44<5:06:31, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.63 ms /    29 runs   (    0.40
ms per token,  2494.19 tokens per second)
llama_print_timings: prompt eval time =   12559.64 ms /    99 tokens (  126.87
ms per token,     7.88 tokens per second)
llama_print_timings:        eval time =    8005.47 ms /    28 runs   (  285.91
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   20740.49 ms /   127 tokens
No. of rows:  17%| | 209/1258 [1:02:04<5:23:08, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.70 ms /    28 runs   (    0.42
ms per token,  2392.96 tokens per second)
llama_print_timings: prompt eval time =   15150.16 ms /   135 tokens (  112.22
ms per token,     8.91 tokens per second)
llama_print_timings:        eval time =    6321.68 ms /    27 runs   (  234.14
ms per token,     4.27 tokens per second)
llama_print_timings:       total time =   21648.97 ms /   162 tokens
No. of rows:  17%| | 210/1258 [1:02:26<5:39:32, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.75 ms /    21 runs   (    0.42
ms per token,  2399.18 tokens per second)
llama_print_timings: prompt eval time =    8688.47 ms /    79 tokens (  109.98
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =    4492.05 ms /    20 runs   (  224.60
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   13312.66 ms /    99 tokens
No. of rows:  17%| | 211/1258 [1:02:39<5:07:09, 17Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.08 ms /    24 runs   (    0.38
ms per token,  2643.75 tokens per second)
llama_print_timings: prompt eval time =    9157.68 ms /    81 tokens (  113.06
ms per token,     8.85 tokens per second)
llama_print_timings:        eval time =    5123.08 ms /    23 runs   (  222.74
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   14428.65 ms /   104 tokens
No. of rows:  17%| | 212/1258 [1:02:54<4:50:17, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.41 ms /    30 runs   (    0.38
ms per token,  2628.58 tokens per second)
llama_print_timings: prompt eval time =   12574.54 ms /    99 tokens (  127.02
ms per token,     7.87 tokens per second)
llama_print_timings:        eval time =    6478.49 ms /    29 runs   (  223.40
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   19234.36 ms /   128 tokens
No. of rows:  17%| | 213/1258 [1:03:13<5:03:36, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.06 ms /    31 runs   (    0.42
ms per token,  2373.66 tokens per second)
llama_print_timings: prompt eval time =   13604.42 ms /   107 tokens (  127.14
ms per token,     7.87 tokens per second)
llama_print_timings:        eval time =    8389.30 ms /    30 runs   (  279.64
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   22188.02 ms /   137 tokens
No. of rows:  17%| | 214/1258 [1:03:35<5:28:09, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.97 ms /    41 runs   (    0.41
ms per token,  2416.31 tokens per second)
llama_print_timings: prompt eval time =   15700.64 ms /   142 tokens (  110.57
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =    9009.61 ms /    40 runs   (  225.24
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   24965.16 ms /   182 tokens
No. of rows:  17%| | 215/1258 [1:04:00<5:59:44, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.32 ms /    47 runs   (    0.41
```

```
ms per token,  2432.71 tokens per second)
llama_print_timings: prompt eval time =   14658.46 ms /   117 tokens (  125.29
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =   10451.91 ms /    46 runs   (  227.22
ms per token,     4.40 tokens per second)
llama_print_timings:       total time =   25400.84 ms /   163 tokens
No. of rows:  17%| | 216/1258 [1:04:26<6:23:56, 22Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.92 ms /    32 runs   (    0.40
ms per token,  2476.01 tokens per second)
llama_print_timings: prompt eval time =   11406.81 ms /   103 tokens (  110.75
ms per token,     9.03 tokens per second)
llama_print_timings:        eval time =    6891.59 ms /    31 runs   (  222.31
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   18492.43 ms /   134 tokens
No. of rows:  17%| | 217/1258 [1:04:44<6:04:44, 21Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.38 ms /    35 runs   (    0.41
ms per token,  2434.78 tokens per second)
llama_print_timings: prompt eval time =   14487.47 ms /   117 tokens (  123.82
ms per token,     8.08 tokens per second)
llama_print_timings:        eval time =    7716.95 ms /    34 runs   (  226.97
ms per token,     4.41 tokens per second)
llama_print_timings:       total time =   22420.07 ms /   151 tokens
No. of rows:  17%| | 218/1258 [1:05:06<6:11:45, 21Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.41 ms /    29 runs   (    0.39
ms per token,  2542.08 tokens per second)
llama_print_timings: prompt eval time =   11022.05 ms /    85 tokens (  129.67
ms per token,     7.71 tokens per second)
llama_print_timings:        eval time =    6217.17 ms /    28 runs   (  222.04
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   17415.43 ms /   113 tokens
No. of rows:  17%| | 219/1258 [1:05:24<5:50:26, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.40 ms /    26 runs   (    0.40
ms per token,  2499.52 tokens per second)
llama_print_timings: prompt eval time =   13234.80 ms /   106 tokens (  124.86
ms per token,     8.01 tokens per second)
llama_print_timings:        eval time =    5553.83 ms /    25 runs   (  222.15
```

```
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   18946.26 ms /   131 tokens
No. of rows:  17%| | 220/1258 [1:05:43<5:43:29, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.14 ms /    32 runs   (    0.41
ms per token,  2435.13 tokens per second)
llama_print_timings: prompt eval time =   14810.82 ms /   119 tokens (  124.46
ms per token,     8.03 tokens per second)
llama_print_timings:        eval time =    6923.94 ms /    31 runs   (  223.35
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   21928.94 ms /   150 tokens
No. of rows:  18%| | 221/1258 [1:06:05<5:53:55, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.95 ms /    29 runs   (    0.41
ms per token,  2427.59 tokens per second)
llama_print_timings: prompt eval time =   13784.59 ms /   108 tokens (  127.64
ms per token,     7.83 tokens per second)
llama_print_timings:        eval time =    6401.98 ms /    28 runs   (  228.64
ms per token,     4.37 tokens per second)
llama_print_timings:       total time =   20368.80 ms /   136 tokens
No. of rows:  18%| | 222/1258 [1:06:25<5:53:05, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.18 ms /    18 runs   (    0.40
ms per token,  2508.36 tokens per second)
llama_print_timings: prompt eval time =    8608.06 ms /    77 tokens (  111.79
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    3816.51 ms /    17 runs   (  224.50
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   12534.20 ms /    94 tokens
No. of rows:  18%| | 223/1258 [1:06:38<5:11:48, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.45 ms /    36 runs   (    0.40
ms per token,  2491.69 tokens per second)
llama_print_timings: prompt eval time =   13004.51 ms /   115 tokens (  113.08
ms per token,     8.84 tokens per second)
llama_print_timings:        eval time =    7795.01 ms /    35 runs   (  222.71
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   21017.33 ms /   150 tokens
No. of rows:  18%| | 224/1258 [1:06:59<5:26:42, 18Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.87 ms /    21 runs   (    0.37
ms per token,  2669.72 tokens per second)
llama_print_timings: prompt eval time =   11199.58 ms /    87 tokens (  128.73
ms per token,     7.77 tokens per second)
llama_print_timings:        eval time =    4447.82 ms /    20 runs   (  222.39
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   15774.00 ms /   107 tokens
No. of rows:  18%| | 225/1258 [1:07:15<5:10:03, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.73 ms /    22 runs   (    0.40
ms per token,  2520.33 tokens per second)
llama_print_timings: prompt eval time =   10723.69 ms /    82 tokens (  130.78
ms per token,     7.65 tokens per second)
llama_print_timings:        eval time =    4664.49 ms /    21 runs   (  222.12
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   15520.33 ms /   103 tokens
No. of rows:  18%| | 226/1258 [1:07:30<4:56:53, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.71 ms /    20 runs   (    0.39
ms per token,  2594.71 tokens per second)
llama_print_timings: prompt eval time =    9592.88 ms /    86 tokens (  111.55
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =    4192.64 ms /    19 runs   (  220.67
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   13905.40 ms /   105 tokens
No. of rows:  18%| | 227/1258 [1:07:44<4:39:23, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.94 ms /    35 runs   (    0.40
ms per token,  2511.30 tokens per second)
llama_print_timings: prompt eval time =   15484.65 ms /   127 tokens (  121.93
ms per token,     8.20 tokens per second)
llama_print_timings:        eval time =    9448.59 ms /    34 runs   (  277.90
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   25147.27 ms /   161 tokens
No. of rows:  18%| | 228/1258 [1:08:09<5:24:56, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.56 ms /    18 runs   (    0.42
ms per token,  2380.32 tokens per second)
```

```
llama_print_timings: prompt eval time =    8441.50 ms /    76 tokens (  111.07
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =    3764.82 ms /    17 runs   (  221.46
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   12323.52 ms /    93 tokens
No. of rows:  18%| | 229/1258 [1:08:21<4:50:35, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.54 ms /    19 runs   (    0.40
ms per token,  2519.23 tokens per second)
llama_print_timings: prompt eval time =    7827.60 ms /    69 tokens (  113.44
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    3988.52 ms /    18 runs   (  221.58
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   11929.87 ms /    87 tokens
No. of rows:  18%| | 230/1258 [1:08:33<4:24:39, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.99 ms /    20 runs   (    0.40
ms per token,  2504.38 tokens per second)
llama_print_timings: prompt eval time =   10425.56 ms /    92 tokens (  113.32
ms per token,     8.82 tokens per second)
llama_print_timings:        eval time =    4654.53 ms /    19 runs   (  244.98
ms per token,     4.08 tokens per second)
llama_print_timings:       total time =   15201.45 ms /   111 tokens
No. of rows:  18%| | 231/1258 [1:08:49<4:23:07, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.30 ms /    26 runs   (    0.40
ms per token,  2525.25 tokens per second)
llama_print_timings: prompt eval time =   11467.76 ms /   104 tokens (  110.27
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    5548.47 ms /    25 runs   (  221.94
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   17175.37 ms /   129 tokens
No. of rows:  18%| | 232/1258 [1:09:06<4:32:10, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.03 ms /    29 runs   (    0.38
ms per token,  2630.15 tokens per second)
llama_print_timings: prompt eval time =   10356.92 ms /    94 tokens (  110.18
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    6210.60 ms /    28 runs   (  221.81
ms per token,     4.51 tokens per second)
```

```
llama_print_timings:       total time =   16741.07 ms /   122 tokens
No. of rows:  19%| | 233/1258 [1:09:23<4:36:11, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.94 ms /    18 runs   (    0.39
ms per token,  2592.54 tokens per second)
llama_print_timings: prompt eval time =   10575.27 ms /    81 tokens (  130.56
ms per token,     7.66 tokens per second)
llama_print_timings:        eval time =    4141.36 ms /    17 runs   (  243.61
ms per token,     4.10 tokens per second)
llama_print_timings:       total time =   14824.88 ms /    98 tokens
No. of rows:  19%| | 234/1258 [1:09:37<4:29:02, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.76 ms /    26 runs   (    0.41
ms per token,  2417.03 tokens per second)
llama_print_timings: prompt eval time =   10598.53 ms /    95 tokens (  111.56
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =    5564.89 ms /    25 runs   (  222.60
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   16323.31 ms /   120 tokens
No. of rows:  19%| | 235/1258 [1:09:54<4:31:40, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.54 ms /    30 runs   (    0.42
ms per token,  2393.11 tokens per second)
llama_print_timings: prompt eval time =   14983.48 ms /   121 tokens (  123.83
ms per token,     8.08 tokens per second)
llama_print_timings:        eval time =    6668.60 ms /    29 runs   (  229.95
ms per token,     4.35 tokens per second)
llama_print_timings:       total time =   21839.92 ms /   150 tokens
No. of rows:  19%| | 236/1258 [1:10:15<5:01:35, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.98 ms /    30 runs   (    0.40
ms per token,  2505.01 tokens per second)
llama_print_timings: prompt eval time =   11087.60 ms /    98 tokens (  113.14
ms per token,     8.84 tokens per second)
llama_print_timings:        eval time =    6438.54 ms /    29 runs   (  222.02
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   17712.89 ms /   127 tokens
No. of rows:  19%| | 237/1258 [1:10:33<5:01:25, 17Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.11 ms /    21 runs   (    0.39
ms per token,  2590.67 tokens per second)
llama_print_timings: prompt eval time =    8851.53 ms /    81 tokens (  109.28
ms per token,     9.15 tokens per second)
llama_print_timings:        eval time =    4460.91 ms /    20 runs   (  223.05
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   13439.71 ms /   101 tokens
No. of rows:  19%| | 238/1258 [1:10:47<4:39:24, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.38 ms /    36 runs   (    0.40
ms per token,  2503.30 tokens per second)
llama_print_timings: prompt eval time =   13744.63 ms /   122 tokens (  112.66
ms per token,     8.88 tokens per second)
llama_print_timings:        eval time =    7876.15 ms /    35 runs   (  225.03
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   21846.02 ms /   157 tokens
No. of rows:  19%| | 239/1258 [1:11:09<5:06:43, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.21 ms /    20 runs   (    0.41
ms per token,  2436.35 tokens per second)
llama_print_timings: prompt eval time =   10333.11 ms /    79 tokens (  130.80
ms per token,     7.65 tokens per second)
llama_print_timings:        eval time =    4664.23 ms /    19 runs   (  245.49
ms per token,     4.07 tokens per second)
llama_print_timings:       total time =   15119.56 ms /    98 tokens
No. of rows:  19%| | 240/1258 [1:11:24<4:51:24, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.94 ms /    28 runs   (    0.39
ms per token,  2558.71 tokens per second)
llama_print_timings: prompt eval time =   11498.13 ms /   105 tokens (  109.51
ms per token,     9.13 tokens per second)
llama_print_timings:        eval time =    6037.41 ms /    27 runs   (  223.61
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   17707.13 ms /   132 tokens
No. of rows:  19%| | 241/1258 [1:11:41<4:53:56, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.87 ms /    20 runs   (    0.39
ms per token,  2541.30 tokens per second)
llama_print_timings: prompt eval time =    9721.23 ms /    74 tokens (  131.37
```

```
ms per token,     7.61 tokens per second)
llama_print_timings:        eval time =    4201.53 ms /    19 runs   (  221.13
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   14041.61 ms /    93 tokens
No. of rows:  19%| | 242/1258 [1:11:55<4:36:51, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.85 ms /    31 runs   (    0.41
ms per token,  2413.39 tokens per second)
llama_print_timings: prompt eval time =   10951.52 ms /    95 tokens (  115.28
ms per token,     8.67 tokens per second)
llama_print_timings:        eval time =    7016.03 ms /    30 runs   (  233.87
ms per token,     4.28 tokens per second)
llama_print_timings:       total time =   18168.73 ms /   125 tokens
No. of rows:  19%| | 243/1258 [1:12:14<4:45:52, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.83 ms /    30 runs   (    0.43
ms per token,  2338.82 tokens per second)
llama_print_timings: prompt eval time =   10411.86 ms /    93 tokens (  111.96
ms per token,     8.93 tokens per second)
llama_print_timings:        eval time =    6907.89 ms /    29 runs   (  238.20
ms per token,     4.20 tokens per second)
llama_print_timings:       total time =   17517.22 ms /   122 tokens
No. of rows:  19%| | 244/1258 [1:12:31<4:48:49, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.37 ms /    37 runs   (    0.42
ms per token,  2407.91 tokens per second)
llama_print_timings: prompt eval time =   12986.41 ms /   116 tokens (  111.95
ms per token,     8.93 tokens per second)
llama_print_timings:        eval time =    9738.36 ms /    36 runs   (  270.51
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   22956.89 ms /   152 tokens
No. of rows:  19%| | 245/1258 [1:12:54<5:18:17, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.66 ms /    24 runs   (    0.40
ms per token,  2484.99 tokens per second)
llama_print_timings: prompt eval time =   11283.49 ms /   103 tokens (  109.55
ms per token,     9.13 tokens per second)
llama_print_timings:        eval time =    6943.98 ms /    23 runs   (  301.91
ms per token,     3.31 tokens per second)
llama_print_timings:       total time =   18376.62 ms /   126 tokens
```

```
No. of rows:  20%| | 246/1258 [1:13:12<5:15:33, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.46 ms /    27 runs   (    0.39
ms per token,  2581.26 tokens per second)
llama_print_timings: prompt eval time =   11784.37 ms /   107 tokens (  110.13
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    5775.09 ms /    26 runs   (  222.12
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   17722.86 ms /   133 tokens
No. of rows:  20%| | 247/1258 [1:13:30<5:10:20, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.50 ms /    19 runs   (    0.39
ms per token,  2534.01 tokens per second)
llama_print_timings: prompt eval time =    8932.93 ms /    81 tokens (  110.28
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    5714.01 ms /    18 runs   (  317.45
ms per token,     3.15 tokens per second)
llama_print_timings:       total time =   14761.61 ms /    99 tokens
No. of rows:  20%| | 248/1258 [1:13:45<4:51:32, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.58 ms /    32 runs   (    0.39
ms per token,  2542.91 tokens per second)
llama_print_timings: prompt eval time =   13354.27 ms /   118 tokens (  113.17
ms per token,     8.84 tokens per second)
llama_print_timings:        eval time =    8893.77 ms /    31 runs   (  286.90
ms per token,     3.49 tokens per second)
llama_print_timings:       total time =   22444.91 ms /   149 tokens
No. of rows:  20%| | 249/1258 [1:14:07<5:17:12, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.90 ms /    22 runs   (    0.40
ms per token,  2472.74 tokens per second)
llama_print_timings: prompt eval time =   10305.54 ms /    90 tokens (  114.51
ms per token,     8.73 tokens per second)
llama_print_timings:        eval time =    4762.78 ms /    21 runs   (  226.80
ms per token,     4.41 tokens per second)
llama_print_timings:       total time =   15204.48 ms /   111 tokens
No. of rows:  20%| | 250/1258 [1:14:23<4:58:29, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       5.68 ms /    15 runs   (    0.38
ms per token,  2639.45 tokens per second)
llama_print_timings: prompt eval time =    8873.11 ms /    80 tokens (  110.91
ms per token,     9.02 tokens per second)
llama_print_timings:        eval time =    3090.39 ms /    14 runs   (  220.74
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   12054.06 ms /    94 tokens
No. of rows:  20%| | 251/1258 [1:14:35<4:29:25, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.09 ms /    28 runs   (    0.40
ms per token,  2523.89 tokens per second)
llama_print_timings: prompt eval time =   15888.58 ms /   143 tokens (  111.11
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =    7909.45 ms /    27 runs   (  292.94
ms per token,     3.41 tokens per second)
llama_print_timings:       total time =   23970.41 ms /   170 tokens
No. of rows:  20%| | 252/1258 [1:14:59<5:09:04, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.06 ms /    42 runs   (    0.41
ms per token,  2462.19 tokens per second)
llama_print_timings: prompt eval time =   11453.14 ms /   103 tokens (  111.20
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    9157.86 ms /    41 runs   (  223.36
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   20867.25 ms /   144 tokens
No. of rows:  20%| | 253/1258 [1:15:20<5:20:59, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.65 ms /    23 runs   (    0.42
ms per token,  2383.42 tokens per second)
llama_print_timings: prompt eval time =    8951.94 ms /    80 tokens (  111.90
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =    4905.69 ms /    22 runs   (  222.99
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   14000.54 ms /   102 tokens
No. of rows:  20%| | 254/1258 [1:15:34<4:54:47, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.70 ms /    33 runs   (    0.38
ms per token,  2599.45 tokens per second)
llama_print_timings: prompt eval time =   14011.33 ms /   112 tokens (  125.10
ms per token,     7.99 tokens per second)
```

```
llama_print_timings:        eval time =    9011.18 ms /    32 runs   (  281.60
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   23227.26 ms /   144 tokens
No. of rows:  20%| | 255/1258 [1:15:57<5:22:42, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.53 ms /    31 runs   (    0.44
ms per token,  2292.05 tokens per second)
llama_print_timings: prompt eval time =   11493.01 ms /   103 tokens (  111.58
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =    6984.91 ms /    30 runs   (  232.83
ms per token,     4.29 tokens per second)
llama_print_timings:       total time =   18681.18 ms /   133 tokens
No. of rows:  20%| | 256/1258 [1:16:15<5:19:15, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.24 ms /    30 runs   (    0.41
ms per token,  2450.78 tokens per second)
llama_print_timings: prompt eval time =   11549.12 ms /   105 tokens (  109.99
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =    6515.92 ms /    29 runs   (  224.69
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   18250.46 ms /   134 tokens
No. of rows:  20%| | 257/1258 [1:16:34<5:14:40, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.91 ms /    33 runs   (    0.42
ms per token,  2372.74 tokens per second)
llama_print_timings: prompt eval time =   13016.60 ms /   117 tokens (  111.25
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    7261.96 ms /    32 runs   (  226.94
ms per token,     4.41 tokens per second)
llama_print_timings:       total time =   20492.49 ms /   149 tokens
No. of rows:  21%| | 258/1258 [1:16:54<5:22:34, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.42 ms /    42 runs   (    0.39
ms per token,  2557.54 tokens per second)
llama_print_timings: prompt eval time =   15591.09 ms /   127 tokens (  122.76
ms per token,     8.15 tokens per second)
llama_print_timings:        eval time =    9219.52 ms /    41 runs   (  224.87
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   25070.94 ms /   168 tokens
No. of rows:  21%| | 259/1258 [1:17:19<5:50:47, 21Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.71 ms /    21 runs   (    0.41
ms per token,  2412.13 tokens per second)
llama_print_timings: prompt eval time =   10681.11 ms /    82 tokens (  130.26
ms per token,     7.68 tokens per second)
llama_print_timings:        eval time =    5091.27 ms /    20 runs   (  254.56
ms per token,     3.93 tokens per second)
llama_print_timings:       total time =   15901.57 ms /   102 tokens
No. of rows:  21%| | 260/1258 [1:17:35<5:24:43, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.58 ms /    20 runs   (    0.38
ms per token,  2638.17 tokens per second)
llama_print_timings: prompt eval time =    8997.34 ms /    79 tokens (  113.89
ms per token,     8.78 tokens per second)
llama_print_timings:        eval time =    4221.84 ms /    19 runs   (  222.20
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   13340.93 ms /    98 tokens
No. of rows:  21%| | 261/1258 [1:17:49<4:53:39, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.58 ms /    26 runs   (    0.41
ms per token,  2458.16 tokens per second)
llama_print_timings: prompt eval time =   11007.15 ms /    99 tokens (  111.18
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    5624.05 ms /    25 runs   (  224.96
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   16793.36 ms /   124 tokens
No. of rows:  21%| | 262/1258 [1:18:05<4:48:58, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.08 ms /    37 runs   (    0.41
ms per token,  2454.07 tokens per second)
llama_print_timings: prompt eval time =   15074.02 ms /   122 tokens (  123.56
ms per token,     8.09 tokens per second)
llama_print_timings:        eval time =    8115.45 ms /    36 runs   (  225.43
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   23424.78 ms /   158 tokens
No. of rows:  21%| | 263/1258 [1:18:29<5:18:39, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.74 ms /    27 runs   (    0.40
```

```
ms per token,  2514.90 tokens per second)
llama_print_timings: prompt eval time =    9397.53 ms /    81 tokens (  116.02
ms per token,     8.62 tokens per second)
llama_print_timings:        eval time =    6250.24 ms /    26 runs   (  240.39
ms per token,     4.16 tokens per second)
llama_print_timings:       total time =   15814.80 ms /   107 tokens
No. of rows:  21%| | 264/1258 [1:18:45<5:01:31, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.75 ms /    33 runs   (    0.39
ms per token,  2589.05 tokens per second)
llama_print_timings: prompt eval time =   12275.53 ms /   110 tokens (  111.60
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =    9028.56 ms /    32 runs   (  282.14
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   21508.72 ms /   142 tokens
No. of rows:  21%| | 265/1258 [1:19:06<5:17:41, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.95 ms /    24 runs   (    0.41
ms per token,  2412.30 tokens per second)
llama_print_timings: prompt eval time =   11973.63 ms /   109 tokens (  109.85
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    5746.14 ms /    23 runs   (  249.83
ms per token,     4.00 tokens per second)
llama_print_timings:       total time =   17868.17 ms /   132 tokens
No. of rows:  21%| | 266/1258 [1:19:24<5:10:48, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.73 ms /    19 runs   (    0.41
ms per token,  2456.68 tokens per second)
llama_print_timings: prompt eval time =    9132.41 ms /    83 tokens (  110.03
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =    4009.73 ms /    18 runs   (  222.76
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   13258.59 ms /   101 tokens
No. of rows:  21%| | 267/1258 [1:19:37<4:43:02, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.32 ms /    31 runs   (    0.40
ms per token,  2515.42 tokens per second)
llama_print_timings: prompt eval time =   11457.35 ms /    93 tokens (  123.20
ms per token,     8.12 tokens per second)
llama_print_timings:        eval time =    8167.33 ms /    30 runs   (  272.24
```

```
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   19813.34 ms /   123 tokens
No. of rows:  21%| | 268/1258 [1:19:57<4:56:05, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.35 ms /    21 runs   (    0.40
ms per token,  2515.87 tokens per second)
llama_print_timings: prompt eval time =    8866.04 ms /    80 tokens (  110.83
ms per token,     9.02 tokens per second)
llama_print_timings:        eval time =    4762.73 ms /    20 runs   (  238.14
ms per token,     4.20 tokens per second)
llama_print_timings:       total time =   13756.97 ms /   100 tokens
No. of rows:  21%| | 269/1258 [1:20:11<4:35:04, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.21 ms /    20 runs   (    0.41
ms per token,  2436.05 tokens per second)
llama_print_timings: prompt eval time =    8902.41 ms /    79 tokens (  112.69
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    4279.72 ms /    19 runs   (  225.25
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   13307.63 ms /    98 tokens
No. of rows:  21%| | 270/1258 [1:20:24<4:18:10, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.91 ms /    29 runs   (    0.41
ms per token,  2435.75 tokens per second)
llama_print_timings: prompt eval time =    9864.88 ms /    87 tokens (  113.39
ms per token,     8.82 tokens per second)
llama_print_timings:        eval time =    6446.42 ms /    28 runs   (  230.23
ms per token,     4.34 tokens per second)
llama_print_timings:       total time =   16490.39 ms /   115 tokens
No. of rows:  22%| | 271/1258 [1:20:41<4:21:54, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.26 ms /    23 runs   (    0.40
ms per token,  2485.14 tokens per second)
llama_print_timings: prompt eval time =    9010.60 ms /    78 tokens (  115.52
ms per token,     8.66 tokens per second)
llama_print_timings:        eval time =    5519.05 ms /    22 runs   (  250.87
ms per token,     3.99 tokens per second)
llama_print_timings:       total time =   14673.18 ms /   100 tokens
No. of rows:  22%| | 272/1258 [1:20:55<4:15:33, 15Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.82 ms /    29 runs   (    0.41
ms per token,  2454.09 tokens per second)
llama_print_timings: prompt eval time =   10185.10 ms /    91 tokens (  111.92
ms per token,     8.93 tokens per second)
llama_print_timings:        eval time =    7889.04 ms /    28 runs   (  281.75
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   18254.29 ms /   119 tokens
No. of rows:  22%| | 273/1258 [1:21:14<4:28:40, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.39 ms /    20 runs   (    0.42
ms per token,  2382.37 tokens per second)
llama_print_timings: prompt eval time =    8878.12 ms /    80 tokens (  110.98
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    4230.40 ms /    19 runs   (  222.65
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   13234.92 ms /    99 tokens
No. of rows:  22%| | 274/1258 [1:21:27<4:13:01, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.76 ms /    41 runs   (    0.41
ms per token,  2445.86 tokens per second)
llama_print_timings: prompt eval time =   14039.11 ms /   118 tokens (  118.98
ms per token,     8.41 tokens per second)
llama_print_timings:        eval time =    9210.75 ms /    40 runs   (  230.27
ms per token,     4.34 tokens per second)
llama_print_timings:       total time =   23510.35 ms /   158 tokens
No. of rows:  22%| | 275/1258 [1:21:50<4:52:29, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.61 ms /    41 runs   (    0.41
ms per token,  2469.14 tokens per second)
llama_print_timings: prompt eval time =   15434.38 ms /   127 tokens (  121.53
ms per token,     8.23 tokens per second)
llama_print_timings:        eval time =    9028.48 ms /    40 runs   (  225.71
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   24715.75 ms /   167 tokens
No. of rows:  22%| | 276/1258 [1:22:15<5:25:55, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.28 ms /    26 runs   (    0.43
ms per token,  2305.78 tokens per second)
```

```
llama_print_timings: prompt eval time =   11298.31 ms /    95 tokens (  118.93
ms per token,     8.41 tokens per second)
llama_print_timings:        eval time =    5658.86 ms /    25 runs   (  226.35
ms per token,     4.42 tokens per second)
llama_print_timings:       total time =   17120.79 ms /   120 tokens
No. of rows:  22%| | 277/1258 [1:22:32<5:11:54, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.81 ms /    43 runs   (    0.39
ms per token,  2557.39 tokens per second)
llama_print_timings: prompt eval time =   12876.86 ms /   115 tokens (  111.97
ms per token,     8.93 tokens per second)
llama_print_timings:        eval time =    9588.41 ms /    42 runs   (  228.30
ms per token,     4.38 tokens per second)
llama_print_timings:       total time =   22733.96 ms /   157 tokens
No. of rows:  22%| | 278/1258 [1:22:55<5:29:34, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.21 ms /    24 runs   (    0.38
ms per token,  2605.86 tokens per second)
llama_print_timings: prompt eval time =   11506.02 ms /    89 tokens (  129.28
ms per token,     7.74 tokens per second)
llama_print_timings:        eval time =    5149.71 ms /    23 runs   (  223.90
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   16803.72 ms /   112 tokens
No. of rows:  22%| | 279/1258 [1:23:12<5:12:46, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.63 ms /    22 runs   (    0.39
ms per token,  2548.95 tokens per second)
llama_print_timings: prompt eval time =    9733.04 ms /    87 tokens (  111.87
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =    4657.68 ms /    21 runs   (  221.79
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   14526.67 ms /   108 tokens
No. of rows:  22%| | 280/1258 [1:23:26<4:49:45, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.45 ms /    32 runs   (    0.42
ms per token,  2379.71 tokens per second)
llama_print_timings: prompt eval time =   10472.90 ms /    96 tokens (  109.09
ms per token,     9.17 tokens per second)
llama_print_timings:        eval time =    6966.09 ms /    31 runs   (  224.71
ms per token,     4.45 tokens per second)
```

```
llama_print_timings:       total time =   17637.64 ms /   127 tokens
No. of rows:  22%| | 281/1258 [1:23:44<4:48:49, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.76 ms /    28 runs   (    0.38
ms per token,  2601.51 tokens per second)
llama_print_timings: prompt eval time =   12682.69 ms /   101 tokens (  125.57
ms per token,     7.96 tokens per second)
llama_print_timings:        eval time =    6048.50 ms /    27 runs   (  224.02
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   18901.12 ms /   128 tokens
No. of rows:  22%| | 282/1258 [1:24:03<4:54:12, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.19 ms /    20 runs   (    0.41
ms per token,  2443.49 tokens per second)
llama_print_timings: prompt eval time =   11528.79 ms /    90 tokens (  128.10
ms per token,     7.81 tokens per second)
llama_print_timings:        eval time =    4519.01 ms /    19 runs   (  237.84
ms per token,     4.20 tokens per second)
llama_print_timings:       total time =   16175.62 ms /   109 tokens
No. of rows:  22%| | 283/1258 [1:24:19<4:44:41, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.51 ms /    50 runs   (    0.39
ms per token,  2563.05 tokens per second)
llama_print_timings: prompt eval time =   14495.41 ms /   131 tokens (  110.65
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =   11182.25 ms /    49 runs   (  228.21
ms per token,     4.38 tokens per second)
llama_print_timings:       total time =   25988.55 ms /   180 tokens
No. of rows:  23%| | 284/1258 [1:24:45<5:25:38, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.42 ms /    23 runs   (    0.41
ms per token,  2441.10 tokens per second)
llama_print_timings: prompt eval time =    8964.32 ms /    78 tokens (  114.93
ms per token,     8.70 tokens per second)
llama_print_timings:        eval time =    4924.62 ms /    22 runs   (  223.85
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   14030.72 ms /   100 tokens
No. of rows:  23%| | 285/1258 [1:24:59<4:55:58, 18Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.73 ms /    31 runs   (    0.38
ms per token,  2643.70 tokens per second)
llama_print_timings: prompt eval time =   11457.47 ms /    91 tokens (  125.91
ms per token,     7.94 tokens per second)
llama_print_timings:        eval time =    8443.89 ms /    30 runs   (  281.46
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   20092.33 ms /   121 tokens
No. of rows:  23%| | 286/1258 [1:25:19<5:04:43, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.07 ms /    20 runs   (    0.40
ms per token,  2478.62 tokens per second)
llama_print_timings: prompt eval time =    8749.43 ms /    79 tokens (  110.75
ms per token,     9.03 tokens per second)
llama_print_timings:        eval time =    4631.90 ms /    19 runs   (  243.78
ms per token,     4.10 tokens per second)
llama_print_timings:       total time =   13502.84 ms /    98 tokens
No. of rows:  23%| | 287/1258 [1:25:33<4:38:39, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.28 ms /    25 runs   (    0.41
ms per token,  2431.43 tokens per second)
llama_print_timings: prompt eval time =   11145.00 ms /   100 tokens (  111.45
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    7031.35 ms /    24 runs   (  292.97
ms per token,     3.41 tokens per second)
llama_print_timings:       total time =   18331.97 ms /   124 tokens
No. of rows:  23%| | 288/1258 [1:25:51<4:43:50, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.37 ms /    27 runs   (    0.38
ms per token,  2604.17 tokens per second)
llama_print_timings: prompt eval time =   10692.40 ms /    96 tokens (  111.38
ms per token,     8.98 tokens per second)
llama_print_timings:        eval time =    5789.55 ms /    26 runs   (  222.68
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   16647.52 ms /   122 tokens
No. of rows:  23%| | 289/1258 [1:26:08<4:39:09, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.23 ms /    20 runs   (    0.46
ms per token,  2167.55 tokens per second)
llama_print_timings: prompt eval time =    8980.52 ms /    81 tokens (  110.87
```

```
ms per token,     9.02 tokens per second)
llama_print_timings:        eval time =    4497.66 ms /    19 runs   (  236.72
ms per token,     4.22 tokens per second)
llama_print_timings:       total time =   13612.81 ms /   100 tokens
No. of rows:  23%| | 290/1258 [1:26:21<4:21:05, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.95 ms /    22 runs   (    0.41
ms per token,  2458.65 tokens per second)
llama_print_timings: prompt eval time =    9205.87 ms /    81 tokens (  113.65
ms per token,     8.80 tokens per second)
llama_print_timings:        eval time =    4781.34 ms /    21 runs   (  227.68
ms per token,     4.39 tokens per second)
llama_print_timings:       total time =   14123.82 ms /   102 tokens
No. of rows:  23%| | 291/1258 [1:26:35<4:10:57, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.36 ms /    26 runs   (    0.40
ms per token,  2510.14 tokens per second)
llama_print_timings: prompt eval time =   11456.09 ms /   104 tokens (  110.15
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    5624.29 ms /    25 runs   (  224.97
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   17243.61 ms /   129 tokens
No. of rows:  23%| | 292/1258 [1:26:53<4:18:45, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.83 ms /    20 runs   (    0.39
ms per token,  2555.58 tokens per second)
llama_print_timings: prompt eval time =   11042.00 ms /    86 tokens (  128.40
ms per token,     7.79 tokens per second)
llama_print_timings:        eval time =    4212.08 ms /    19 runs   (  221.69
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   15376.93 ms /   105 tokens
No. of rows:  23%| | 293/1258 [1:27:08<4:15:10, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.44 ms /    33 runs   (    0.41
ms per token,  2454.99 tokens per second)
llama_print_timings: prompt eval time =   12276.09 ms /   111 tokens (  110.60
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =    7207.95 ms /    32 runs   (  225.25
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   19688.33 ms /   143 tokens
```

```
No. of rows:  23%| | 294/1258 [1:27:28<4:33:25, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.57 ms /    32 runs   (    0.39
ms per token,  2545.34 tokens per second)
llama_print_timings: prompt eval time =   12783.10 ms /   102 tokens (  125.32
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =    7232.85 ms /    31 runs   (  233.32
ms per token,     4.29 tokens per second)
llama_print_timings:       total time =   20212.42 ms /   133 tokens
No. of rows:  23%| | 295/1258 [1:27:48<4:48:30, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.18 ms /    24 runs   (    0.38
ms per token,  2614.95 tokens per second)
llama_print_timings: prompt eval time =   10758.01 ms /    96 tokens (  112.06
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    5136.36 ms /    23 runs   (  223.32
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   16039.59 ms /   119 tokens
No. of rows:  24%| | 296/1258 [1:28:04<4:38:55, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.47 ms /    28 runs   (    0.45
ms per token,  2245.03 tokens per second)
llama_print_timings: prompt eval time =   12065.83 ms /    95 tokens (  127.01
ms per token,     7.87 tokens per second)
llama_print_timings:        eval time =    6307.89 ms /    27 runs   (  233.63
ms per token,     4.28 tokens per second)
llama_print_timings:       total time =   18550.42 ms /   122 tokens
No. of rows:  24%| | 297/1258 [1:28:23<4:44:16, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.14 ms /    33 runs   (    0.40
ms per token,  2511.80 tokens per second)
llama_print_timings: prompt eval time =   10706.04 ms /    96 tokens (  111.52
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    7278.31 ms /    32 runs   (  227.45
ms per token,     4.40 tokens per second)
llama_print_timings:       total time =   18185.92 ms /   128 tokens
No. of rows:  24%| | 298/1258 [1:28:41<4:46:04, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       8.98 ms /    23 runs   (    0.39
ms per token,  2561.25 tokens per second)
llama_print_timings: prompt eval time =   10117.33 ms /    91 tokens (  111.18
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    4885.00 ms /    22 runs   (  222.05
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   15142.87 ms /   113 tokens
No. of rows:  24%| | 299/1258 [1:28:56<4:32:43, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.25 ms /    21 runs   (    0.39
ms per token,  2545.15 tokens per second)
llama_print_timings: prompt eval time =    9957.01 ms /    89 tokens (  111.88
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =    4449.89 ms /    20 runs   (  222.49
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   14539.08 ms /   109 tokens
No. of rows:  24%| | 300/1258 [1:29:11<4:20:22, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.02 ms /    25 runs   (    0.40
ms per token,  2495.76 tokens per second)
llama_print_timings: prompt eval time =   10752.36 ms /    82 tokens (  131.13
ms per token,     7.63 tokens per second)
llama_print_timings:        eval time =    5341.60 ms /    24 runs   (  222.57
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   16248.26 ms /   106 tokens
No. of rows:  24%| | 301/1258 [1:29:27<4:19:49, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.05 ms /    25 runs   (    0.68
ms per token,  1466.62 tokens per second)
llama_print_timings: prompt eval time =   11871.75 ms /    94 tokens (  126.30
ms per token,     7.92 tokens per second)
llama_print_timings:        eval time =    8360.01 ms /    24 runs   (  348.33
ms per token,     2.87 tokens per second)
llama_print_timings:       total time =   20490.35 ms /   118 tokens
No. of rows:  24%| | 302/1258 [1:29:47<4:39:44, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.86 ms /    30 runs   (    0.36
ms per token,  2761.67 tokens per second)
llama_print_timings: prompt eval time =   21332.44 ms /   125 tokens (  170.66
ms per token,     5.86 tokens per second)
```

```
llama_print_timings:        eval time = 10803149.23 ms /    29 runs   (372522.39
ms per token,     0.00 tokens per second)
llama_print_timings:       total time = 10824643.44 ms /   154 tokens
No. of rows:  24%| | 303/1258 [4:30:12<864:43:11, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.38 ms /    40 runs   (    0.31
ms per token,  3229.71 tokens per second)
llama_print_timings: prompt eval time = 20253732.13 ms /   133 tokens (152283.70
ms per token,     0.01 tokens per second)
llama_print_timings:        eval time =   18428.96 ms /    39 runs   (  472.54
ms per token,     2.12 tokens per second)
llama_print_timings:       total time = 20272358.16 ms /   172 tokens
No. of rows:  24%| | 304/1258 [10:07:56<2215:41:47Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.05 ms /    50 runs   (    0.36
ms per token,  2770.85 tokens per second)
llama_print_timings: prompt eval time =   11274.99 ms /   112 tokens (  100.67
ms per token,     9.93 tokens per second)
llama_print_timings:        eval time =   10101.78 ms /    49 runs   (  206.16
ms per token,     4.85 tokens per second)
llama_print_timings:       total time =   21647.55 ms /   161 tokens
No. of rows:  24%| | 305/1258 [10:08:18<1551:04:54Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.63 ms /    24 runs   (    0.44
ms per token,  2258.40 tokens per second)
llama_print_timings: prompt eval time =   10561.14 ms /    93 tokens (  113.56
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    6233.28 ms /    23 runs   (  271.01
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   16958.29 ms /   116 tokens
No. of rows:  24%| | 306/1258 [10:08:35<1085:57:50Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.65 ms /    22 runs   (    0.35
ms per token,  2874.31 tokens per second)
llama_print_timings: prompt eval time =    7612.67 ms /    79 tokens (   96.36
ms per token,    10.38 tokens per second)
llama_print_timings:        eval time =    4265.94 ms /    21 runs   (  203.14
ms per token,     4.92 tokens per second)
llama_print_timings:       total time =   11996.41 ms /   100 tokens
No. of rows:  24%| | 307/1258 [10:08:47<760:19:34,Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.63 ms /    28 runs   (    0.34
ms per token,  2906.37 tokens per second)
llama_print_timings: prompt eval time =    8536.24 ms /    98 tokens (   87.10
ms per token,    11.48 tokens per second)
llama_print_timings:        eval time =    5361.27 ms /    27 runs   (  198.57
ms per token,     5.04 tokens per second)
llama_print_timings:       total time =   14042.82 ms /   125 tokens
No. of rows:  24%| | 308/1258 [10:09:01<532:46:55,Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.07 ms /    18 runs   (    0.34
ms per token,  2963.45 tokens per second)
llama_print_timings: prompt eval time =    6631.25 ms /    79 tokens (   83.94
ms per token,    11.91 tokens per second)
llama_print_timings:        eval time =    3371.35 ms /    17 runs   (  198.31
ms per token,     5.04 tokens per second)
llama_print_timings:       total time =   10096.19 ms /    96 tokens
No. of rows:  25%| | 309/1258 [10:09:11<373:21:14,Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.48 ms /    29 runs   (    0.40
ms per token,  2525.69 tokens per second)
llama_print_timings: prompt eval time =    9664.90 ms /    98 tokens (   98.62
ms per token,    10.14 tokens per second)
llama_print_timings:        eval time =    5883.04 ms /    28 runs   (  210.11
ms per token,     4.76 tokens per second)
llama_print_timings:       total time =   15702.66 ms /   126 tokens
No. of rows:  25%| | 310/1258 [10:09:27<262:18:48,Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.96 ms /    23 runs   (    0.39
ms per token,  2567.82 tokens per second)
llama_print_timings: prompt eval time =    8148.70 ms /    93 tokens (   87.62
ms per token,    11.41 tokens per second)
llama_print_timings:        eval time =    4513.62 ms /    22 runs   (  205.16
ms per token,     4.87 tokens per second)
llama_print_timings:       total time =   12787.85 ms /   115 tokens
No. of rows:  25%| | 311/1258 [10:09:40<184:26:07,Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.86 ms /    24 runs   (    0.33
```

```
ms per token,  3051.88 tokens per second)
llama_print_timings: prompt eval time =    8265.60 ms /    97 tokens (   85.21
ms per token,    11.74 tokens per second)
llama_print_timings:        eval time =    4629.05 ms /    23 runs   (  201.26
ms per token,     4.97 tokens per second)
llama_print_timings:       total time =   13019.63 ms /   120 tokens
No. of rows:  25%| | 312/1258 [10:09:53<129:59:43,Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.48 ms /    25 runs   (    0.34
ms per token,  2948.11 tokens per second)
llama_print_timings: prompt eval time =    8742.24 ms /   106 tokens (   82.47
ms per token,    12.13 tokens per second)
llama_print_timings:        eval time =    4783.88 ms /    24 runs   (  199.33
ms per token,     5.02 tokens per second)
llama_print_timings:       total time =   13655.77 ms /   130 tokens
No. of rows:  25%| | 313/1258 [10:10:06<91:58:35, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.78 ms /    38 runs   (    0.28
ms per token,  3525.70 tokens per second)
llama_print_timings: prompt eval time =   12063.64 ms /   127 tokens (   94.99
ms per token,    10.53 tokens per second)
llama_print_timings:        eval time =    7895.23 ms /    37 runs   (  213.38
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   20136.40 ms /   164 tokens
No. of rows:  25%| | 314/1258 [10:10:26<65:54:00, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.39 ms /    22 runs   (    0.29
ms per token,  3442.34 tokens per second)
llama_print_timings: prompt eval time =    7549.06 ms /    86 tokens (   87.78
ms per token,    11.39 tokens per second)
llama_print_timings:        eval time =    4216.57 ms /    21 runs   (  200.79
ms per token,     4.98 tokens per second)
llama_print_timings:       total time =   11866.21 ms /   107 tokens
No. of rows:  25%| | 315/1258 [10:10:38<47:00:52, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.08 ms /    26 runs   (    0.27
ms per token,  3673.87 tokens per second)
llama_print_timings: prompt eval time =    8390.69 ms /    93 tokens (   90.22
ms per token,    11.08 tokens per second)
llama_print_timings:        eval time =    5021.93 ms /    25 runs   (  200.88
```

```
ms per token,     4.98 tokens per second)
llama_print_timings:       total time =   13530.80 ms /   118 tokens
No. of rows:  25%| | 316/1258 [10:10:52<33:56:16, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.31 ms /    22 runs   (    0.29
ms per token,  3484.87 tokens per second)
llama_print_timings: prompt eval time =    8200.55 ms /    81 tokens (  101.24
ms per token,     9.88 tokens per second)
llama_print_timings:        eval time =    4256.93 ms /    21 runs   (  202.71
ms per token,     4.93 tokens per second)
llama_print_timings:       total time =   12560.77 ms /   102 tokens
No. of rows:  25%| | 317/1258 [10:11:04<24:43:00, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.76 ms /    25 runs   (    0.27
ms per token,  3699.32 tokens per second)
llama_print_timings: prompt eval time =    9125.91 ms /   106 tokens (   86.09
ms per token,    11.62 tokens per second)
llama_print_timings:        eval time =    4820.60 ms /    24 runs   (  200.86
ms per token,     4.98 tokens per second)
llama_print_timings:       total time =   14064.06 ms /   130 tokens
No. of rows:  25%| | 318/1258 [10:11:19<18:23:07, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.01 ms /    24 runs   (    0.29
ms per token,  3423.19 tokens per second)
llama_print_timings: prompt eval time =    9263.96 ms /    90 tokens (  102.93
ms per token,     9.72 tokens per second)
llama_print_timings:        eval time =    4842.17 ms /    23 runs   (  210.53
ms per token,     4.75 tokens per second)
llama_print_timings:       total time =   14227.39 ms /   113 tokens
No. of rows:  25%| | 319/1258 [10:11:33<13:58:11, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.76 ms /    22 runs   (    0.31
ms per token,  3256.36 tokens per second)
llama_print_timings: prompt eval time =    8254.35 ms /    89 tokens (   92.75
ms per token,    10.78 tokens per second)
llama_print_timings:        eval time =    4610.61 ms /    21 runs   (  219.55
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   12974.26 ms /   110 tokens
No. of rows:  25%| | 320/1258 [10:11:46<10:46:59, Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.32 ms /    20 runs   (    0.52
ms per token,  1937.80 tokens per second)
llama_print_timings: prompt eval time =    7957.39 ms /    83 tokens (   95.87
ms per token,    10.43 tokens per second)
llama_print_timings:        eval time =    5160.85 ms /    19 runs   (  271.62
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   13271.57 ms /   102 tokens
No. of rows:  26%| | 321/1258 [10:11:59<8:34:33, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.81 ms /    24 runs   (    0.33
ms per token,  3072.20 tokens per second)
llama_print_timings: prompt eval time =    8502.56 ms /    92 tokens (   92.42
ms per token,    10.82 tokens per second)
llama_print_timings:        eval time =    5181.66 ms /    23 runs   (  225.29
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   13807.35 ms /   115 tokens
No. of rows:  26%| | 322/1258 [10:12:13<7:04:27, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.50 ms /    18 runs   (    0.36
ms per token,  2769.66 tokens per second)
llama_print_timings: prompt eval time =    8144.12 ms /    77 tokens (  105.77
ms per token,     9.45 tokens per second)
llama_print_timings:        eval time =    3849.51 ms /    17 runs   (  226.44
ms per token,     4.42 tokens per second)
llama_print_timings:       total time =   12096.02 ms /    94 tokens
No. of rows:  26%| | 323/1258 [10:12:25<5:53:23, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.56 ms /    18 runs   (    0.36
ms per token,  2744.32 tokens per second)
llama_print_timings: prompt eval time =    7751.30 ms /    81 tokens (   95.70
ms per token,    10.45 tokens per second)
llama_print_timings:        eval time =    3696.65 ms /    17 runs   (  217.45
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   11548.15 ms /    98 tokens
No. of rows:  26%| | 324/1258 [10:12:37<5:01:07, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.32 ms /    20 runs   (    0.37
ms per token,  2730.75 tokens per second)
```

```
llama_print_timings: prompt eval time =    8598.19 ms /    82 tokens (  104.86
ms per token,     9.54 tokens per second)
llama_print_timings:        eval time =    4071.77 ms /    19 runs   (  214.30
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   12787.14 ms /   101 tokens
No. of rows:  26%| | 325/1258 [10:12:49<4:30:12, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.84 ms /    29 runs   (    0.37
ms per token,  2676.02 tokens per second)
llama_print_timings: prompt eval time =   12397.60 ms /   105 tokens (  118.07
ms per token,     8.47 tokens per second)
llama_print_timings:        eval time =    5997.50 ms /    28 runs   (  214.20
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   18564.04 ms /   133 tokens
No. of rows:  26%| | 326/1258 [10:13:08<4:35:31, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.51 ms /    22 runs   (    0.34
ms per token,  2928.26 tokens per second)
llama_print_timings: prompt eval time =    9582.89 ms /    82 tokens (  116.86
ms per token,     8.56 tokens per second)
llama_print_timings:        eval time =    4344.98 ms /    21 runs   (  206.90
ms per token,     4.83 tokens per second)
llama_print_timings:       total time =   14045.21 ms /   103 tokens
No. of rows:  26%| | 327/1258 [10:13:22<4:18:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.26 ms /    30 runs   (    0.48
ms per token,  2103.93 tokens per second)
llama_print_timings: prompt eval time =   14861.69 ms /   119 tokens (  124.89
ms per token,     8.01 tokens per second)
llama_print_timings:        eval time =    7561.27 ms /    29 runs   (  260.73
ms per token,     3.84 tokens per second)
llama_print_timings:       total time =   22650.51 ms /   148 tokens
No. of rows:  26%| | 328/1258 [10:13:45<4:45:46, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      21.09 ms /    48 runs   (    0.44
ms per token,  2275.96 tokens per second)
llama_print_timings: prompt eval time =   11126.48 ms /   110 tokens (  101.15
ms per token,     9.89 tokens per second)
llama_print_timings:        eval time =   11017.11 ms /    47 runs   (  234.41
ms per token,     4.27 tokens per second)
```

```
llama_print_timings:       total time =   22439.17 ms /   157 tokens
No. of rows:  26%| | 329/1258 [10:14:07<5:04:07, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.36 ms /    22 runs   (    0.38
ms per token,  2631.89 tokens per second)
llama_print_timings: prompt eval time =    9150.11 ms /    94 tokens (   97.34
ms per token,    10.27 tokens per second)
llama_print_timings:        eval time =    4547.16 ms /    21 runs   (  216.53
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   13825.59 ms /   115 tokens
No. of rows:  26%| | 330/1258 [10:14:21<4:36:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.23 ms /    20 runs   (    0.51
ms per token,  1954.08 tokens per second)
llama_print_timings: prompt eval time =    9448.54 ms /    77 tokens (  122.71
ms per token,     8.15 tokens per second)
llama_print_timings:        eval time =    4476.98 ms /    19 runs   (  235.63
ms per token,     4.24 tokens per second)
llama_print_timings:       total time =   14065.55 ms /    96 tokens
No. of rows:  26%| | 331/1258 [10:14:35<4:18:47, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.73 ms /    18 runs   (    0.43
ms per token,  2330.10 tokens per second)
llama_print_timings: prompt eval time =    8611.23 ms /    78 tokens (  110.40
ms per token,     9.06 tokens per second)
llama_print_timings:        eval time =    4149.59 ms /    17 runs   (  244.09
ms per token,     4.10 tokens per second)
llama_print_timings:       total time =   12880.45 ms /    95 tokens
No. of rows:  26%| | 332/1258 [10:14:48<4:00:41, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.61 ms /    16 runs   (    0.48
ms per token,  2102.77 tokens per second)
llama_print_timings: prompt eval time =    9407.58 ms /    81 tokens (  116.14
ms per token,     8.61 tokens per second)
llama_print_timings:        eval time =    3674.06 ms /    15 runs   (  244.94
ms per token,     4.08 tokens per second)
llama_print_timings:       total time =   13196.29 ms /    96 tokens
No. of rows:  26%| | 333/1258 [10:15:01<3:49:20, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.17 ms /    25 runs   (    0.37
ms per token,  2726.88 tokens per second)
llama_print_timings: prompt eval time =   10847.46 ms /    87 tokens (  124.68
ms per token,     8.02 tokens per second)
llama_print_timings:        eval time =    5035.45 ms /    24 runs   (  209.81
ms per token,     4.77 tokens per second)
llama_print_timings:       total time =   16023.71 ms /   111 tokens
No. of rows:  27%| | 334/1258 [10:15:17<3:54:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.94 ms /    28 runs   (    0.43
ms per token,  2344.27 tokens per second)
llama_print_timings: prompt eval time =   11609.26 ms /    87 tokens (  133.44
ms per token,     7.49 tokens per second)
llama_print_timings:        eval time =    6938.42 ms /    27 runs   (  256.98
ms per token,     3.89 tokens per second)
llama_print_timings:       total time =   18728.97 ms /   114 tokens
No. of rows:  27%| | 335/1258 [10:15:36<4:10:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.87 ms /    26 runs   (    0.38
ms per token,  2634.51 tokens per second)
llama_print_timings: prompt eval time =   11800.59 ms /   101 tokens (  116.84
ms per token,     8.56 tokens per second)
llama_print_timings:        eval time =    5554.26 ms /    25 runs   (  222.17
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   17514.13 ms /   126 tokens
No. of rows:  27%| | 336/1258 [10:15:53<4:15:52, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.78 ms /    35 runs   (    0.39
ms per token,  2540.65 tokens per second)
llama_print_timings: prompt eval time =   12454.94 ms /   112 tokens (  111.20
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    7678.32 ms /    34 runs   (  225.83
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   20341.66 ms /   146 tokens
No. of rows:  27%| | 337/1258 [10:16:14<4:32:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.62 ms /    22 runs   (    0.48
ms per token,  2070.98 tokens per second)
llama_print_timings: prompt eval time =   11820.02 ms /    93 tokens (  127.10
```

```
ms per token,     7.87 tokens per second)
llama_print_timings:        eval time =    6673.05 ms /    21 runs   (  317.76
ms per token,     3.15 tokens per second)
llama_print_timings:       total time =   18654.19 ms /   114 tokens
No. of rows:  27%| | 338/1258 [10:16:32<4:36:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.61 ms /    21 runs   (    0.36
ms per token,  2761.34 tokens per second)
llama_print_timings: prompt eval time =    9630.87 ms /    93 tokens (  103.56
ms per token,     9.66 tokens per second)
llama_print_timings:        eval time =    4367.71 ms /    20 runs   (  218.39
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   14120.56 ms /   113 tokens
No. of rows:  27%| | 339/1258 [10:16:46<4:18:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.75 ms /    25 runs   (    0.47
ms per token,  2127.84 tokens per second)
llama_print_timings: prompt eval time =   11021.67 ms /    99 tokens (  111.33
ms per token,     8.98 tokens per second)
llama_print_timings:        eval time =    5117.47 ms /    24 runs   (  213.23
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   16292.48 ms /   123 tokens
No. of rows:  27%| | 340/1258 [10:17:03<4:15:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.28 ms /    34 runs   (    0.51
ms per token,  1968.16 tokens per second)
llama_print_timings: prompt eval time =   11760.36 ms /   101 tokens (  116.44
ms per token,     8.59 tokens per second)
llama_print_timings:        eval time =   10744.66 ms /    33 runs   (  325.60
ms per token,     3.07 tokens per second)
llama_print_timings:       total time =   22758.15 ms /   134 tokens
No. of rows:  27%| | 341/1258 [10:17:26<4:43:00, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.65 ms /    25 runs   (    0.59
ms per token,  1706.14 tokens per second)
llama_print_timings: prompt eval time =   17188.42 ms /    94 tokens (  182.86
ms per token,     5.47 tokens per second)
llama_print_timings:        eval time =    7589.54 ms /    24 runs   (  316.23
ms per token,     3.16 tokens per second)
llama_print_timings:       total time =   24982.54 ms /   118 tokens
```

```
No. of rows:  27%| | 342/1258 [10:17:51<5:12:18, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.21 ms /    16 runs   (    0.45
ms per token,  2218.22 tokens per second)
llama_print_timings: prompt eval time =   11095.14 ms /    82 tokens (  135.31
ms per token,     7.39 tokens per second)
llama_print_timings:        eval time =    3507.34 ms /    15 runs   (  233.82
ms per token,     4.28 tokens per second)
llama_print_timings:       total time =   14714.51 ms /    97 tokens
No. of rows:  27%| | 343/1258 [10:18:05<4:45:47, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.41 ms /    32 runs   (    0.45
ms per token,  2220.83 tokens per second)
llama_print_timings: prompt eval time =   14750.93 ms /   108 tokens (  136.58
ms per token,     7.32 tokens per second)
llama_print_timings:        eval time =    7905.79 ms /    31 runs   (  255.03
ms per token,     3.92 tokens per second)
llama_print_timings:       total time =   22882.54 ms /   139 tokens
No. of rows:  27%| | 344/1258 [10:18:28<5:04:28, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.09 ms /    30 runs   (    0.47
ms per token,  2129.32 tokens per second)
llama_print_timings: prompt eval time =   13885.90 ms /   109 tokens (  127.39
ms per token,     7.85 tokens per second)
llama_print_timings:        eval time =    9749.78 ms /    29 runs   (  336.20
ms per token,     2.97 tokens per second)
llama_print_timings:       total time =   23857.84 ms /   138 tokens
No. of rows:  27%| | 345/1258 [10:18:52<5:21:49, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.35 ms /    15 runs   (    0.56
ms per token,  1795.55 tokens per second)
llama_print_timings: prompt eval time =   12206.07 ms /    85 tokens (  143.60
ms per token,     6.96 tokens per second)
llama_print_timings:        eval time =    5448.95 ms /    14 runs   (  389.21
ms per token,     2.57 tokens per second)
llama_print_timings:       total time =   17775.24 ms /    99 tokens
No. of rows:  28%| | 346/1258 [10:19:10<5:06:06, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =      13.28 ms /    27 runs   (    0.49
ms per token,  2033.44 tokens per second)
llama_print_timings: prompt eval time =   11574.06 ms /    91 tokens (  127.19
ms per token,     7.86 tokens per second)
llama_print_timings:        eval time =    8419.73 ms /    26 runs   (  323.84
ms per token,     3.09 tokens per second)
llama_print_timings:       total time =   20213.30 ms /   117 tokens
No. of rows:  28%| | 347/1258 [10:19:30<5:06:06, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.31 ms /    22 runs   (    0.51
ms per token,  1945.70 tokens per second)
llama_print_timings: prompt eval time =   10709.05 ms /    82 tokens (  130.60
ms per token,     7.66 tokens per second)
llama_print_timings:        eval time =    7501.40 ms /    21 runs   (  357.21
ms per token,     2.80 tokens per second)
llama_print_timings:       total time =   18391.24 ms /   103 tokens
No. of rows:  28%| | 348/1258 [10:19:48<4:57:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      21.49 ms /    30 runs   (    0.72
ms per token,  1396.19 tokens per second)
llama_print_timings: prompt eval time =   15441.36 ms /   121 tokens (  127.61
ms per token,     7.84 tokens per second)
llama_print_timings:        eval time =    9957.07 ms /    29 runs   (  343.35
ms per token,     2.91 tokens per second)
llama_print_timings:       total time =   25722.66 ms /   150 tokens
No. of rows:  28%| | 349/1258 [10:20:14<5:25:12, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.94 ms /    22 runs   (    0.45
ms per token,  2212.39 tokens per second)
llama_print_timings: prompt eval time =   10362.80 ms /    91 tokens (  113.88
ms per token,     8.78 tokens per second)
llama_print_timings:        eval time =    5405.53 ms /    21 runs   (  257.41
ms per token,     3.88 tokens per second)
llama_print_timings:       total time =   15931.03 ms /   112 tokens
No. of rows:  28%| | 350/1258 [10:20:30<4:59:45, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.46 ms /    25 runs   (    0.42
ms per token,  2389.14 tokens per second)
llama_print_timings: prompt eval time =   11066.94 ms /    95 tokens (  116.49
ms per token,     8.58 tokens per second)
```

```
llama_print_timings:        eval time =    5733.75 ms /    24 runs   (  238.91
ms per token,     4.19 tokens per second)
llama_print_timings:       total time =   16964.31 ms /   119 tokens
No. of rows:  28%| | 351/1258 [10:20:47<4:46:34, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.51 ms /    25 runs   (    0.46
ms per token,  2172.59 tokens per second)
llama_print_timings: prompt eval time =   10951.99 ms /    98 tokens (  111.76
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    6104.41 ms /    24 runs   (  254.35
ms per token,     3.93 tokens per second)
llama_print_timings:       total time =   17233.21 ms /   122 tokens
No. of rows:  28%| | 352/1258 [10:21:04<4:38:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.52 ms /    22 runs   (    0.43
ms per token,  2309.95 tokens per second)
llama_print_timings: prompt eval time =   11143.28 ms /    85 tokens (  131.10
ms per token,     7.63 tokens per second)
llama_print_timings:        eval time =    6582.69 ms /    21 runs   (  313.46
ms per token,     3.19 tokens per second)
llama_print_timings:       total time =   17871.36 ms /   106 tokens
No. of rows:  28%| | 353/1258 [10:21:22<4:35:37, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.28 ms /    22 runs   (    0.42
ms per token,  2369.92 tokens per second)
llama_print_timings: prompt eval time =   10440.87 ms /    88 tokens (  118.65
ms per token,     8.43 tokens per second)
llama_print_timings:        eval time =    4983.53 ms /    21 runs   (  237.31
ms per token,     4.21 tokens per second)
llama_print_timings:       total time =   15573.44 ms /   109 tokens
No. of rows:  28%| | 354/1258 [10:21:38<4:23:09, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.71 ms /    38 runs   (    0.44
ms per token,  2273.95 tokens per second)
llama_print_timings: prompt eval time =   12531.39 ms /   106 tokens (  118.22
ms per token,     8.46 tokens per second)
llama_print_timings:        eval time =    9043.86 ms /    37 runs   (  244.43
ms per token,     4.09 tokens per second)
llama_print_timings:       total time =   21832.62 ms /   143 tokens
No. of rows:  28%| | 355/1258 [10:22:00<4:42:37, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.94 ms /    24 runs   (    0.58
ms per token,  1721.29 tokens per second)
llama_print_timings: prompt eval time =   10297.33 ms /    85 tokens (  121.15
ms per token,     8.25 tokens per second)
llama_print_timings:        eval time =    8182.11 ms /    23 runs   (  355.74
ms per token,     2.81 tokens per second)
llama_print_timings:       total time =   18680.71 ms /   108 tokens
No. of rows:  28%| | 356/1258 [10:22:18<4:41:53, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.22 ms /    22 runs   (    0.46
ms per token,  2152.01 tokens per second)
llama_print_timings: prompt eval time =   11931.15 ms /    91 tokens (  131.11
ms per token,     7.63 tokens per second)
llama_print_timings:        eval time =    6631.52 ms /    21 runs   (  315.79
ms per token,     3.17 tokens per second)
llama_print_timings:       total time =   18707.82 ms /   112 tokens
No. of rows:  28%| | 357/1258 [10:22:37<4:41:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      25.83 ms /    44 runs   (    0.59
ms per token,  1703.71 tokens per second)
llama_print_timings: prompt eval time =   15957.02 ms /   116 tokens (  137.56
ms per token,     7.27 tokens per second)
llama_print_timings:        eval time =   13169.55 ms /    43 runs   (  306.27
ms per token,     3.27 tokens per second)
llama_print_timings:       total time =   29543.54 ms /   159 tokens
No. of rows:  28%| | 358/1258 [10:23:07<5:29:45, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.25 ms /    30 runs   (    0.41
ms per token,  2449.58 tokens per second)
llama_print_timings: prompt eval time =   11632.69 ms /    97 tokens (  119.92
ms per token,     8.34 tokens per second)
llama_print_timings:        eval time =    8236.24 ms /    29 runs   (  284.01
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   20060.92 ms /   126 tokens
No. of rows:  29%| | 359/1258 [10:23:27<5:20:47, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.75 ms /    28 runs   (    0.46
```

```
ms per token,  2195.73 tokens per second)
llama_print_timings: prompt eval time =   10414.54 ms /    99 tokens (  105.20
ms per token,     9.51 tokens per second)
llama_print_timings:        eval time =    6874.73 ms /    27 runs   (  254.62
ms per token,     3.93 tokens per second)
llama_print_timings:       total time =   17485.75 ms /   126 tokens
No. of rows:  29%| | 360/1258 [10:23:44<5:02:50, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.65 ms /    22 runs   (    0.44
ms per token,  2279.56 tokens per second)
llama_print_timings: prompt eval time =    9641.04 ms /    89 tokens (  108.33
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =    5010.92 ms /    21 runs   (  238.62
ms per token,     4.19 tokens per second)
llama_print_timings:       total time =   14805.77 ms /   110 tokens
No. of rows:  29%| | 361/1258 [10:23:59<4:38:11, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.43 ms /    31 runs   (    0.47
ms per token,  2149.05 tokens per second)
llama_print_timings: prompt eval time =   16319.26 ms /   105 tokens (  155.42
ms per token,     6.43 tokens per second)
llama_print_timings:        eval time =    7414.63 ms /    30 runs   (  247.15
ms per token,     4.05 tokens per second)
llama_print_timings:       total time =   23954.38 ms /   135 tokens
No. of rows:  29%| | 362/1258 [10:24:23<5:01:50, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.93 ms /    23 runs   (    0.48
ms per token,  2104.49 tokens per second)
llama_print_timings: prompt eval time =   10778.70 ms /    99 tokens (  108.88
ms per token,     9.18 tokens per second)
llama_print_timings:        eval time =    6113.06 ms /    22 runs   (  277.87
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   17058.67 ms /   121 tokens
No. of rows:  29%| | 363/1258 [10:24:40<4:47:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.80 ms /    29 runs   (    0.58
ms per token,  1725.88 tokens per second)
llama_print_timings: prompt eval time =   16325.31 ms /   120 tokens (  136.04
ms per token,     7.35 tokens per second)
llama_print_timings:        eval time =    8148.39 ms /    28 runs   (  291.01
```

```
ms per token,     3.44 tokens per second)
llama_print_timings:       total time =   24718.52 ms /   148 tokens
No. of rows:  29%| | 364/1258 [10:25:05<5:11:30, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.19 ms /    20 runs   (    0.41
ms per token,  2443.49 tokens per second)
llama_print_timings: prompt eval time =    8837.60 ms /    79 tokens (  111.87
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =    4746.91 ms /    19 runs   (  249.84
ms per token,     4.00 tokens per second)
llama_print_timings:       total time =   13734.11 ms /    98 tokens
No. of rows:  29%| | 365/1258 [10:25:18<4:39:10, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.63 ms /    25 runs   (    0.47
ms per token,  2149.43 tokens per second)
llama_print_timings: prompt eval time =   11168.00 ms /    92 tokens (  121.39
ms per token,     8.24 tokens per second)
llama_print_timings:        eval time =    7982.70 ms /    24 runs   (  332.61
ms per token,     3.01 tokens per second)
llama_print_timings:       total time =   19324.58 ms /   116 tokens
No. of rows:  29%| | 366/1258 [10:25:38<4:41:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.03 ms /    25 runs   (    0.40
ms per token,  2493.02 tokens per second)
llama_print_timings: prompt eval time =   11814.59 ms /    99 tokens (  119.34
ms per token,     8.38 tokens per second)
llama_print_timings:        eval time =    5862.70 ms /    24 runs   (  244.28
ms per token,     4.09 tokens per second)
llama_print_timings:       total time =   17830.78 ms /   123 tokens
No. of rows:  29%| | 367/1258 [10:25:56<4:36:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.96 ms /    26 runs   (    0.38
ms per token,  2609.92 tokens per second)
llama_print_timings: prompt eval time =   10460.49 ms /    95 tokens (  110.11
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    5506.10 ms /    25 runs   (  220.24
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   16125.22 ms /   120 tokens
No. of rows:  29%| | 368/1258 [10:26:12<4:24:59, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.98 ms /    20 runs   (    0.80
ms per token,  1251.80 tokens per second)
llama_print_timings: prompt eval time =   12711.21 ms /    90 tokens (  141.24
ms per token,     7.08 tokens per second)
llama_print_timings:        eval time =    6760.72 ms /    19 runs   (  355.83
ms per token,     2.81 tokens per second)
llama_print_timings:       total time =   19666.26 ms /   109 tokens
No. of rows:  29%| | 369/1258 [10:26:31<4:32:45, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.23 ms /    17 runs   (    0.43
ms per token,  2350.99 tokens per second)
llama_print_timings: prompt eval time =   11561.73 ms /    68 tokens (  170.03
ms per token,     5.88 tokens per second)
llama_print_timings:        eval time =    4810.45 ms /    16 runs   (  300.65
ms per token,     3.33 tokens per second)
llama_print_timings:       total time =   16490.04 ms /    84 tokens
No. of rows:  29%| | 370/1258 [10:26:48<4:23:55, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.53 ms /    33 runs   (    0.38
ms per token,  2634.10 tokens per second)
llama_print_timings: prompt eval time =   13682.31 ms /   127 tokens (  107.73
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    6967.80 ms /    32 runs   (  217.74
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   20845.41 ms /   159 tokens
No. of rows:  29%| | 371/1258 [10:27:09<4:37:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.17 ms /    21 runs   (    0.39
ms per token,  2570.38 tokens per second)
llama_print_timings: prompt eval time =   10600.36 ms /    82 tokens (  129.27
ms per token,     7.74 tokens per second)
llama_print_timings:        eval time =    4401.48 ms /    20 runs   (  220.07
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   15128.22 ms /   102 tokens
No. of rows:  30%| | 372/1258 [10:27:24<4:20:47, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.21 ms /    28 runs   (    0.40
ms per token,  2498.22 tokens per second)
```

```
llama_print_timings: prompt eval time =   12025.78 ms /   101 tokens (  119.07
ms per token,     8.40 tokens per second)
llama_print_timings:        eval time =    7741.32 ms /    27 runs   (  286.72
ms per token,     3.49 tokens per second)
llama_print_timings:       total time =   19940.53 ms /   128 tokens
No. of rows:  30%| | 373/1258 [10:27:44<4:30:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.77 ms /    23 runs   (    0.38
ms per token,  2621.68 tokens per second)
llama_print_timings: prompt eval time =   10041.11 ms /    92 tokens (  109.14
ms per token,     9.16 tokens per second)
llama_print_timings:        eval time =    4757.73 ms /    22 runs   (  216.26
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   14935.98 ms /   114 tokens
No. of rows:  30%| | 374/1258 [10:27:59<4:15:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.62 ms /    32 runs   (    0.39
ms per token,  2536.66 tokens per second)
llama_print_timings: prompt eval time =   12809.91 ms /   118 tokens (  108.56
ms per token,     9.21 tokens per second)
llama_print_timings:        eval time =    8429.69 ms /    31 runs   (  271.93
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   21433.00 ms /   149 tokens
No. of rows:  30%| | 375/1258 [10:28:20<4:33:09, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.99 ms /    21 runs   (    0.38
ms per token,  2626.97 tokens per second)
llama_print_timings: prompt eval time =   10515.15 ms /   100 tokens (  105.15
ms per token,     9.51 tokens per second)
llama_print_timings:        eval time =    5914.43 ms /    20 runs   (  295.72
ms per token,     3.38 tokens per second)
llama_print_timings:       total time =   16554.08 ms /   120 tokens
No. of rows:  30%| | 376/1258 [10:28:37<4:24:00, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.95 ms /    23 runs   (    0.39
ms per token,  2569.26 tokens per second)
llama_print_timings: prompt eval time =   10330.20 ms /    96 tokens (  107.61
ms per token,     9.29 tokens per second)
llama_print_timings:        eval time =    5074.09 ms /    22 runs   (  230.64
ms per token,     4.34 tokens per second)
```

```
llama_print_timings:       total time =   15539.97 ms /   118 tokens
No. of rows:  30%| | 377/1258 [10:28:52<4:13:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.48 ms /    39 runs   (    0.40
ms per token,  2519.05 tokens per second)
llama_print_timings: prompt eval time =   13461.84 ms /   126 tokens (  106.84
ms per token,     9.36 tokens per second)
llama_print_timings:        eval time =   10248.93 ms /    38 runs   (  269.71
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   23953.02 ms /   164 tokens
No. of rows:  30%| | 378/1258 [10:29:16<4:42:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.08 ms /    33 runs   (    0.40
ms per token,  2522.55 tokens per second)
llama_print_timings: prompt eval time =   13046.38 ms /   123 tokens (  106.07
ms per token,     9.43 tokens per second)
llama_print_timings:        eval time =    8733.37 ms /    32 runs   (  272.92
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   21987.55 ms /   155 tokens
No. of rows:  30%| | 379/1258 [10:29:38<4:54:05, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.32 ms /    27 runs   (    0.38
ms per token,  2617.04 tokens per second)
llama_print_timings: prompt eval time =   11094.68 ms /   101 tokens (  109.85
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    5688.14 ms /    26 runs   (  218.77
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   16944.13 ms /   127 tokens
No. of rows:  30%| | 380/1258 [10:29:55<4:40:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.62 ms /    20 runs   (    0.38
ms per token,  2625.02 tokens per second)
llama_print_timings: prompt eval time =    8291.99 ms /    77 tokens (  107.69
ms per token,     9.29 tokens per second)
llama_print_timings:        eval time =    4065.71 ms /    19 runs   (  213.98
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   12476.46 ms /    96 tokens
No. of rows:  30%| | 381/1258 [10:30:08<4:10:35, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.66 ms /    19 runs   (    0.40
ms per token,  2482.04 tokens per second)
llama_print_timings: prompt eval time =    9070.20 ms /    85 tokens (  106.71
ms per token,     9.37 tokens per second)
llama_print_timings:        eval time =    4380.42 ms /    18 runs   (  243.36
ms per token,     4.11 tokens per second)
llama_print_timings:       total time =   13562.54 ms /   103 tokens
No. of rows:  30%| | 382/1258 [10:30:21<3:54:38, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.31 ms /    27 runs   (    0.38
ms per token,  2617.80 tokens per second)
llama_print_timings: prompt eval time =   10272.18 ms /    93 tokens (  110.45
ms per token,     9.05 tokens per second)
llama_print_timings:        eval time =    7430.26 ms /    26 runs   (  285.78
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   17860.90 ms /   119 tokens
No. of rows:  30%| | 383/1258 [10:30:39<4:02:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.27 ms /    29 runs   (    0.39
ms per token,  2572.29 tokens per second)
llama_print_timings: prompt eval time =   11201.09 ms /   104 tokens (  107.70
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    6521.96 ms /    28 runs   (  232.93
ms per token,     4.29 tokens per second)
llama_print_timings:       total time =   17899.19 ms /   132 tokens
No. of rows:  31%| | 384/1258 [10:30:57<4:07:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.60 ms /    27 runs   (    0.39
ms per token,  2548.37 tokens per second)
llama_print_timings: prompt eval time =    9649.07 ms /    88 tokens (  109.65
ms per token,     9.12 tokens per second)
llama_print_timings:        eval time =    7067.64 ms /    26 runs   (  271.83
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   16878.81 ms /   114 tokens
No. of rows:  31%| | 385/1258 [10:31:14<4:06:48, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.76 ms /    25 runs   (    0.39
ms per token,  2561.21 tokens per second)
llama_print_timings: prompt eval time =   10726.79 ms /   100 tokens (  107.27
```

```
ms per token,     9.32 tokens per second)
llama_print_timings:        eval time =    5129.55 ms /    24 runs   (  213.73
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   16005.13 ms /   124 tokens
No. of rows:  31%| | 386/1258 [10:31:30<4:02:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.62 ms /    29 runs   (    0.40
ms per token,  2495.91 tokens per second)
llama_print_timings: prompt eval time =   14056.68 ms /   117 tokens (  120.14
ms per token,     8.32 tokens per second)
llama_print_timings:        eval time =    6146.06 ms /    28 runs   (  219.50
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   20381.33 ms /   145 tokens
No. of rows:  31%| | 387/1258 [10:31:50<4:18:14, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.88 ms /    20 runs   (    0.39
ms per token,  2539.04 tokens per second)
llama_print_timings: prompt eval time =    8312.51 ms /    79 tokens (  105.22
ms per token,     9.50 tokens per second)
llama_print_timings:        eval time =    4164.91 ms /    19 runs   (  219.21
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   12600.19 ms /    98 tokens
No. of rows:  31%| | 388/1258 [10:32:03<3:55:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.29 ms /    24 runs   (    0.39
ms per token,  2584.54 tokens per second)
llama_print_timings: prompt eval time =    8947.76 ms /    84 tokens (  106.52
ms per token,     9.39 tokens per second)
llama_print_timings:        eval time =    4928.40 ms /    23 runs   (  214.28
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   14018.64 ms /   107 tokens
No. of rows:  31%| | 389/1258 [10:32:17<3:45:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.89 ms /    25 runs   (    0.40
ms per token,  2528.83 tokens per second)
llama_print_timings: prompt eval time =   10854.25 ms /   103 tokens (  105.38
ms per token,     9.49 tokens per second)
llama_print_timings:        eval time =    5240.73 ms /    24 runs   (  218.36
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   16251.85 ms /   127 tokens
```

```
No. of rows:  31%| | 390/1258 [10:32:33<3:48:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.88 ms /    21 runs   (    0.38
ms per token,  2665.31 tokens per second)
llama_print_timings: prompt eval time =   10279.91 ms /    81 tokens (  126.91
ms per token,     7.88 tokens per second)
llama_print_timings:        eval time =    4658.21 ms /    20 runs   (  232.91
ms per token,     4.29 tokens per second)
llama_print_timings:       total time =   15063.91 ms /   101 tokens
No. of rows:  31%| | 391/1258 [10:32:48<3:44:54, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.96 ms /    34 runs   (    0.38
ms per token,  2622.65 tokens per second)
llama_print_timings: prompt eval time =    9984.75 ms /    95 tokens (  105.10
ms per token,     9.51 tokens per second)
llama_print_timings:        eval time =    7461.87 ms /    33 runs   (  226.12
ms per token,     4.42 tokens per second)
llama_print_timings:       total time =   17655.42 ms /   128 tokens
No. of rows:  31%| | 392/1258 [10:33:06<3:53:44, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.38 ms /    31 runs   (    0.40
ms per token,  2503.23 tokens per second)
llama_print_timings: prompt eval time =   12423.05 ms /   114 tokens (  108.97
ms per token,     9.18 tokens per second)
llama_print_timings:        eval time =    8170.82 ms /    30 runs   (  272.36
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   20788.89 ms /   144 tokens
No. of rows:  31%| | 393/1258 [10:33:27<4:13:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.04 ms /    22 runs   (    0.41
ms per token,  2434.17 tokens per second)
llama_print_timings: prompt eval time =   10519.16 ms /    99 tokens (  106.25
ms per token,     9.41 tokens per second)
llama_print_timings:        eval time =    4662.83 ms /    21 runs   (  222.04
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   15322.46 ms /   120 tokens
No. of rows:  31%| | 394/1258 [10:33:42<4:03:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       7.38 ms /    19 runs   (    0.39
ms per token,  2575.92 tokens per second)
llama_print_timings: prompt eval time =   10864.45 ms /    87 tokens (  124.88
ms per token,     8.01 tokens per second)
llama_print_timings:        eval time =    3914.85 ms /    18 runs   (  217.49
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   14890.16 ms /   105 tokens
No. of rows:  31%| | 395/1258 [10:33:57<3:54:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.63 ms /    25 runs   (    0.39
ms per token,  2597.13 tokens per second)
llama_print_timings: prompt eval time =   10108.92 ms /    94 tokens (  107.54
ms per token,     9.30 tokens per second)
llama_print_timings:        eval time =    5272.35 ms /    24 runs   (  219.68
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   15534.21 ms /   118 tokens
No. of rows:  31%| | 396/1258 [10:34:13<3:50:53, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.02 ms /    19 runs   (    0.37
ms per token,  2707.32 tokens per second)
llama_print_timings: prompt eval time =    8391.72 ms /    79 tokens (  106.22
ms per token,     9.41 tokens per second)
llama_print_timings:        eval time =    3925.80 ms /    18 runs   (  218.10
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   12430.94 ms /    97 tokens
No. of rows:  32%| | 397/1258 [10:34:25<3:35:00, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.65 ms /    20 runs   (    0.38
ms per token,  2614.04 tokens per second)
llama_print_timings: prompt eval time =   10594.01 ms /    99 tokens (  107.01
ms per token,     9.34 tokens per second)
llama_print_timings:        eval time =    4096.94 ms /    19 runs   (  215.63
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14810.09 ms /   118 tokens
No. of rows:  32%| | 398/1258 [10:34:40<3:34:03, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.82 ms /    28 runs   (    0.39
ms per token,  2587.56 tokens per second)
llama_print_timings: prompt eval time =   11701.69 ms /    97 tokens (  120.64
ms per token,     8.29 tokens per second)
```

```
llama_print_timings:        eval time =    7566.18 ms /    27 runs   (  280.23
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   19436.69 ms /   124 tokens
No. of rows:  32%| | 399/1258 [10:34:59<3:53:10, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.14 ms /    29 runs   (    0.38
ms per token,  2604.17 tokens per second)
llama_print_timings: prompt eval time =   10426.44 ms /    99 tokens (  105.32
ms per token,     9.50 tokens per second)
llama_print_timings:        eval time =    5975.45 ms /    28 runs   (  213.41
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   16571.31 ms /   127 tokens
No. of rows:  32%| | 400/1258 [10:35:16<3:54:10, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.04 ms /    38 runs   (    0.37
ms per token,  2706.55 tokens per second)
llama_print_timings: prompt eval time =   13827.62 ms /   117 tokens (  118.18
ms per token,     8.46 tokens per second)
llama_print_timings:        eval time =    8006.67 ms /    37 runs   (  216.40
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   22064.96 ms /   154 tokens
No. of rows:  32%| | 401/1258 [10:35:38<4:18:14, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.95 ms /    27 runs   (    0.37
ms per token,  2712.75 tokens per second)
llama_print_timings: prompt eval time =   12251.91 ms /    98 tokens (  125.02
ms per token,     8.00 tokens per second)
llama_print_timings:        eval time =    5597.42 ms /    26 runs   (  215.29
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   18008.36 ms /   124 tokens
No. of rows:  32%| | 402/1258 [10:35:56<4:17:43, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.72 ms /    27 runs   (    0.40
ms per token,  2518.42 tokens per second)
llama_print_timings: prompt eval time =    9853.66 ms /    93 tokens (  105.95
ms per token,     9.44 tokens per second)
llama_print_timings:        eval time =    5613.83 ms /    26 runs   (  215.92
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   15633.80 ms /   119 tokens
No. of rows:  32%| | 403/1258 [10:36:12<4:07:01, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.98 ms /    36 runs   (    0.36
ms per token,  2773.93 tokens per second)
llama_print_timings: prompt eval time =   13373.28 ms /   127 tokens (  105.30
ms per token,     9.50 tokens per second)
llama_print_timings:        eval time =    7520.33 ms /    35 runs   (  214.87
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   21103.01 ms /   162 tokens
No. of rows:  32%| | 404/1258 [10:36:33<4:22:53, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.07 ms /    19 runs   (    0.37
ms per token,  2688.17 tokens per second)
llama_print_timings: prompt eval time =   10788.66 ms /    89 tokens (  121.22
ms per token,     8.25 tokens per second)
llama_print_timings:        eval time =    3825.28 ms /    18 runs   (  212.52
ms per token,     4.71 tokens per second)
llama_print_timings:       total time =   14724.67 ms /   107 tokens
No. of rows:  32%| | 405/1258 [10:36:47<4:06:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.86 ms /    30 runs   (    0.40
ms per token,  2528.87 tokens per second)
llama_print_timings: prompt eval time =   12750.23 ms /   111 tokens (  114.87
ms per token,     8.71 tokens per second)
llama_print_timings:        eval time =    8291.27 ms /    29 runs   (  285.91
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   21225.48 ms /   140 tokens
No. of rows:  32%| | 406/1258 [10:37:09<4:22:54, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.15 ms /    28 runs   (    0.40
ms per token,  2511.89 tokens per second)
llama_print_timings: prompt eval time =    9869.63 ms /    94 tokens (  105.00
ms per token,     9.52 tokens per second)
llama_print_timings:        eval time =    5962.00 ms /    27 runs   (  220.81
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   16001.13 ms /   121 tokens
No. of rows:  32%| | 407/1258 [10:37:25<4:11:55, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       4.10 ms /    11 runs   (    0.37
```

```
ms per token,  2682.27 tokens per second)
llama_print_timings: prompt eval time =    6952.17 ms /    66 tokens (  105.34
ms per token,     9.49 tokens per second)
llama_print_timings:        eval time =    2188.11 ms /    10 runs   (  218.81
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =    9205.83 ms /    76 tokens
No. of rows:  32%| | 408/1258 [10:37:34<3:35:19, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.86 ms /    24 runs   (    0.37
ms per token,  2708.50 tokens per second)
llama_print_timings: prompt eval time =    9530.09 ms /    88 tokens (  108.30
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =    5009.57 ms /    23 runs   (  217.81
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   14685.05 ms /   111 tokens
No. of rows:  33%| | 409/1258 [10:37:49<3:32:53, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.40 ms /    32 runs   (    0.39
ms per token,  2581.06 tokens per second)
llama_print_timings: prompt eval time =   13008.76 ms /   108 tokens (  120.45
ms per token,     8.30 tokens per second)
llama_print_timings:        eval time =    6675.01 ms /    31 runs   (  215.32
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   19879.47 ms /   139 tokens
No. of rows:  33%| | 410/1258 [10:38:08<3:53:08, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.87 ms /    31 runs   (    0.38
ms per token,  2612.73 tokens per second)
llama_print_timings: prompt eval time =   10619.35 ms /   100 tokens (  106.19
ms per token,     9.42 tokens per second)
llama_print_timings:        eval time =    6428.87 ms /    30 runs   (  214.30
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   17231.48 ms /   130 tokens
No. of rows:  33%| | 411/1258 [10:38:26<3:56:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.55 ms /    20 runs   (    0.38
ms per token,  2649.01 tokens per second)
llama_print_timings: prompt eval time =    8948.62 ms /    85 tokens (  105.28
ms per token,     9.50 tokens per second)
llama_print_timings:        eval time =    5902.64 ms /    19 runs   (  310.67
```

```
ms per token,     3.22 tokens per second)
llama_print_timings:       total time =   14969.35 ms /   104 tokens
No. of rows:  33%| | 412/1258 [10:38:41<3:48:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.98 ms /    26 runs   (    0.38
ms per token,  2605.73 tokens per second)
llama_print_timings: prompt eval time =    9920.99 ms /    93 tokens (  106.68
ms per token,     9.37 tokens per second)
llama_print_timings:        eval time =    7060.57 ms /    25 runs   (  282.42
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   17133.98 ms /   118 tokens
No. of rows:  33%| | 413/1258 [10:38:58<3:52:08, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.00 ms /    34 runs   (    0.38
ms per token,  2615.18 tokens per second)
llama_print_timings: prompt eval time =   12222.51 ms /   118 tokens (  103.58
ms per token,     9.65 tokens per second)
llama_print_timings:        eval time =    7073.79 ms /    33 runs   (  214.36
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   19504.78 ms /   151 tokens
No. of rows:  33%| | 414/1258 [10:39:17<4:04:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.94 ms /    28 runs   (    0.36
ms per token,  2816.33 tokens per second)
llama_print_timings: prompt eval time =   11959.57 ms /   100 tokens (  119.60
ms per token,     8.36 tokens per second)
llama_print_timings:        eval time =    7552.15 ms /    27 runs   (  279.71
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   19680.02 ms /   127 tokens
No. of rows:  33%| | 415/1258 [10:39:37<4:14:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.29 ms /    20 runs   (    0.36
ms per token,  2745.37 tokens per second)
llama_print_timings: prompt eval time =    9323.02 ms /    86 tokens (  108.41
ms per token,     9.22 tokens per second)
llama_print_timings:        eval time =    4547.60 ms /    19 runs   (  239.35
ms per token,     4.18 tokens per second)
llama_print_timings:       total time =   13991.32 ms /   105 tokens
No. of rows:  33%| | 416/1258 [10:39:51<3:56:32, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.18 ms /    26 runs   (    0.43
ms per token,  2326.21 tokens per second)
llama_print_timings: prompt eval time =   10771.43 ms /   100 tokens (  107.71
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    5505.12 ms /    25 runs   (  220.20
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   16434.46 ms /   125 tokens
No. of rows:  33%| | 417/1258 [10:40:07<3:54:31, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.60 ms /    28 runs   (    0.38
ms per token,  2641.51 tokens per second)
llama_print_timings: prompt eval time =    9048.55 ms /    84 tokens (  107.72
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    5788.52 ms /    27 runs   (  214.39
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   15001.52 ms /   111 tokens
No. of rows:  33%| | 418/1258 [10:40:22<3:47:00, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.65 ms /    30 runs   (    0.39
ms per token,  2574.67 tokens per second)
llama_print_timings: prompt eval time =   14136.76 ms /   118 tokens (  119.80
ms per token,     8.35 tokens per second)
llama_print_timings:        eval time =    7967.28 ms /    29 runs   (  274.73
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   22285.31 ms /   147 tokens
No. of rows:  33%| | 419/1258 [10:40:45<4:12:12, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.61 ms /    41 runs   (    0.38
ms per token,  2626.69 tokens per second)
llama_print_timings: prompt eval time =   10307.08 ms /    99 tokens (  104.11
ms per token,     9.61 tokens per second)
llama_print_timings:        eval time =    8593.38 ms /    40 runs   (  214.83
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   19149.12 ms /   139 tokens
No. of rows:  33%| | 420/1258 [10:41:04<4:16:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.29 ms /    24 runs   (    0.39
ms per token,  2582.59 tokens per second)
```

```
llama_print_timings: prompt eval time =   10004.55 ms /    93 tokens (  107.58
ms per token,     9.30 tokens per second)
llama_print_timings:        eval time =    4914.65 ms /    23 runs   (  213.68
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   15060.97 ms /   116 tokens
No. of rows:  33%| | 421/1258 [10:41:19<4:02:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.90 ms /    29 runs   (    0.38
ms per token,  2661.04 tokens per second)
llama_print_timings: prompt eval time =   11703.35 ms /    92 tokens (  127.21
ms per token,     7.86 tokens per second)
llama_print_timings:        eval time =    6117.40 ms /    28 runs   (  218.48
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   17997.82 ms /   120 tokens
No. of rows:  34%| | 422/1258 [10:41:37<4:04:46, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.38 ms /    21 runs   (    0.40
ms per token,  2505.07 tokens per second)
llama_print_timings: prompt eval time =   10193.90 ms /    94 tokens (  108.45
ms per token,     9.22 tokens per second)
llama_print_timings:        eval time =    4390.03 ms /    20 runs   (  219.50
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   14711.42 ms /   114 tokens
No. of rows:  34%| | 423/1258 [10:41:52<3:52:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.82 ms /    31 runs   (    0.38
ms per token,  2622.67 tokens per second)
llama_print_timings: prompt eval time =   12995.44 ms /   107 tokens (  121.45
ms per token,     8.23 tokens per second)
llama_print_timings:        eval time =    6486.59 ms /    30 runs   (  216.22
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   19671.66 ms /   137 tokens
No. of rows:  34%| | 424/1258 [10:42:11<4:04:41, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.42 ms /    24 runs   (    0.39
ms per token,  2549.12 tokens per second)
llama_print_timings: prompt eval time =    9311.44 ms /    87 tokens (  107.03
ms per token,     9.34 tokens per second)
llama_print_timings:        eval time =    5113.49 ms /    23 runs   (  222.33
ms per token,     4.50 tokens per second)
```

```
llama_print_timings:       total time =   14575.94 ms /   110 tokens
No. of rows:  34%| | 425/1258 [10:42:26<3:51:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.55 ms /    28 runs   (    0.38
ms per token,  2652.77 tokens per second)
llama_print_timings: prompt eval time =   11657.11 ms /    95 tokens (  122.71
ms per token,     8.15 tokens per second)
llama_print_timings:        eval time =    5811.32 ms /    27 runs   (  215.23
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   17633.19 ms /   122 tokens
No. of rows:  34%| | 426/1258 [10:42:44<3:55:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.66 ms /    38 runs   (    0.39
ms per token,  2591.56 tokens per second)
llama_print_timings: prompt eval time =   13346.50 ms /   111 tokens (  120.24
ms per token,     8.32 tokens per second)
llama_print_timings:        eval time =    8025.65 ms /    37 runs   (  216.91
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   21597.27 ms /   148 tokens
No. of rows:  34%| | 427/1258 [10:43:05<4:14:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.81 ms /    18 runs   (    0.38
ms per token,  2642.01 tokens per second)
llama_print_timings: prompt eval time =    8801.55 ms /    82 tokens (  107.34
ms per token,     9.32 tokens per second)
llama_print_timings:        eval time =    3622.52 ms /    17 runs   (  213.09
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   12529.17 ms /    99 tokens
No. of rows:  34%| | 428/1258 [10:43:18<3:49:52, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.37 ms /    36 runs   (    0.40
ms per token,  2504.70 tokens per second)
llama_print_timings: prompt eval time =   13041.39 ms /   109 tokens (  119.65
ms per token,     8.36 tokens per second)
llama_print_timings:        eval time =    9276.68 ms /    35 runs   (  265.05
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   22536.33 ms /   144 tokens
No. of rows:  34%| | 429/1258 [10:43:40<4:14:12, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.68 ms /    33 runs   (    0.38
ms per token,  2602.93 tokens per second)
llama_print_timings: prompt eval time =   12293.94 ms /   113 tokens (  108.80
ms per token,     9.19 tokens per second)
llama_print_timings:        eval time =    7103.16 ms /    32 runs   (  221.97
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   19594.42 ms /   145 tokens
No. of rows:  34%| | 430/1258 [10:44:00<4:18:53, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.83 ms /    29 runs   (    0.37
ms per token,  2677.25 tokens per second)
llama_print_timings: prompt eval time =   11561.34 ms /   109 tokens (  106.07
ms per token,     9.43 tokens per second)
llama_print_timings:        eval time =    6115.98 ms /    28 runs   (  218.43
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   17850.65 ms /   137 tokens
No. of rows:  34%| | 431/1258 [10:44:18<4:14:50, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.69 ms /    25 runs   (    0.39
ms per token,  2578.91 tokens per second)
llama_print_timings: prompt eval time =   10682.59 ms /   100 tokens (  106.83
ms per token,     9.36 tokens per second)
llama_print_timings:        eval time =    5131.08 ms /    24 runs   (  213.80
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   15964.00 ms /   124 tokens
No. of rows:  34%| | 432/1258 [10:44:34<4:04:08, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.12 ms /    19 runs   (    0.37
ms per token,  2667.04 tokens per second)
llama_print_timings: prompt eval time =    8760.59 ms /    83 tokens (  105.55
ms per token,     9.47 tokens per second)
llama_print_timings:        eval time =    3827.35 ms /    18 runs   (  212.63
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   12703.92 ms /   101 tokens
No. of rows:  34%| | 433/1258 [10:44:46<3:43:06, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.87 ms /    25 runs   (    0.39
ms per token,  2533.95 tokens per second)
llama_print_timings: prompt eval time =   12141.76 ms /   100 tokens (  121.42
```

```
ms per token,     8.24 tokens per second)
llama_print_timings:        eval time =    5506.99 ms /    24 runs   (  229.46
ms per token,     4.36 tokens per second)
llama_print_timings:       total time =   17803.76 ms /   124 tokens
No. of rows:  34%| | 434/1258 [10:45:04<3:49:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.51 ms /    25 runs   (    0.38
ms per token,  2629.36 tokens per second)
llama_print_timings: prompt eval time =   10258.52 ms /    98 tokens (  104.68
ms per token,     9.55 tokens per second)
llama_print_timings:        eval time =    5128.48 ms /    24 runs   (  213.69
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   15541.67 ms /   122 tokens
No. of rows:  35%| | 435/1258 [10:45:20<3:44:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      23.23 ms /    50 runs   (    0.46
ms per token,  2151.93 tokens per second)
llama_print_timings: prompt eval time =   15941.38 ms /   139 tokens (  114.69
ms per token,     8.72 tokens per second)
llama_print_timings:        eval time =   11140.61 ms /    49 runs   (  227.36
ms per token,     4.40 tokens per second)
llama_print_timings:       total time =   27397.45 ms /   188 tokens
No. of rows:  35%| | 436/1258 [10:45:47<4:29:31, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.46 ms /    19 runs   (    0.39
ms per token,  2546.58 tokens per second)
llama_print_timings: prompt eval time =    9432.40 ms /    89 tokens (  105.98
ms per token,     9.44 tokens per second)
llama_print_timings:        eval time =    4080.16 ms /    18 runs   (  226.68
ms per token,     4.41 tokens per second)
llama_print_timings:       total time =   13637.11 ms /   107 tokens
No. of rows:  35%| | 437/1258 [10:46:01<4:04:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.18 ms /    23 runs   (    0.44
ms per token,  2258.67 tokens per second)
llama_print_timings: prompt eval time =    9614.85 ms /    89 tokens (  108.03
ms per token,     9.26 tokens per second)
llama_print_timings:        eval time =    4946.37 ms /    22 runs   (  224.83
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   14703.04 ms /   111 tokens
```

```
No. of rows:  35%| | 438/1258 [10:46:16<3:51:13, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.65 ms /    31 runs   (    0.41
ms per token,  2449.82 tokens per second)
llama_print_timings: prompt eval time =   11136.03 ms /   107 tokens (  104.08
ms per token,     9.61 tokens per second)
llama_print_timings:        eval time =    8147.53 ms /    30 runs   (  271.58
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   19472.55 ms /   137 tokens
No. of rows:  35%| | 439/1258 [10:46:35<4:01:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.96 ms /    24 runs   (    0.50
ms per token,  2005.85 tokens per second)
llama_print_timings: prompt eval time =    9227.01 ms /    87 tokens (  106.06
ms per token,     9.43 tokens per second)
llama_print_timings:        eval time =    4965.41 ms /    23 runs   (  215.89
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   14342.05 ms /   110 tokens
No. of rows:  35%| | 440/1258 [10:46:49<3:47:28, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.45 ms /    22 runs   (    0.38
ms per token,  2604.17 tokens per second)
llama_print_timings: prompt eval time =    8569.47 ms /    82 tokens (  104.51
ms per token,     9.57 tokens per second)
llama_print_timings:        eval time =    4509.41 ms /    21 runs   (  214.73
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   13211.40 ms /   103 tokens
No. of rows:  35%| | 441/1258 [10:47:03<3:33:03, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.22 ms /    50 runs   (    0.38
ms per token,  2601.73 tokens per second)
llama_print_timings: prompt eval time =   14571.61 ms /   125 tokens (  116.57
ms per token,     8.58 tokens per second)
llama_print_timings:        eval time =   12322.74 ms /    49 runs   (  251.48
ms per token,     3.98 tokens per second)
llama_print_timings:       total time =   27196.89 ms /   174 tokens
No. of rows:  35%| | 442/1258 [10:47:30<4:19:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =      13.33 ms /    34 runs   (    0.39
ms per token,  2550.45 tokens per second)
llama_print_timings: prompt eval time =   12287.84 ms /   116 tokens (  105.93
ms per token,     9.44 tokens per second)
llama_print_timings:        eval time =    7275.65 ms /    33 runs   (  220.47
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   19771.79 ms /   149 tokens
No. of rows:  35%| | 443/1258 [10:47:50<4:22:19, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.13 ms /    31 runs   (    0.39
ms per token,  2556.49 tokens per second)
llama_print_timings: prompt eval time =   15478.44 ms /   130 tokens (  119.06
ms per token,     8.40 tokens per second)
llama_print_timings:        eval time =    6624.32 ms /    30 runs   (  220.81
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   22291.24 ms /   160 tokens
No. of rows:  35%| | 444/1258 [10:48:12<4:34:07, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.93 ms /    25 runs   (    0.36
ms per token,  2798.93 tokens per second)
llama_print_timings: prompt eval time =   11723.84 ms /   111 tokens (  105.62
ms per token,     9.47 tokens per second)
llama_print_timings:        eval time =    5688.06 ms /    24 runs   (  237.00
ms per token,     4.22 tokens per second)
llama_print_timings:       total time =   17557.82 ms /   135 tokens
No. of rows:  35%| | 445/1258 [10:48:29<4:23:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.43 ms /    22 runs   (    0.38
ms per token,  2610.04 tokens per second)
llama_print_timings: prompt eval time =    8701.40 ms /    81 tokens (  107.42
ms per token,     9.31 tokens per second)
llama_print_timings:        eval time =    4827.86 ms /    21 runs   (  229.90
ms per token,     4.35 tokens per second)
llama_print_timings:       total time =   13658.97 ms /   102 tokens
No. of rows:  35%| | 446/1258 [10:48:43<3:59:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.27 ms /    21 runs   (    0.39
ms per token,  2538.68 tokens per second)
llama_print_timings: prompt eval time =    8679.33 ms /    82 tokens (  105.85
ms per token,     9.45 tokens per second)
```

```
llama_print_timings:        eval time =    4378.74 ms /    20 runs   (  218.94
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   13185.87 ms /   102 tokens
No. of rows:  36%| | 447/1258 [10:48:56<3:40:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.41 ms /    19 runs   (    0.39
ms per token,  2565.83 tokens per second)
llama_print_timings: prompt eval time =    9277.10 ms /    71 tokens (  130.66
ms per token,     7.65 tokens per second)
llama_print_timings:        eval time =    4217.78 ms /    18 runs   (  234.32
ms per token,     4.27 tokens per second)
llama_print_timings:       total time =   13608.57 ms /    89 tokens
No. of rows:  36%| | 448/1258 [10:49:10<3:29:33, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.59 ms /    18 runs   (    0.37
ms per token,  2732.66 tokens per second)
llama_print_timings: prompt eval time =    8853.68 ms /    83 tokens (  106.67
ms per token,     9.37 tokens per second)
llama_print_timings:        eval time =    3656.99 ms /    17 runs   (  215.12
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   12616.65 ms /   100 tokens
No. of rows:  36%| | 449/1258 [10:49:23<3:17:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.74 ms /    26 runs   (    0.37
ms per token,  2669.95 tokens per second)
llama_print_timings: prompt eval time =    9523.27 ms /    91 tokens (  104.65
ms per token,     9.56 tokens per second)
llama_print_timings:        eval time =    7096.31 ms /    25 runs   (  283.85
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   16781.63 ms /   116 tokens
No. of rows:  36%| | 450/1258 [10:49:39<3:25:59, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.00 ms /    20 runs   (    0.40
ms per token,  2500.00 tokens per second)
llama_print_timings: prompt eval time =    9276.78 ms /    85 tokens (  109.14
ms per token,     9.16 tokens per second)
llama_print_timings:        eval time =    4087.11 ms /    19 runs   (  215.11
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   13485.01 ms /   104 tokens
No. of rows:  36%| | 451/1258 [10:49:53<3:18:25, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.19 ms /    45 runs   (    0.38
ms per token,  2618.11 tokens per second)
llama_print_timings: prompt eval time =   13552.49 ms /   112 tokens (  121.00
ms per token,     8.26 tokens per second)
llama_print_timings:        eval time =    9495.15 ms /    44 runs   (  215.80
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   23316.50 ms /   156 tokens
No. of rows:  36%| | 452/1258 [10:50:16<3:52:43, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.86 ms /    24 runs   (    0.37
ms per token,  2708.80 tokens per second)
llama_print_timings: prompt eval time =    9017.22 ms /    86 tokens (  104.85
ms per token,     9.54 tokens per second)
llama_print_timings:        eval time =    4923.13 ms /    23 runs   (  214.05
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   14079.39 ms /   109 tokens
No. of rows:  36%| | 453/1258 [10:50:30<3:39:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.05 ms /    33 runs   (    0.37
ms per token,  2738.36 tokens per second)
llama_print_timings: prompt eval time =   10854.52 ms /   102 tokens (  106.42
ms per token,     9.40 tokens per second)
llama_print_timings:        eval time =    7198.63 ms /    32 runs   (  224.96
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   18247.25 ms /   134 tokens
No. of rows:  36%| | 454/1258 [10:50:48<3:46:46, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.19 ms /    24 runs   (    0.38
ms per token,  2612.96 tokens per second)
llama_print_timings: prompt eval time =   10534.75 ms /   101 tokens (  104.30
ms per token,     9.59 tokens per second)
llama_print_timings:        eval time =    4933.72 ms /    23 runs   (  214.51
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   15613.94 ms /   124 tokens
No. of rows:  36%| | 455/1258 [10:51:04<3:41:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.84 ms /    37 runs   (    0.37
```

```
ms per token,  2673.99 tokens per second)
llama_print_timings: prompt eval time =   13758.63 ms /   117 tokens (  117.60
ms per token,     8.50 tokens per second)
llama_print_timings:        eval time =    8800.61 ms /    36 runs   (  244.46
ms per token,     4.09 tokens per second)
llama_print_timings:       total time =   22777.57 ms /   153 tokens
No. of rows:  36%| | 456/1258 [10:51:27<4:06:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.00 ms /    21 runs   (    0.38
ms per token,  2626.64 tokens per second)
llama_print_timings: prompt eval time =    8220.18 ms /    78 tokens (  105.39
ms per token,     9.49 tokens per second)
llama_print_timings:        eval time =    4609.04 ms /    20 runs   (  230.45
ms per token,     4.34 tokens per second)
llama_print_timings:       total time =   12954.11 ms /    98 tokens
No. of rows:  36%| | 457/1258 [10:51:40<3:43:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.40 ms /    19 runs   (    0.39
ms per token,  2568.61 tokens per second)
llama_print_timings: prompt eval time =    9857.00 ms /    89 tokens (  110.75
ms per token,     9.03 tokens per second)
llama_print_timings:        eval time =    3869.01 ms /    18 runs   (  214.94
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   13842.03 ms /   107 tokens
No. of rows:  36%| | 458/1258 [10:51:54<3:31:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.23 ms /    20 runs   (    0.41
ms per token,  2430.13 tokens per second)
llama_print_timings: prompt eval time =    8404.86 ms /    79 tokens (  106.39
ms per token,     9.40 tokens per second)
llama_print_timings:        eval time =    4119.02 ms /    19 runs   (  216.79
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   12649.47 ms /    98 tokens
No. of rows:  36%| | 459/1258 [10:52:06<3:18:45, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.35 ms /    27 runs   (    0.38
ms per token,  2608.95 tokens per second)
llama_print_timings: prompt eval time =   11814.50 ms /   112 tokens (  105.49
ms per token,     9.48 tokens per second)
llama_print_timings:        eval time =    5644.95 ms /    26 runs   (  217.11
```

```
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   17622.18 ms /   138 tokens
No. of rows:  37%| | 460/1258 [10:52:24<3:29:18, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.71 ms /    23 runs   (    0.38
ms per token,  2640.34 tokens per second)
llama_print_timings: prompt eval time =    9144.00 ms /    86 tokens (  106.33
ms per token,     9.41 tokens per second)
llama_print_timings:        eval time =    4752.72 ms /    22 runs   (  216.03
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   14031.35 ms /   108 tokens
No. of rows:  37%| | 461/1258 [10:52:38<3:22:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.18 ms /    23 runs   (    0.40
ms per token,  2506.81 tokens per second)
llama_print_timings: prompt eval time =   10575.19 ms /   101 tokens (  104.70
ms per token,     9.55 tokens per second)
llama_print_timings:        eval time =    4738.65 ms /    22 runs   (  215.39
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   15455.15 ms /   123 tokens
No. of rows:  37%| | 462/1258 [10:52:53<3:22:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.04 ms /    19 runs   (    0.37
ms per token,  2696.95 tokens per second)
llama_print_timings: prompt eval time =   10162.73 ms /    81 tokens (  125.47
ms per token,     7.97 tokens per second)
llama_print_timings:        eval time =    3825.01 ms /    18 runs   (  212.50
ms per token,     4.71 tokens per second)
llama_print_timings:       total time =   14099.59 ms /    99 tokens
No. of rows:  37%| | 463/1258 [10:53:08<3:17:59, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.56 ms /    30 runs   (    0.39
ms per token,  2594.71 tokens per second)
llama_print_timings: prompt eval time =   11140.22 ms /   100 tokens (  111.40
ms per token,     8.98 tokens per second)
llama_print_timings:        eval time =    6237.77 ms /    29 runs   (  215.10
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   17556.19 ms /   129 tokens
No. of rows:  37%| | 464/1258 [10:53:25<3:28:06, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.00 ms /    16 runs   (    0.38
ms per token,  2666.67 tokens per second)
llama_print_timings: prompt eval time =    8038.57 ms /    77 tokens (  104.40
ms per token,     9.58 tokens per second)
llama_print_timings:        eval time =    3210.14 ms /    15 runs   (  214.01
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   11344.87 ms /    92 tokens
No. of rows:  37%| | 465/1258 [10:53:37<3:10:29, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.40 ms /    26 runs   (    0.36
ms per token,  2764.78 tokens per second)
llama_print_timings: prompt eval time =   10011.79 ms /    92 tokens (  108.82
ms per token,     9.19 tokens per second)
llama_print_timings:        eval time =    5388.01 ms /    25 runs   (  215.52
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   15551.26 ms /   117 tokens
No. of rows:  37%| | 466/1258 [10:53:52<3:14:48, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.31 ms /    18 runs   (    0.41
ms per token,  2461.03 tokens per second)
llama_print_timings: prompt eval time =   11043.35 ms /    88 tokens (  125.49
ms per token,     7.97 tokens per second)
llama_print_timings:        eval time =    3731.59 ms /    17 runs   (  219.51
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   14886.59 ms /   105 tokens
No. of rows:  37%| | 467/1258 [10:54:07<3:15:06, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.04 ms /    27 runs   (    0.37
ms per token,  2688.44 tokens per second)
llama_print_timings: prompt eval time =    9272.05 ms /    86 tokens (  107.81
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    5564.18 ms /    26 runs   (  214.01
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   14996.21 ms /   112 tokens
No. of rows:  37%| | 468/1258 [10:54:22<3:15:40, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.42 ms /    27 runs   (    0.39
ms per token,  2591.17 tokens per second)
```

```
llama_print_timings: prompt eval time =    9099.51 ms /    85 tokens (  107.05
ms per token,     9.34 tokens per second)
llama_print_timings:        eval time =    5577.23 ms /    26 runs   (  214.51
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   14841.85 ms /   111 tokens
No. of rows:  37%| | 469/1258 [10:54:37<3:15:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.52 ms /    28 runs   (    0.38
ms per token,  2660.84 tokens per second)
llama_print_timings: prompt eval time =   11409.20 ms /    93 tokens (  122.68
ms per token,     8.15 tokens per second)
llama_print_timings:        eval time =    6874.07 ms /    27 runs   (  254.60
ms per token,     3.93 tokens per second)
llama_print_timings:       total time =   18448.26 ms /   120 tokens
No. of rows:  37%| | 470/1258 [10:54:55<3:29:16, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.82 ms /    37 runs   (    0.37
ms per token,  2676.89 tokens per second)
llama_print_timings: prompt eval time =   13230.16 ms /   125 tokens (  105.84
ms per token,     9.45 tokens per second)
llama_print_timings:        eval time =    9316.47 ms /    36 runs   (  258.79
ms per token,     3.86 tokens per second)
llama_print_timings:       total time =   22768.82 ms /   161 tokens
No. of rows:  37%| | 471/1258 [10:55:18<3:55:55, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.37 ms /    29 runs   (    0.39
ms per token,  2550.80 tokens per second)
llama_print_timings: prompt eval time =   10011.64 ms /    96 tokens (  104.29
ms per token,     9.59 tokens per second)
llama_print_timings:        eval time =    6002.95 ms /    28 runs   (  214.39
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   16189.36 ms /   124 tokens
No. of rows:  38%| | 472/1258 [10:55:34<3:48:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.02 ms /    37 runs   (    0.38
ms per token,  2638.33 tokens per second)
llama_print_timings: prompt eval time =   12496.35 ms /   103 tokens (  121.32
ms per token,     8.24 tokens per second)
llama_print_timings:        eval time =    9450.63 ms /    36 runs   (  262.52
ms per token,     3.81 tokens per second)
```

```
llama_print_timings:       total time =   22168.03 ms /   139 tokens
No. of rows:  38%| | 473/1258 [10:55:56<4:06:54, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.89 ms /    31 runs   (    0.38
ms per token,  2606.79 tokens per second)
llama_print_timings: prompt eval time =   12581.97 ms /   119 tokens (  105.73
ms per token,     9.46 tokens per second)
llama_print_timings:        eval time =    6549.50 ms /    30 runs   (  218.32
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   19318.32 ms /   149 tokens
No. of rows:  38%| | 474/1258 [10:56:16<4:08:20, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.98 ms /    38 runs   (    0.37
ms per token,  2718.56 tokens per second)
llama_print_timings: prompt eval time =   12556.97 ms /   120 tokens (  104.64
ms per token,     9.56 tokens per second)
llama_print_timings:        eval time =    7954.89 ms /    37 runs   (  215.00
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   20737.36 ms /   157 tokens
No. of rows:  38%| | 475/1258 [10:56:36<4:14:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.75 ms /    26 runs   (    0.37
ms per token,  2667.21 tokens per second)
llama_print_timings: prompt eval time =   10451.18 ms /    97 tokens (  107.74
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    7092.20 ms /    25 runs   (  283.69
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   17698.02 ms /   122 tokens
No. of rows:  38%| | 476/1258 [10:56:54<4:07:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.95 ms /    28 runs   (    0.39
ms per token,  2558.25 tokens per second)
llama_print_timings: prompt eval time =    9791.05 ms /    92 tokens (  106.42
ms per token,     9.40 tokens per second)
llama_print_timings:        eval time =    7523.33 ms /    27 runs   (  278.64
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   17480.76 ms /   119 tokens
No. of rows:  38%| | 477/1258 [10:57:12<4:01:12, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.62 ms /    26 runs   (    0.37
ms per token,  2702.70 tokens per second)
llama_print_timings: prompt eval time =   10432.69 ms /    97 tokens (  107.55
ms per token,     9.30 tokens per second)
llama_print_timings:        eval time =    5356.78 ms /    25 runs   (  214.27
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   15943.92 ms /   122 tokens
No. of rows:  38%| | 478/1258 [10:57:28<3:50:52, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.13 ms /    26 runs   (    0.39
ms per token,  2567.65 tokens per second)
llama_print_timings: prompt eval time =   11766.73 ms /    97 tokens (  121.31
ms per token,     8.24 tokens per second)
llama_print_timings:        eval time =    7116.04 ms /    25 runs   (  284.64
ms per token,     3.51 tokens per second)
llama_print_timings:       total time =   19038.16 ms /   122 tokens
No. of rows:  38%| | 479/1258 [10:57:47<3:55:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.72 ms /    28 runs   (    0.38
ms per token,  2612.43 tokens per second)
llama_print_timings: prompt eval time =    9789.60 ms /    92 tokens (  106.41
ms per token,     9.40 tokens per second)
llama_print_timings:        eval time =    5827.94 ms /    27 runs   (  215.85
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   15783.95 ms /   119 tokens
No. of rows:  38%| | 480/1258 [10:58:02<3:46:05, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.80 ms /    28 runs   (    0.39
ms per token,  2591.87 tokens per second)
llama_print_timings: prompt eval time =   12561.13 ms /   119 tokens (  105.56
ms per token,     9.47 tokens per second)
llama_print_timings:        eval time =    5932.11 ms /    27 runs   (  219.71
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   18667.38 ms /   146 tokens
No. of rows:  38%| | 481/1258 [10:58:21<3:50:38, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.89 ms /    24 runs   (    0.37
ms per token,  2699.97 tokens per second)
llama_print_timings: prompt eval time =    8912.91 ms /    84 tokens (  106.11
```

```
ms per token,     9.42 tokens per second)
llama_print_timings:        eval time =    5865.40 ms /    23 runs   (  255.02
ms per token,     3.92 tokens per second)
llama_print_timings:       total time =   14919.84 ms /   107 tokens
No. of rows:  38%| | 482/1258 [10:58:36<3:39:09, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.88 ms /    24 runs   (    0.37
ms per token,  2703.31 tokens per second)
llama_print_timings: prompt eval time =    9110.39 ms /    86 tokens (  105.93
ms per token,     9.44 tokens per second)
llama_print_timings:        eval time =    4914.05 ms /    23 runs   (  213.65
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   14163.96 ms /   109 tokens
No. of rows:  38%| | 483/1258 [10:58:50<3:28:08, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.88 ms /    38 runs   (    0.37
ms per token,  2737.16 tokens per second)
llama_print_timings: prompt eval time =   12721.59 ms /   108 tokens (  117.79
ms per token,     8.49 tokens per second)
llama_print_timings:        eval time =    7950.70 ms /    37 runs   (  214.88
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   20901.12 ms /   145 tokens
No. of rows:  38%| | 484/1258 [10:59:11<3:46:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.43 ms /    24 runs   (    0.39
ms per token,  2546.42 tokens per second)
llama_print_timings: prompt eval time =   10979.81 ms /    90 tokens (  122.00
ms per token,     8.20 tokens per second)
llama_print_timings:        eval time =    4931.22 ms /    23 runs   (  214.40
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   16054.41 ms /   113 tokens
No. of rows:  39%| | 485/1258 [10:59:27<3:40:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.35 ms /    27 runs   (    0.38
ms per token,  2609.45 tokens per second)
llama_print_timings: prompt eval time =    9933.02 ms /    95 tokens (  104.56
ms per token,     9.56 tokens per second)
llama_print_timings:        eval time =    7326.66 ms /    26 runs   (  281.79
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   17425.67 ms /   121 tokens
```

```
No. of rows:  39%| | 486/1258 [10:59:45<3:41:18, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.54 ms /    29 runs   (    0.40
ms per token,  2513.43 tokens per second)
llama_print_timings: prompt eval time =   10706.39 ms /    98 tokens (  109.25
ms per token,     9.15 tokens per second)
llama_print_timings:        eval time =    7781.61 ms /    28 runs   (  277.91
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   18666.65 ms /   126 tokens
No. of rows:  39%| | 487/1258 [11:00:03<3:46:42, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.17 ms /    30 runs   (    0.37
ms per token,  2686.97 tokens per second)
llama_print_timings: prompt eval time =   10627.92 ms /    99 tokens (  107.35
ms per token,     9.32 tokens per second)
llama_print_timings:        eval time =    6289.23 ms /    29 runs   (  216.87
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   17098.90 ms /   128 tokens
No. of rows:  39%| | 488/1258 [11:00:20<3:44:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.75 ms /    26 runs   (    0.37
ms per token,  2667.49 tokens per second)
llama_print_timings: prompt eval time =    9959.12 ms /    95 tokens (  104.83
ms per token,     9.54 tokens per second)
llama_print_timings:        eval time =    5346.24 ms /    25 runs   (  213.85
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   15462.18 ms /   120 tokens
No. of rows:  39%| | 489/1258 [11:00:36<3:36:19, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.22 ms /    41 runs   (    0.37
ms per token,  2694.18 tokens per second)
llama_print_timings: prompt eval time =   12093.83 ms /   112 tokens (  107.98
ms per token,     9.26 tokens per second)
llama_print_timings:        eval time =    8585.25 ms /    40 runs   (  214.63
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   20924.70 ms /   152 tokens
No. of rows:  39%| | 490/1258 [11:00:57<3:51:39, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       7.79 ms /    21 runs   (    0.37
ms per token,  2695.42 tokens per second)
llama_print_timings: prompt eval time =    9579.49 ms /    87 tokens (  110.11
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    4285.95 ms /    20 runs   (  214.30
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   13991.95 ms /   107 tokens
No. of rows:  39%| | 491/1258 [11:01:11<3:35:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.34 ms /    31 runs   (    0.37
ms per token,  2732.72 tokens per second)
llama_print_timings: prompt eval time =   13375.89 ms /   112 tokens (  119.43
ms per token,     8.37 tokens per second)
llama_print_timings:        eval time =    6463.02 ms /    30 runs   (  215.43
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   20019.20 ms /   142 tokens
No. of rows:  39%| | 492/1258 [11:01:31<3:47:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.07 ms /    35 runs   (    0.40
ms per token,  2487.21 tokens per second)
llama_print_timings: prompt eval time =   12607.82 ms /   104 tokens (  121.23
ms per token,     8.25 tokens per second)
llama_print_timings:        eval time =    7475.26 ms /    34 runs   (  219.86
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   20298.73 ms /   138 tokens
No. of rows:  39%| | 493/1258 [11:01:51<3:56:39, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.54 ms /    31 runs   (    0.37
ms per token,  2686.77 tokens per second)
llama_print_timings: prompt eval time =   13908.17 ms /   114 tokens (  122.00
ms per token,     8.20 tokens per second)
llama_print_timings:        eval time =    6570.34 ms /    30 runs   (  219.01
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   20664.67 ms /   144 tokens
No. of rows:  39%| | 494/1258 [11:02:12<4:04:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.01 ms /    24 runs   (    0.38
ms per token,  2663.12 tokens per second)
llama_print_timings: prompt eval time =   11270.35 ms /   107 tokens (  105.33
ms per token,     9.49 tokens per second)
```

```
llama_print_timings:        eval time =    4927.78 ms /    23 runs   (  214.25
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   16339.11 ms /   130 tokens
No. of rows:  39%| | 495/1258 [11:02:28<3:53:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.14 ms /    20 runs   (    0.41
ms per token,  2456.40 tokens per second)
llama_print_timings: prompt eval time =    9052.60 ms /    84 tokens (  107.77
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    4727.21 ms /    19 runs   (  248.80
ms per token,     4.02 tokens per second)
llama_print_timings:       total time =   13901.31 ms /   103 tokens
No. of rows:  39%| | 496/1258 [11:02:42<3:36:00, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.58 ms /    23 runs   (    0.37
ms per token,  2680.97 tokens per second)
llama_print_timings: prompt eval time =    8802.30 ms /    82 tokens (  107.35
ms per token,     9.32 tokens per second)
llama_print_timings:        eval time =    4701.03 ms /    22 runs   (  213.68
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   13638.30 ms /   104 tokens
No. of rows:  40%| | 497/1258 [11:02:56<3:22:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.64 ms /    21 runs   (    0.36
ms per token,  2747.97 tokens per second)
llama_print_timings: prompt eval time =    9266.35 ms /    77 tokens (  120.34
ms per token,     8.31 tokens per second)
llama_print_timings:        eval time =    4275.59 ms /    20 runs   (  213.78
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   13665.44 ms /    97 tokens
No. of rows:  40%| | 498/1258 [11:03:09<3:13:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.85 ms /    22 runs   (    0.36
ms per token,  2801.48 tokens per second)
llama_print_timings: prompt eval time =    9812.50 ms /    92 tokens (  106.66
ms per token,     9.38 tokens per second)
llama_print_timings:        eval time =    4487.92 ms /    21 runs   (  213.71
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   14428.98 ms /   113 tokens
No. of rows:  40%| | 499/1258 [11:03:24<3:10:20, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.54 ms /    28 runs   (    0.38
ms per token,  2655.79 tokens per second)
llama_print_timings: prompt eval time =   10716.80 ms /    87 tokens (  123.18
ms per token,     8.12 tokens per second)
llama_print_timings:        eval time =    5807.43 ms /    27 runs   (  215.09
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   16691.06 ms /   114 tokens
No. of rows:  40%| | 500/1258 [11:03:41<3:16:19, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.65 ms /    20 runs   (    0.38
ms per token,  2614.38 tokens per second)
llama_print_timings: prompt eval time =   10585.46 ms /    82 tokens (  129.09
ms per token,     7.75 tokens per second)
llama_print_timings:        eval time =    4122.25 ms /    19 runs   (  216.96
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   14826.77 ms /   101 tokens
No. of rows:  40%| | 501/1258 [11:03:55<3:13:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.74 ms /    20 runs   (    0.39
ms per token,  2583.31 tokens per second)
llama_print_timings: prompt eval time =   10321.00 ms /    98 tokens (  105.32
ms per token,     9.50 tokens per second)
llama_print_timings:        eval time =    4051.01 ms /    19 runs   (  213.21
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   14491.28 ms /   117 tokens
No. of rows:  40%| | 502/1258 [11:04:10<3:09:58, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.60 ms /    20 runs   (    0.38
ms per token,  2631.58 tokens per second)
llama_print_timings: prompt eval time =    9795.43 ms /    77 tokens (  127.21
ms per token,     7.86 tokens per second)
llama_print_timings:        eval time =    4056.76 ms /    19 runs   (  213.51
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   13970.38 ms /    96 tokens
No. of rows:  40%| | 503/1258 [11:04:24<3:05:33, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.34 ms /    36 runs   (    0.40
```

```
ms per token,  2511.16 tokens per second)
llama_print_timings: prompt eval time =   14485.10 ms /   139 tokens (  104.21
ms per token,     9.60 tokens per second)
llama_print_timings:        eval time =    9515.42 ms /    35 runs   (  271.87
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   24225.34 ms /   174 tokens
No. of rows:  40%| | 504/1258 [11:04:48<3:41:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.01 ms /    48 runs   (    0.38
ms per token,  2665.78 tokens per second)
llama_print_timings: prompt eval time =   15748.11 ms /   150 tokens (  104.99
ms per token,     9.52 tokens per second)
llama_print_timings:        eval time =   10184.41 ms /    47 runs   (  216.69
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   26216.49 ms /   197 tokens
No. of rows:  40%| | 505/1258 [11:05:14<4:13:18, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.26 ms /    33 runs   (    0.40
ms per token,  2489.44 tokens per second)
llama_print_timings: prompt eval time =   12853.63 ms /   109 tokens (  117.92
ms per token,     8.48 tokens per second)
llama_print_timings:        eval time =    7011.24 ms /    32 runs   (  219.10
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   20069.20 ms /   141 tokens
No. of rows:  40%| | 506/1258 [11:05:34<4:12:34, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.92 ms /    21 runs   (    0.38
ms per token,  2650.51 tokens per second)
llama_print_timings: prompt eval time =    9703.48 ms /    77 tokens (  126.02
ms per token,     7.94 tokens per second)
llama_print_timings:        eval time =    4792.74 ms /    20 runs   (  239.64
ms per token,     4.17 tokens per second)
llama_print_timings:       total time =   14629.91 ms /    97 tokens
No. of rows:  40%| | 507/1258 [11:05:49<3:51:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.59 ms /    23 runs   (    0.37
ms per token,  2677.84 tokens per second)
llama_print_timings: prompt eval time =   10512.67 ms /    99 tokens (  106.19
ms per token,     9.42 tokens per second)
llama_print_timings:        eval time =    4744.77 ms /    22 runs   (  215.67
```

```
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   15393.23 ms /   121 tokens
No. of rows:  40%| | 508/1258 [11:06:04<3:39:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.26 ms /    34 runs   (    0.39
ms per token,  2564.68 tokens per second)
llama_print_timings: prompt eval time =   13683.41 ms /   113 tokens (  121.09
ms per token,     8.26 tokens per second)
llama_print_timings:        eval time =    7068.87 ms /    33 runs   (  214.21
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   20955.08 ms /   146 tokens
No. of rows:  40%| | 509/1258 [11:06:25<3:52:00, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.18 ms /    19 runs   (    0.38
ms per token,  2645.50 tokens per second)
llama_print_timings: prompt eval time =    9051.98 ms /    86 tokens (  105.26
ms per token,     9.50 tokens per second)
llama_print_timings:        eval time =    3891.93 ms /    18 runs   (  216.22
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   13058.44 ms /   104 tokens
No. of rows:  41%| | 510/1258 [11:06:38<3:31:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.44 ms /    33 runs   (    0.38
ms per token,  2652.31 tokens per second)
llama_print_timings: prompt eval time =   11515.36 ms /   110 tokens (  104.69
ms per token,     9.55 tokens per second)
llama_print_timings:        eval time =    8560.47 ms /    32 runs   (  267.51
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   20274.22 ms /   142 tokens
No. of rows:  41%| | 511/1258 [11:06:59<3:43:19, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.53 ms /    25 runs   (    0.38
ms per token,  2624.40 tokens per second)
llama_print_timings: prompt eval time =    9844.76 ms /    91 tokens (  108.18
ms per token,     9.24 tokens per second)
llama_print_timings:        eval time =    6803.99 ms /    24 runs   (  283.50
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   16801.19 ms /   115 tokens
No. of rows:  41%| | 512/1258 [11:07:16<3:38:47, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.19 ms /    17 runs   (    0.42
ms per token,  2365.05 tokens per second)
llama_print_timings: prompt eval time =   10357.80 ms /    95 tokens (  109.03
ms per token,     9.17 tokens per second)
llama_print_timings:        eval time =    3834.36 ms /    16 runs   (  239.65
ms per token,     4.17 tokens per second)
llama_print_timings:       total time =   14308.81 ms /   111 tokens
No. of rows:  41%| | 513/1258 [11:07:30<3:26:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.57 ms /    23 runs   (    0.37
ms per token,  2684.41 tokens per second)
llama_print_timings: prompt eval time =   10296.69 ms /    83 tokens (  124.06
ms per token,     8.06 tokens per second)
llama_print_timings:        eval time =    4792.40 ms /    22 runs   (  217.84
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   15225.41 ms /   105 tokens
No. of rows:  41%| | 514/1258 [11:07:45<3:20:53, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.99 ms /    28 runs   (    0.46
ms per token,  2156.17 tokens per second)
llama_print_timings: prompt eval time =   11804.35 ms /   109 tokens (  108.30
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =    7554.66 ms /    27 runs   (  279.80
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   19529.82 ms /   136 tokens
No. of rows:  41%| | 515/1258 [11:08:05<3:32:58, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.21 ms /    28 runs   (    0.40
ms per token,  2498.22 tokens per second)
llama_print_timings: prompt eval time =   11971.64 ms /   112 tokens (  106.89
ms per token,     9.36 tokens per second)
llama_print_timings:        eval time =    5862.96 ms /    27 runs   (  217.15
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   18004.17 ms /   139 tokens
No. of rows:  41%| | 516/1258 [11:08:23<3:35:45, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.93 ms /    29 runs   (    0.38
ms per token,  2653.73 tokens per second)
```

```
llama_print_timings: prompt eval time =   10561.21 ms /   101 tokens (  104.57
ms per token,     9.56 tokens per second)
llama_print_timings:        eval time =    6071.69 ms /    28 runs   (  216.85
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   16810.46 ms /   129 tokens
No. of rows:  41%| | 517/1258 [11:08:39<3:33:05, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.82 ms /    33 runs   (    0.39
ms per token,  2574.71 tokens per second)
llama_print_timings: prompt eval time =   13594.94 ms /   129 tokens (  105.39
ms per token,     9.49 tokens per second)
llama_print_timings:        eval time =    6904.12 ms /    32 runs   (  215.75
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   20697.94 ms /   161 tokens
No. of rows:  41%| | 518/1258 [11:09:00<3:45:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.50 ms /    21 runs   (    0.36
ms per token,  2800.00 tokens per second)
llama_print_timings: prompt eval time =   10576.06 ms /    86 tokens (  122.98
ms per token,     8.13 tokens per second)
llama_print_timings:        eval time =    4274.44 ms /    20 runs   (  213.72
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   14975.50 ms /   106 tokens
No. of rows:  41%| | 519/1258 [11:09:15<3:33:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.90 ms /    24 runs   (    0.37
ms per token,  2696.63 tokens per second)
llama_print_timings: prompt eval time =   10598.09 ms /    89 tokens (  119.08
ms per token,     8.40 tokens per second)
llama_print_timings:        eval time =    4909.34 ms /    23 runs   (  213.45
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   15649.80 ms /   112 tokens
No. of rows:  41%| | 520/1258 [11:09:31<3:26:42, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.00 ms /    23 runs   (    0.39
ms per token,  2555.27 tokens per second)
llama_print_timings: prompt eval time =   11597.22 ms /    96 tokens (  120.80
ms per token,     8.28 tokens per second)
llama_print_timings:        eval time =    6401.09 ms /    22 runs   (  290.96
ms per token,     3.44 tokens per second)
```

```
llama_print_timings:       total time =   18136.20 ms /   118 tokens
No. of rows:  41%| | 521/1258 [11:09:49<3:31:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.07 ms /    32 runs   (    0.38
ms per token,  2651.20 tokens per second)
llama_print_timings: prompt eval time =   10923.28 ms /   102 tokens (  107.09
ms per token,     9.34 tokens per second)
llama_print_timings:        eval time =    6654.29 ms /    31 runs   (  214.65
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   17769.51 ms /   133 tokens
No. of rows:  41%| | 522/1258 [11:10:07<3:33:11, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.94 ms /    50 runs   (    0.38
ms per token,  2639.78 tokens per second)
llama_print_timings: prompt eval time =   14967.41 ms /   127 tokens (  117.85
ms per token,     8.49 tokens per second)
llama_print_timings:        eval time =   10571.94 ms /    49 runs   (  215.75
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   25847.38 ms /   176 tokens
No. of rows:  42%| | 523/1258 [11:10:33<4:04:03, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.06 ms /    23 runs   (    0.39
ms per token,  2539.47 tokens per second)
llama_print_timings: prompt eval time =    8524.03 ms /    80 tokens (  106.55
ms per token,     9.39 tokens per second)
llama_print_timings:        eval time =    4700.24 ms /    22 runs   (  213.65
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   13363.47 ms /   102 tokens
No. of rows:  42%| | 524/1258 [11:10:46<3:39:38, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.58 ms /    38 runs   (    0.38
ms per token,  2607.03 tokens per second)
llama_print_timings: prompt eval time =   14821.92 ms /   130 tokens (  114.01
ms per token,     8.77 tokens per second)
llama_print_timings:        eval time =    7974.00 ms /    37 runs   (  215.51
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   23025.50 ms /   167 tokens
No. of rows:  42%| | 525/1258 [11:11:09<3:57:55, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.42 ms /    19 runs   (    0.39
ms per token,  2561.34 tokens per second)
llama_print_timings: prompt eval time =    8816.10 ms /    85 tokens (  103.72
ms per token,     9.64 tokens per second)
llama_print_timings:        eval time =    3893.60 ms /    18 runs   (  216.31
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   12823.63 ms /   103 tokens
No. of rows:  42%| | 526/1258 [11:11:22<3:33:18, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.68 ms /    22 runs   (    0.39
ms per token,  2534.85 tokens per second)
llama_print_timings: prompt eval time =   11595.62 ms /    96 tokens (  120.79
ms per token,     8.28 tokens per second)
llama_print_timings:        eval time =    5950.49 ms /    21 runs   (  283.36
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   17678.49 ms /   117 tokens
No. of rows:  42%| | 527/1258 [11:11:40<3:33:46, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.65 ms /    24 runs   (    0.36
ms per token,  2775.21 tokens per second)
llama_print_timings: prompt eval time =    9480.57 ms /    87 tokens (  108.97
ms per token,     9.18 tokens per second)
llama_print_timings:        eval time =    4931.70 ms /    23 runs   (  214.42
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   14551.43 ms /   110 tokens
No. of rows:  42%| | 528/1258 [11:11:54<3:22:31, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.74 ms /    21 runs   (    0.37
ms per token,  2712.48 tokens per second)
llama_print_timings: prompt eval time =   10907.85 ms /    87 tokens (  125.38
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =    4276.84 ms /    20 runs   (  213.84
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   15307.42 ms /   107 tokens
No. of rows:  42%| | 529/1258 [11:12:09<3:17:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.10 ms /    23 runs   (    0.40
ms per token,  2526.64 tokens per second)
llama_print_timings: prompt eval time =   10670.34 ms /    86 tokens (  124.07
```

```
ms per token,     8.06 tokens per second)
llama_print_timings:        eval time =    5008.39 ms /    22 runs   (  227.65
ms per token,     4.39 tokens per second)
llama_print_timings:       total time =   15823.83 ms /   108 tokens
No. of rows:  42%| | 530/1258 [11:12:25<3:15:37, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.54 ms /    20 runs   (    0.38
ms per token,  2651.46 tokens per second)
llama_print_timings: prompt eval time =    9486.16 ms /    88 tokens (  107.80
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    4091.72 ms /    19 runs   (  215.35
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   13695.97 ms /   107 tokens
No. of rows:  42%| | 531/1258 [11:12:39<3:06:34, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.71 ms /    44 runs   (    0.38
ms per token,  2633.94 tokens per second)
llama_print_timings: prompt eval time =   11754.35 ms /   108 tokens (  108.84
ms per token,     9.19 tokens per second)
llama_print_timings:        eval time =   11180.43 ms /    43 runs   (  260.01
ms per token,     3.85 tokens per second)
llama_print_timings:       total time =   23198.51 ms /   151 tokens
No. of rows:  42%| | 532/1258 [11:13:02<3:34:39, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.68 ms /    37 runs   (    0.40
ms per token,  2520.26 tokens per second)
llama_print_timings: prompt eval time =   13045.05 ms /   124 tokens (  105.20
ms per token,     9.51 tokens per second)
llama_print_timings:        eval time =    7800.30 ms /    36 runs   (  216.67
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   21076.01 ms /   160 tokens
No. of rows:  42%| | 533/1258 [11:13:23<3:46:29, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.18 ms /    21 runs   (    0.39
ms per token,  2567.24 tokens per second)
llama_print_timings: prompt eval time =   10951.75 ms /    86 tokens (  127.35
ms per token,     7.85 tokens per second)
llama_print_timings:        eval time =    4272.00 ms /    20 runs   (  213.60
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   15348.99 ms /   106 tokens
```

```
No. of rows:  42%| | 534/1258 [11:13:39<3:33:54, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.26 ms /    29 runs   (    0.39
ms per token,  2575.95 tokens per second)
llama_print_timings: prompt eval time =   12229.64 ms /   102 tokens (  119.90
ms per token,     8.34 tokens per second)
llama_print_timings:        eval time =    6379.37 ms /    28 runs   (  227.83
ms per token,     4.39 tokens per second)
llama_print_timings:       total time =   18781.11 ms /   130 tokens
No. of rows:  43%| | 535/1258 [11:13:57<3:37:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.40 ms /    23 runs   (    0.37
ms per token,  2739.07 tokens per second)
llama_print_timings: prompt eval time =   10389.80 ms /    97 tokens (  107.11
ms per token,     9.34 tokens per second)
llama_print_timings:        eval time =    4675.23 ms /    22 runs   (  212.51
ms per token,     4.71 tokens per second)
llama_print_timings:       total time =   15197.93 ms /   119 tokens
No. of rows:  43%| | 536/1258 [11:14:13<3:26:52, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.14 ms /    31 runs   (    0.42
ms per token,  2358.67 tokens per second)
llama_print_timings: prompt eval time =   12580.68 ms /   104 tokens (  120.97
ms per token,     8.27 tokens per second)
llama_print_timings:        eval time =    7726.86 ms /    30 runs   (  257.56
ms per token,     3.88 tokens per second)
llama_print_timings:       total time =   20527.33 ms /   134 tokens
No. of rows:  43%| | 537/1258 [11:14:33<3:38:38, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.63 ms /    26 runs   (    0.37
ms per token,  2698.50 tokens per second)
llama_print_timings: prompt eval time =   11043.55 ms /    99 tokens (  111.55
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =    5356.62 ms /    25 runs   (  214.26
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   16552.76 ms /   124 tokens
No. of rows:  43%| | 538/1258 [11:14:50<3:32:29, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       7.30 ms /    20 runs   (    0.36
ms per token,  2740.48 tokens per second)
llama_print_timings: prompt eval time =   10079.23 ms /    80 tokens (  125.99
ms per token,     7.94 tokens per second)
llama_print_timings:        eval time =    4071.62 ms /    19 runs   (  214.30
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   14269.26 ms /    99 tokens
No. of rows:  43%| | 539/1258 [11:15:04<3:19:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.35 ms /    21 runs   (    0.40
ms per token,  2514.37 tokens per second)
llama_print_timings: prompt eval time =    9663.19 ms /    86 tokens (  112.36
ms per token,     8.90 tokens per second)
llama_print_timings:        eval time =    4880.86 ms /    20 runs   (  244.04
ms per token,     4.10 tokens per second)
llama_print_timings:       total time =   14667.55 ms /   106 tokens
No. of rows:  43%| | 540/1258 [11:15:19<3:12:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.67 ms /    26 runs   (    0.37
ms per token,  2689.84 tokens per second)
llama_print_timings: prompt eval time =   11265.74 ms /   105 tokens (  107.29
ms per token,     9.32 tokens per second)
llama_print_timings:        eval time =    5369.12 ms /    25 runs   (  214.76
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   16789.27 ms /   130 tokens
No. of rows:  43%| | 541/1258 [11:15:35<3:14:41, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.34 ms /    36 runs   (    0.37
ms per token,  2698.25 tokens per second)
llama_print_timings: prompt eval time =   13991.46 ms /   115 tokens (  121.66
ms per token,     8.22 tokens per second)
llama_print_timings:        eval time =    7581.92 ms /    35 runs   (  216.63
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   21787.73 ms /   150 tokens
No. of rows:  43%| | 542/1258 [11:15:57<3:34:08, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.22 ms /    22 runs   (    0.42
ms per token,  2386.89 tokens per second)
llama_print_timings: prompt eval time =   12862.19 ms /   106 tokens (  121.34
ms per token,     8.24 tokens per second)
```

```
llama_print_timings:        eval time =    6263.06 ms /    21 runs   (  298.24
ms per token,     3.35 tokens per second)
llama_print_timings:       total time =   19257.75 ms /   127 tokens
No. of rows:  43%| | 543/1258 [11:16:16<3:38:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.29 ms /    32 runs   (    0.38
ms per token,  2603.74 tokens per second)
llama_print_timings: prompt eval time =   11511.81 ms /   104 tokens (  110.69
ms per token,     9.03 tokens per second)
llama_print_timings:        eval time =    6651.96 ms /    31 runs   (  214.58
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   18357.48 ms /   135 tokens
No. of rows:  43%| | 544/1258 [11:16:35<3:38:20, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.72 ms /    21 runs   (    0.37
ms per token,  2721.26 tokens per second)
llama_print_timings: prompt eval time =    9441.47 ms /    85 tokens (  111.08
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =    4291.00 ms /    20 runs   (  214.55
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   13856.62 ms /   105 tokens
No. of rows:  43%| | 545/1258 [11:16:49<3:22:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.98 ms /    24 runs   (    0.37
ms per token,  2672.61 tokens per second)
llama_print_timings: prompt eval time =   10774.94 ms /    89 tokens (  121.07
ms per token,     8.26 tokens per second)
llama_print_timings:        eval time =    6303.10 ms /    23 runs   (  274.05
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   17224.71 ms /   112 tokens
No. of rows:  43%| | 546/1258 [11:17:06<3:22:34, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.41 ms /    29 runs   (    0.39
ms per token,  2541.18 tokens per second)
llama_print_timings: prompt eval time =    9881.34 ms /    91 tokens (  108.59
ms per token,     9.21 tokens per second)
llama_print_timings:        eval time =    6083.30 ms /    28 runs   (  217.26
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   16144.82 ms /   119 tokens
No. of rows:  43%| | 547/1258 [11:17:22<3:19:01, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.52 ms /    23 runs   (    0.37
ms per token,  2700.48 tokens per second)
llama_print_timings: prompt eval time =   12017.78 ms /    97 tokens (  123.89
ms per token,     8.07 tokens per second)
llama_print_timings:        eval time =    4836.08 ms /    22 runs   (  219.82
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   16991.58 ms /   119 tokens
No. of rows:  44%| | 548/1258 [11:17:39<3:19:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.20 ms /    24 runs   (    0.38
ms per token,  2608.70 tokens per second)
llama_print_timings: prompt eval time =   11411.05 ms /    92 tokens (  124.03
ms per token,     8.06 tokens per second)
llama_print_timings:        eval time =    6442.93 ms /    23 runs   (  280.13
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   17996.11 ms /   115 tokens
No. of rows:  44%| | 549/1258 [11:17:57<3:23:12, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.63 ms /    23 runs   (    0.38
ms per token,  2666.05 tokens per second)
llama_print_timings: prompt eval time =    8936.00 ms /    81 tokens (  110.32
ms per token,     9.06 tokens per second)
llama_print_timings:        eval time =    4729.53 ms /    22 runs   (  214.98
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   13801.96 ms /   103 tokens
No. of rows:  44%| | 550/1258 [11:18:11<3:10:58, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.12 ms /    24 runs   (    0.38
ms per token,  2633.02 tokens per second)
llama_print_timings: prompt eval time =   11187.65 ms /    91 tokens (  122.94
ms per token,     8.13 tokens per second)
llama_print_timings:        eval time =    4925.66 ms /    23 runs   (  214.16
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   16255.35 ms /   114 tokens
No. of rows:  44%| | 551/1258 [11:18:27<3:10:58, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.86 ms /    19 runs   (    0.36
```

```
ms per token,  2768.47 tokens per second)
llama_print_timings: prompt eval time =    8882.45 ms /    80 tokens (  111.03
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    3872.06 ms /    18 runs   (  215.11
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   12866.98 ms /    98 tokens
No. of rows:  44%| | 552/1258 [11:18:40<2:58:53, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.79 ms /    23 runs   (    0.38
ms per token,  2617.80 tokens per second)
llama_print_timings: prompt eval time =   10406.00 ms /    92 tokens (  113.11
ms per token,     8.84 tokens per second)
llama_print_timings:        eval time =    4750.08 ms /    22 runs   (  215.91
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   15297.61 ms /   114 tokens
No. of rows:  44%| | 553/1258 [11:18:55<2:59:00, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.29 ms /    29 runs   (    0.39
ms per token,  2569.33 tokens per second)
llama_print_timings: prompt eval time =   11149.67 ms /    99 tokens (  112.62
ms per token,     8.88 tokens per second)
llama_print_timings:        eval time =    6211.50 ms /    28 runs   (  221.84
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   17542.44 ms /   127 tokens
No. of rows:  44%| | 554/1258 [11:19:13<3:06:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.49 ms /    29 runs   (    0.40
ms per token,  2523.28 tokens per second)
llama_print_timings: prompt eval time =   11415.98 ms /   107 tokens (  106.69
ms per token,     9.37 tokens per second)
llama_print_timings:        eval time =    6072.29 ms /    28 runs   (  216.87
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   17665.13 ms /   135 tokens
No. of rows:  44%| | 555/1258 [11:19:31<3:12:47, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.00 ms /    24 runs   (    0.37
ms per token,  2668.15 tokens per second)
llama_print_timings: prompt eval time =   11851.77 ms /    94 tokens (  126.08
ms per token,     7.93 tokens per second)
llama_print_timings:        eval time =    5046.94 ms /    23 runs   (  219.43
```

```
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   17044.68 ms /   117 tokens
No. of rows:  44%| | 556/1258 [11:19:48<3:14:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.56 ms /    20 runs   (    0.38
ms per token,  2644.80 tokens per second)
llama_print_timings: prompt eval time =   11487.49 ms /    92 tokens (  124.86
ms per token,     8.01 tokens per second)
llama_print_timings:        eval time =    5370.31 ms /    19 runs   (  282.65
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   16977.02 ms /   111 tokens
No. of rows:  44%| | 557/1258 [11:20:05<3:15:33, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.42 ms /    33 runs   (    0.41
ms per token,  2458.65 tokens per second)
llama_print_timings: prompt eval time =   12493.46 ms /   116 tokens (  107.70
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    7222.96 ms /    32 runs   (  225.72
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   19929.13 ms /   148 tokens
No. of rows:  44%| | 558/1258 [11:20:25<3:26:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.26 ms /    30 runs   (    0.38
ms per token,  2664.77 tokens per second)
llama_print_timings: prompt eval time =   10649.83 ms /    99 tokens (  107.57
ms per token,     9.30 tokens per second)
llama_print_timings:        eval time =    6258.30 ms /    29 runs   (  215.80
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   17085.35 ms /   128 tokens
No. of rows:  44%| | 559/1258 [11:20:42<3:24:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.29 ms /    37 runs   (    0.39
ms per token,  2588.68 tokens per second)
llama_print_timings: prompt eval time =   12209.16 ms /   114 tokens (  107.10
ms per token,     9.34 tokens per second)
llama_print_timings:        eval time =    8035.57 ms /    36 runs   (  223.21
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   20465.62 ms /   150 tokens
No. of rows:  45%| | 560/1258 [11:21:02<3:34:06, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.65 ms /    20 runs   (    0.38
ms per token,  2615.40 tokens per second)
llama_print_timings: prompt eval time =    9374.86 ms /    73 tokens (  128.42
ms per token,     7.79 tokens per second)
llama_print_timings:        eval time =    4076.37 ms /    19 runs   (  214.55
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   13568.27 ms /    92 tokens
No. of rows:  45%| | 561/1258 [11:21:16<3:16:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.15 ms /    25 runs   (    0.37
ms per token,  2732.54 tokens per second)
llama_print_timings: prompt eval time =   11893.55 ms /   108 tokens (  110.13
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    5244.44 ms /    24 runs   (  218.52
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   17286.88 ms /   132 tokens
No. of rows:  45%| | 562/1258 [11:21:33<3:17:52, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.03 ms /    22 runs   (    0.37
ms per token,  2739.04 tokens per second)
llama_print_timings: prompt eval time =    9513.68 ms /    87 tokens (  109.35
ms per token,     9.14 tokens per second)
llama_print_timings:        eval time =    6180.66 ms /    21 runs   (  294.32
ms per token,     3.40 tokens per second)
llama_print_timings:       total time =   15823.00 ms /   108 tokens
No. of rows:  45%| | 563/1258 [11:21:49<3:13:19, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.80 ms /    36 runs   (    0.38
ms per token,  2608.51 tokens per second)
llama_print_timings: prompt eval time =   13388.05 ms /   120 tokens (  111.57
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =    7637.44 ms /    35 runs   (  218.21
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   21246.84 ms /   155 tokens
No. of rows:  45%| | 564/1258 [11:22:10<3:28:52, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.11 ms /    38 runs   (    0.37
ms per token,  2692.55 tokens per second)
```

```
llama_print_timings: prompt eval time =   13618.91 ms /   111 tokens (  122.69
ms per token,     8.15 tokens per second)
llama_print_timings:        eval time =    8276.88 ms /    37 runs   (  223.70
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   22122.66 ms /   148 tokens
No. of rows:  45%| | 565/1258 [11:22:32<3:42:40, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.82 ms /    31 runs   (    0.38
ms per token,  2622.01 tokens per second)
llama_print_timings: prompt eval time =   10132.41 ms /    93 tokens (  108.95
ms per token,     9.18 tokens per second)
llama_print_timings:        eval time =    8229.26 ms /    30 runs   (  274.31
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   18546.37 ms /   123 tokens
No. of rows:  45%| | 566/1258 [11:22:51<3:39:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.40 ms /    21 runs   (    0.40
ms per token,  2500.60 tokens per second)
llama_print_timings: prompt eval time =    8461.35 ms /    79 tokens (  107.11
ms per token,     9.34 tokens per second)
llama_print_timings:        eval time =    4651.60 ms /    20 runs   (  232.58
ms per token,     4.30 tokens per second)
llama_print_timings:       total time =   13242.37 ms /    99 tokens
No. of rows:  45%| | 567/1258 [11:23:04<3:19:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.17 ms /    25 runs   (    0.37
ms per token,  2726.58 tokens per second)
llama_print_timings: prompt eval time =    9210.45 ms /    85 tokens (  108.36
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =    5150.59 ms /    24 runs   (  214.61
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   14506.79 ms /   109 tokens
No. of rows:  45%| | 568/1258 [11:23:18<3:09:29, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.43 ms /    33 runs   (    0.38
ms per token,  2655.94 tokens per second)
llama_print_timings: prompt eval time =   12641.69 ms /   105 tokens (  120.40
ms per token,     8.31 tokens per second)
llama_print_timings:        eval time =    7291.65 ms /    32 runs   (  227.86
ms per token,     4.39 tokens per second)
```

```
llama_print_timings:       total time =   20128.82 ms /   137 tokens
No. of rows:  45%| | 569/1258 [11:23:39<3:21:48, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.69 ms /    22 runs   (    0.40
ms per token,  2530.48 tokens per second)
llama_print_timings: prompt eval time =    9324.77 ms /    86 tokens (  108.43
ms per token,     9.22 tokens per second)
llama_print_timings:        eval time =    4666.78 ms /    21 runs   (  222.23
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   14127.12 ms /   107 tokens
No. of rows:  45%| | 570/1258 [11:23:53<3:09:42, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.81 ms /    35 runs   (    0.37
ms per token,  2732.67 tokens per second)
llama_print_timings: prompt eval time =   12898.18 ms /   106 tokens (  121.68
ms per token,     8.22 tokens per second)
llama_print_timings:        eval time =    7312.13 ms /    34 runs   (  215.06
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   20420.54 ms /   140 tokens
No. of rows:  45%| | 571/1258 [11:24:13<3:22:46, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.92 ms /    25 runs   (    0.40
ms per token,  2519.65 tokens per second)
llama_print_timings: prompt eval time =    9312.33 ms /    84 tokens (  110.86
ms per token,     9.02 tokens per second)
llama_print_timings:        eval time =    5294.14 ms /    24 runs   (  220.59
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   14758.97 ms /   108 tokens
No. of rows:  45%| | 572/1258 [11:24:28<3:12:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.56 ms /    31 runs   (    0.37
ms per token,  2680.50 tokens per second)
llama_print_timings: prompt eval time =   10639.51 ms /    98 tokens (  108.57
ms per token,     9.21 tokens per second)
llama_print_timings:        eval time =    6485.82 ms /    30 runs   (  216.19
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   17311.97 ms /   128 tokens
No. of rows:  46%| | 573/1258 [11:24:45<3:13:46, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.53 ms /    22 runs   (    0.39
ms per token,  2580.04 tokens per second)
llama_print_timings: prompt eval time =    9475.53 ms /    88 tokens (  107.68
ms per token,     9.29 tokens per second)
llama_print_timings:        eval time =    5489.76 ms /    21 runs   (  261.42
ms per token,     3.83 tokens per second)
llama_print_timings:       total time =   15096.50 ms /   109 tokens
No. of rows:  46%| | 574/1258 [11:25:00<3:07:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.61 ms /    34 runs   (    0.37
ms per token,  2696.06 tokens per second)
llama_print_timings: prompt eval time =   12428.72 ms /   114 tokens (  109.02
ms per token,     9.17 tokens per second)
llama_print_timings:        eval time =    8776.41 ms /    33 runs   (  265.95
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   21406.16 ms /   147 tokens
No. of rows:  46%| | 575/1258 [11:25:22<3:23:55, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.03 ms /    24 runs   (    0.38
ms per token,  2658.69 tokens per second)
llama_print_timings: prompt eval time =    9221.27 ms /    84 tokens (  109.78
ms per token,     9.11 tokens per second)
llama_print_timings:        eval time =    4965.62 ms /    23 runs   (  215.90
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   14333.77 ms /   107 tokens
No. of rows:  46%| | 576/1258 [11:25:36<3:11:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.09 ms /    19 runs   (    0.37
ms per token,  2680.21 tokens per second)
llama_print_timings: prompt eval time =    9801.75 ms /    78 tokens (  125.66
ms per token,     7.96 tokens per second)
llama_print_timings:        eval time =    3869.49 ms /    18 runs   (  214.97
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   13781.77 ms /    96 tokens
No. of rows:  46%| | 577/1258 [11:25:50<3:00:45, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.66 ms /    25 runs   (    0.39
ms per token,  2586.92 tokens per second)
llama_print_timings: prompt eval time =   11020.43 ms /    97 tokens (  113.61
```

```
ms per token,     8.80 tokens per second)
llama_print_timings:        eval time =    6890.78 ms /    24 runs   (  287.12
ms per token,     3.48 tokens per second)
llama_print_timings:       total time =   18065.23 ms /   121 tokens
No. of rows:  46%| | 578/1258 [11:26:08<3:07:48, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.02 ms /    22 runs   (    0.41
ms per token,  2439.02 tokens per second)
llama_print_timings: prompt eval time =    9901.32 ms /    91 tokens (  108.81
ms per token,     9.19 tokens per second)
llama_print_timings:        eval time =    5099.35 ms /    21 runs   (  242.83
ms per token,     4.12 tokens per second)
llama_print_timings:       total time =   15139.94 ms /   112 tokens
No. of rows:  46%| | 579/1258 [11:26:23<3:02:42, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.23 ms /    20 runs   (    0.36
ms per token,  2766.25 tokens per second)
llama_print_timings: prompt eval time =    9971.95 ms /    91 tokens (  109.58
ms per token,     9.13 tokens per second)
llama_print_timings:        eval time =    4096.86 ms /    19 runs   (  215.62
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14188.70 ms /   110 tokens
No. of rows:  46%| | 580/1258 [11:26:37<2:55:50, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.28 ms /    28 runs   (    0.37
ms per token,  2725.06 tokens per second)
llama_print_timings: prompt eval time =   11426.11 ms /   102 tokens (  112.02
ms per token,     8.93 tokens per second)
llama_print_timings:        eval time =    7496.95 ms /    27 runs   (  277.66
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   19088.38 ms /   129 tokens
No. of rows:  46%| | 581/1258 [11:26:56<3:07:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.97 ms /    27 runs   (    0.37
ms per token,  2707.85 tokens per second)
llama_print_timings: prompt eval time =   11819.45 ms /   109 tokens (  108.44
ms per token,     9.22 tokens per second)
llama_print_timings:        eval time =    5600.76 ms /    26 runs   (  215.41
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   17578.12 ms /   135 tokens
```

```
No. of rows:  46%| | 582/1258 [11:27:14<3:10:31, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.81 ms /    21 runs   (    0.37
ms per token,  2688.52 tokens per second)
llama_print_timings: prompt eval time =   10545.80 ms /    84 tokens (  125.55
ms per token,     7.97 tokens per second)
llama_print_timings:        eval time =    4301.89 ms /    20 runs   (  215.09
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   14973.92 ms /   104 tokens
No. of rows:  46%| | 583/1258 [11:27:29<3:03:43, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.95 ms /    31 runs   (    0.39
ms per token,  2593.49 tokens per second)
llama_print_timings: prompt eval time =   11846.65 ms /    97 tokens (  122.13
ms per token,     8.19 tokens per second)
llama_print_timings:        eval time =    6437.01 ms /    30 runs   (  214.57
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   18468.67 ms /   127 tokens
No. of rows:  46%| | 584/1258 [11:27:47<3:10:41, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.19 ms /    28 runs   (    0.36
ms per token,  2748.33 tokens per second)
llama_print_timings: prompt eval time =   13290.64 ms /   107 tokens (  124.21
ms per token,     8.05 tokens per second)
llama_print_timings:        eval time =    5853.21 ms /    27 runs   (  216.79
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   19308.26 ms /   134 tokens
No. of rows:  47%| | 585/1258 [11:28:07<3:18:14, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.70 ms /    28 runs   (    0.38
ms per token,  2615.84 tokens per second)
llama_print_timings: prompt eval time =   13779.75 ms /   114 tokens (  120.87
ms per token,     8.27 tokens per second)
llama_print_timings:        eval time =    6061.15 ms /    27 runs   (  224.49
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   20010.46 ms /   141 tokens
No. of rows:  47%| | 586/1258 [11:28:27<3:25:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =      11.91 ms /    31 runs   (    0.38
ms per token,  2603.29 tokens per second)
llama_print_timings: prompt eval time =   11202.51 ms /   104 tokens (  107.72
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    6432.13 ms /    30 runs   (  214.40
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   17818.47 ms /   134 tokens
No. of rows:  47%| | 587/1258 [11:28:45<3:23:41, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.79 ms /    28 runs   (    0.39
ms per token,  2594.03 tokens per second)
llama_print_timings: prompt eval time =   10003.35 ms /    91 tokens (  109.93
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    5809.26 ms /    27 runs   (  215.16
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   15983.31 ms /   118 tokens
No. of rows:  47%| | 588/1258 [11:29:01<3:15:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.48 ms /    50 runs   (    0.37
ms per token,  2705.77 tokens per second)
llama_print_timings: prompt eval time =   14259.93 ms /   131 tokens (  108.85
ms per token,     9.19 tokens per second)
llama_print_timings:        eval time =   10599.99 ms /    49 runs   (  216.33
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   25161.06 ms /   180 tokens
No. of rows:  47%| | 589/1258 [11:29:26<3:41:08, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.78 ms /    50 runs   (    0.40
ms per token,  2527.93 tokens per second)
llama_print_timings: prompt eval time =   15796.36 ms /   132 tokens (  119.67
ms per token,     8.36 tokens per second)
llama_print_timings:        eval time =   10750.69 ms /    49 runs   (  219.40
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   26854.21 ms /   181 tokens
No. of rows:  47%| | 590/1258 [11:29:53<4:04:14, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.69 ms /    34 runs   (    0.37
ms per token,  2678.43 tokens per second)
llama_print_timings: prompt eval time =   15720.10 ms /   131 tokens (  120.00
ms per token,     8.33 tokens per second)
```

```
llama_print_timings:        eval time =    7120.33 ms /    33 runs   (  215.77
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   23047.02 ms /   164 tokens
No. of rows:  47%| | 591/1258 [11:30:16<4:07:38, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.65 ms /    22 runs   (    0.39
ms per token,  2543.94 tokens per second)
llama_print_timings: prompt eval time =    9533.95 ms /    87 tokens (  109.59
ms per token,     9.13 tokens per second)
llama_print_timings:        eval time =    5079.64 ms /    21 runs   (  241.89
ms per token,     4.13 tokens per second)
llama_print_timings:       total time =   14754.06 ms /   108 tokens
No. of rows:  47%| | 592/1258 [11:30:30<3:42:14, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.73 ms /    23 runs   (    0.38
ms per token,  2633.39 tokens per second)
llama_print_timings: prompt eval time =    9304.07 ms /    84 tokens (  110.76
ms per token,     9.03 tokens per second)
llama_print_timings:        eval time =    4813.20 ms /    22 runs   (  218.78
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   14255.73 ms /   106 tokens
No. of rows:  47%| | 593/1258 [11:30:45<3:22:42, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.41 ms /    20 runs   (    0.37
ms per token,  2700.51 tokens per second)
llama_print_timings: prompt eval time =    9779.25 ms /    89 tokens (  109.88
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    4066.81 ms /    19 runs   (  214.04
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   13962.20 ms /   108 tokens
No. of rows:  47%| | 594/1258 [11:30:59<3:08:06, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.41 ms /    25 runs   (    0.38
ms per token,  2656.75 tokens per second)
llama_print_timings: prompt eval time =   10764.97 ms /    86 tokens (  125.17
ms per token,     7.99 tokens per second)
llama_print_timings:        eval time =    5390.36 ms /    24 runs   (  224.60
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   16307.61 ms /   110 tokens
No. of rows:  47%| | 595/1258 [11:31:15<3:05:33, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.74 ms /    29 runs   (    0.37
ms per token,  2700.94 tokens per second)
llama_print_timings: prompt eval time =   11083.67 ms /    88 tokens (  125.95
ms per token,     7.94 tokens per second)
llama_print_timings:        eval time =    7819.21 ms /    28 runs   (  279.26
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   19075.43 ms /   116 tokens
No. of rows:  47%| | 596/1258 [11:31:34<3:12:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.72 ms /    28 runs   (    0.38
ms per token,  2610.72 tokens per second)
llama_print_timings: prompt eval time =    9265.36 ms /    84 tokens (  110.30
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    7370.65 ms /    27 runs   (  272.99
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   16800.98 ms /   111 tokens
No. of rows:  47%| | 597/1258 [11:31:51<3:10:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.38 ms /    27 runs   (    0.38
ms per token,  2599.90 tokens per second)
llama_print_timings: prompt eval time =   10876.80 ms /    97 tokens (  112.13
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    5617.73 ms /    26 runs   (  216.07
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   16657.55 ms /   123 tokens
No. of rows:  48%| | 598/1258 [11:32:08<3:08:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.49 ms /    26 runs   (    0.36
ms per token,  2740.59 tokens per second)
llama_print_timings: prompt eval time =    9383.83 ms /    86 tokens (  109.11
ms per token,     9.16 tokens per second)
llama_print_timings:        eval time =    7072.46 ms /    25 runs   (  282.90
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   16610.25 ms /   111 tokens
No. of rows:  48%| | 599/1258 [11:32:24<3:06:11, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.74 ms /    23 runs   (    0.42
```

```
ms per token,  2361.15 tokens per second)
llama_print_timings: prompt eval time =   10444.53 ms /    93 tokens (  112.31
ms per token,     8.90 tokens per second)
llama_print_timings:        eval time =    4806.32 ms /    22 runs   (  218.47
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   15397.22 ms /   115 tokens
No. of rows:  48%| | 600/1258 [11:32:40<3:00:46, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.06 ms /    27 runs   (    0.37
ms per token,  2684.70 tokens per second)
llama_print_timings: prompt eval time =   10054.65 ms /    91 tokens (  110.49
ms per token,     9.05 tokens per second)
llama_print_timings:        eval time =    5590.02 ms /    26 runs   (  215.00
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   15813.81 ms /   117 tokens
No. of rows:  48%| | 601/1258 [11:32:55<2:58:20, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      20.41 ms /    50 runs   (    0.41
ms per token,  2450.38 tokens per second)
llama_print_timings: prompt eval time =   17180.97 ms /   160 tokens (  107.38
ms per token,     9.31 tokens per second)
llama_print_timings:        eval time =   12316.82 ms /    49 runs   (  251.36
ms per token,     3.98 tokens per second)
llama_print_timings:       total time =   29807.94 ms /   209 tokens
No. of rows:  48%| | 602/1258 [11:33:25<3:42:25, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.58 ms /    30 runs   (    0.39
ms per token,  2590.45 tokens per second)
llama_print_timings: prompt eval time =   12398.43 ms /   113 tokens (  109.72
ms per token,     9.11 tokens per second)
llama_print_timings:        eval time =    6249.06 ms /    29 runs   (  215.48
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   18830.49 ms /   142 tokens
No. of rows:  48%| | 603/1258 [11:33:44<3:37:12, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.91 ms /    34 runs   (    0.38
ms per token,  2633.21 tokens per second)
llama_print_timings: prompt eval time =   12766.47 ms /   106 tokens (  120.44
ms per token,     8.30 tokens per second)
llama_print_timings:        eval time =    7255.50 ms /    33 runs   (  219.86
```

```
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   20232.10 ms /   139 tokens
No. of rows:  48%| | 604/1258 [11:34:04<3:37:59, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.78 ms /    22 runs   (    0.40
ms per token,  2506.27 tokens per second)
llama_print_timings: prompt eval time =   10524.26 ms /    95 tokens (  110.78
ms per token,     9.03 tokens per second)
llama_print_timings:        eval time =    4709.84 ms /    21 runs   (  224.28
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   15370.98 ms /   116 tokens
No. of rows:  48%| | 605/1258 [11:34:20<3:22:34, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.20 ms /    42 runs   (    0.39
ms per token,  2593.07 tokens per second)
llama_print_timings: prompt eval time =   15323.27 ms /   128 tokens (  119.71
ms per token,     8.35 tokens per second)
llama_print_timings:        eval time =    9067.87 ms /    41 runs   (  221.17
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   24652.28 ms /   169 tokens
No. of rows:  48%| | 606/1258 [11:34:44<3:41:58, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.58 ms /    33 runs   (    0.38
ms per token,  2623.42 tokens per second)
llama_print_timings: prompt eval time =   11832.19 ms /   108 tokens (  109.56
ms per token,     9.13 tokens per second)
llama_print_timings:        eval time =    8585.92 ms /    32 runs   (  268.31
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   20616.19 ms /   140 tokens
No. of rows:  48%| | 607/1258 [11:35:05<3:42:16, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.19 ms /    35 runs   (    0.41
ms per token,  2465.83 tokens per second)
llama_print_timings: prompt eval time =   11473.61 ms /   107 tokens (  107.23
ms per token,     9.33 tokens per second)
llama_print_timings:        eval time =    7313.15 ms /    34 runs   (  215.09
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   19000.52 ms /   141 tokens
No. of rows:  48%| | 608/1258 [11:35:24<3:37:08, 2Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.39 ms /    27 runs   (    0.38
ms per token,  2599.15 tokens per second)
llama_print_timings: prompt eval time =   11493.91 ms /    94 tokens (  122.28
ms per token,     8.18 tokens per second)
llama_print_timings:        eval time =    6302.43 ms /    26 runs   (  242.40
ms per token,     4.13 tokens per second)
llama_print_timings:       total time =   17964.93 ms /   120 tokens
No. of rows:  48%| | 609/1258 [11:35:42<3:30:05, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.19 ms /    29 runs   (    0.39
ms per token,  2590.91 tokens per second)
llama_print_timings: prompt eval time =   10633.70 ms /    98 tokens (  108.51
ms per token,     9.22 tokens per second)
llama_print_timings:        eval time =    6491.27 ms /    28 runs   (  231.83
ms per token,     4.31 tokens per second)
llama_print_timings:       total time =   17306.78 ms /   126 tokens
No. of rows:  48%| | 610/1258 [11:35:59<3:22:54, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.89 ms /    26 runs   (    0.38
ms per token,  2628.65 tokens per second)
llama_print_timings: prompt eval time =   10269.93 ms /    93 tokens (  110.43
ms per token,     9.06 tokens per second)
llama_print_timings:        eval time =    7057.01 ms /    25 runs   (  282.28
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   17486.33 ms /   118 tokens
No. of rows:  49%| | 611/1258 [11:36:17<3:18:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.49 ms /    40 runs   (    0.39
ms per token,  2582.14 tokens per second)
llama_print_timings: prompt eval time =   14163.95 ms /   130 tokens (  108.95
ms per token,     9.18 tokens per second)
llama_print_timings:        eval time =    8516.20 ms /    39 runs   (  218.36
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   22929.07 ms /   169 tokens
No. of rows:  49%| | 612/1258 [11:36:40<3:32:46, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.98 ms /    23 runs   (    0.39
ms per token,  2561.53 tokens per second)
```

```
llama_print_timings: prompt eval time =    9051.37 ms /    81 tokens (  111.75
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    4720.95 ms /    22 runs   (  214.59
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   13912.11 ms /   103 tokens
No. of rows:  49%| | 613/1258 [11:36:54<3:13:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.87 ms /    21 runs   (    0.37
ms per token,  2667.01 tokens per second)
llama_print_timings: prompt eval time =   10407.75 ms /    85 tokens (  122.44
ms per token,     8.17 tokens per second)
llama_print_timings:        eval time =    4283.47 ms /    20 runs   (  214.17
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   14817.54 ms /   105 tokens
No. of rows:  49%| | 614/1258 [11:37:08<3:03:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.59 ms /    31 runs   (    0.37
ms per token,  2675.64 tokens per second)
llama_print_timings: prompt eval time =   10823.60 ms /   100 tokens (  108.24
ms per token,     9.24 tokens per second)
llama_print_timings:        eval time =    8280.08 ms /    30 runs   (  276.00
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   19292.33 ms /   130 tokens
No. of rows:  49%| | 615/1258 [11:37:28<3:09:58, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.21 ms /    24 runs   (    0.38
ms per token,  2606.71 tokens per second)
llama_print_timings: prompt eval time =    8378.78 ms /    75 tokens (  111.72
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    4953.17 ms /    23 runs   (  215.36
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   13472.75 ms /    98 tokens
No. of rows:  49%| | 616/1258 [11:37:41<2:56:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.02 ms /    37 runs   (    0.38
ms per token,  2639.65 tokens per second)
llama_print_timings: prompt eval time =   15008.43 ms /   129 tokens (  116.34
ms per token,     8.60 tokens per second)
llama_print_timings:        eval time =    7777.12 ms /    36 runs   (  216.03
ms per token,     4.63 tokens per second)
```

```
llama_print_timings:       total time =   23008.42 ms /   165 tokens
No. of rows:  49%| | 617/1258 [11:38:04<3:16:47, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       5.99 ms /    16 runs   (    0.37
ms per token,  2673.35 tokens per second)
llama_print_timings: prompt eval time =   10330.26 ms /    81 tokens (  127.53
ms per token,     7.84 tokens per second)
llama_print_timings:        eval time =    3552.62 ms /    15 runs   (  236.84
ms per token,     4.22 tokens per second)
llama_print_timings:       total time =   13977.75 ms /    96 tokens
No. of rows:  49%| | 618/1258 [11:38:18<3:02:17, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.44 ms /    36 runs   (    0.37
ms per token,  2678.17 tokens per second)
llama_print_timings: prompt eval time =   13024.34 ms /   119 tokens (  109.45
ms per token,     9.14 tokens per second)
llama_print_timings:        eval time =    7713.90 ms /    35 runs   (  220.40
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   20956.16 ms /   154 tokens
No. of rows:  49%| | 619/1258 [11:38:39<3:14:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       4.51 ms /    12 runs   (    0.38
ms per token,  2659.57 tokens per second)
llama_print_timings: prompt eval time =    7587.12 ms /    68 tokens (  111.58
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =    2363.83 ms /    11 runs   (  214.89
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   10020.22 ms /    79 tokens
No. of rows:  49%| | 620/1258 [11:38:49<2:47:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.71 ms /    40 runs   (    0.39
ms per token,  2546.96 tokens per second)
llama_print_timings: prompt eval time =   14439.16 ms /   133 tokens (  108.57
ms per token,     9.21 tokens per second)
llama_print_timings:        eval time =    8487.78 ms /    39 runs   (  217.64
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   23172.40 ms /   172 tokens
No. of rows:  49%| | 621/1258 [11:39:12<3:11:09, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.41 ms /    33 runs   (    0.38
ms per token,  2658.50 tokens per second)
llama_print_timings: prompt eval time =   13067.13 ms /   110 tokens (  118.79
ms per token,     8.42 tokens per second)
llama_print_timings:        eval time =    7068.44 ms /    32 runs   (  220.89
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   20340.91 ms /   142 tokens
No. of rows:  49%| | 622/1258 [11:39:33<3:18:18, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.49 ms /    22 runs   (    0.39
ms per token,  2590.67 tokens per second)
llama_print_timings: prompt eval time =   10321.60 ms /    81 tokens (  127.43
ms per token,     7.85 tokens per second)
llama_print_timings:        eval time =    5127.38 ms /    21 runs   (  244.16
ms per token,     4.10 tokens per second)
llama_print_timings:       total time =   15581.31 ms /   102 tokens
No. of rows:  50%| | 623/1258 [11:39:48<3:08:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.53 ms /    20 runs   (    0.38
ms per token,  2654.98 tokens per second)
llama_print_timings: prompt eval time =    9359.78 ms /    81 tokens (  115.55
ms per token,     8.65 tokens per second)
llama_print_timings:        eval time =    4107.69 ms /    19 runs   (  216.19
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   13584.74 ms /   100 tokens
No. of rows:  50%| | 624/1258 [11:40:02<2:54:33, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.13 ms /    32 runs   (    0.38
ms per token,  2637.44 tokens per second)
llama_print_timings: prompt eval time =   13636.45 ms /   116 tokens (  117.56
ms per token,     8.51 tokens per second)
llama_print_timings:        eval time =    6721.09 ms /    31 runs   (  216.81
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   20547.93 ms /   147 tokens
No. of rows:  50%| | 625/1258 [11:40:22<3:07:03, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.61 ms /    29 runs   (    0.47
ms per token,  2130.63 tokens per second)
llama_print_timings: prompt eval time =   12477.07 ms /   101 tokens (  123.54
```

```
ms per token,     8.09 tokens per second)
llama_print_timings:        eval time =    6063.44 ms /    28 runs   (  216.55
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   18720.53 ms /   129 tokens
No. of rows:  50%| | 626/1258 [11:40:41<3:09:53, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.15 ms /    28 runs   (    0.40
ms per token,  2511.44 tokens per second)
llama_print_timings: prompt eval time =   10607.96 ms /    97 tokens (  109.36
ms per token,     9.14 tokens per second)
llama_print_timings:        eval time =    5786.78 ms /    27 runs   (  214.33
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   16567.85 ms /   124 tokens
No. of rows:  50%| | 627/1258 [11:40:58<3:05:00, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.77 ms /    31 runs   (    0.38
ms per token,  2633.81 tokens per second)
llama_print_timings: prompt eval time =   12514.02 ms /   102 tokens (  122.69
ms per token,     8.15 tokens per second)
llama_print_timings:        eval time =    6439.83 ms /    30 runs   (  214.66
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   19136.08 ms /   132 tokens
No. of rows:  50%| | 628/1258 [11:41:17<3:09:37, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.86 ms /    43 runs   (    0.37
ms per token,  2710.88 tokens per second)
llama_print_timings: prompt eval time =   12229.26 ms /   100 tokens (  122.29
ms per token,     8.18 tokens per second)
llama_print_timings:        eval time =    9090.66 ms /    42 runs   (  216.44
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   21574.09 ms /   142 tokens
No. of rows:  50%| | 629/1258 [11:41:39<3:20:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.12 ms /    33 runs   (    0.40
ms per token,  2515.82 tokens per second)
llama_print_timings: prompt eval time =   12398.86 ms /   101 tokens (  122.76
ms per token,     8.15 tokens per second)
llama_print_timings:        eval time =    7018.64 ms /    32 runs   (  219.33
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   19619.89 ms /   133 tokens
```

```
No. of rows:  50%| | 630/1258 [11:41:58<3:21:41, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.46 ms /    28 runs   (    0.37
ms per token,  2677.12 tokens per second)
llama_print_timings: prompt eval time =   11464.39 ms /    94 tokens (  121.96
ms per token,     8.20 tokens per second)
llama_print_timings:        eval time =    5800.14 ms /    27 runs   (  214.82
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   17431.09 ms /   121 tokens
No. of rows:  50%| | 631/1258 [11:42:16<3:15:38, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.19 ms /    29 runs   (    0.39
ms per token,  2591.60 tokens per second)
llama_print_timings: prompt eval time =   12668.44 ms /   105 tokens (  120.65
ms per token,     8.29 tokens per second)
llama_print_timings:        eval time =    6304.33 ms /    28 runs   (  225.15
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   19152.53 ms /   133 tokens
No. of rows:  50%| | 632/1258 [11:42:35<3:16:41, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.98 ms /    22 runs   (    0.36
ms per token,  2755.51 tokens per second)
llama_print_timings: prompt eval time =   11245.84 ms /   101 tokens (  111.34
ms per token,     8.98 tokens per second)
llama_print_timings:        eval time =    4591.51 ms /    21 runs   (  218.64
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   15968.39 ms /   122 tokens
No. of rows:  50%| | 633/1258 [11:42:51<3:07:20, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.09 ms /    22 runs   (    0.37
ms per token,  2719.41 tokens per second)
llama_print_timings: prompt eval time =    9252.27 ms /    83 tokens (  111.47
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    5173.95 ms /    21 runs   (  246.38
ms per token,     4.06 tokens per second)
llama_print_timings:       total time =   14555.50 ms /   104 tokens
No. of rows:  50%| | 634/1258 [11:43:05<2:56:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =      14.61 ms /    36 runs   (    0.41
ms per token,  2464.91 tokens per second)
llama_print_timings: prompt eval time =   13143.01 ms /   121 tokens (  108.62
ms per token,     9.21 tokens per second)
llama_print_timings:        eval time =    7570.53 ms /    35 runs   (  216.30
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   20935.16 ms /   156 tokens
No. of rows:  50%| | 635/1258 [11:43:26<3:08:31, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.29 ms /    31 runs   (    0.36
ms per token,  2746.52 tokens per second)
llama_print_timings: prompt eval time =   13092.37 ms /   107 tokens (  122.36
ms per token,     8.17 tokens per second)
llama_print_timings:        eval time =    6547.97 ms /    30 runs   (  218.27
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   19822.49 ms /   137 tokens
No. of rows:  51%| | 636/1258 [11:43:46<3:13:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.40 ms /    23 runs   (    0.37
ms per token,  2738.10 tokens per second)
llama_print_timings: prompt eval time =    9996.30 ms /    90 tokens (  111.07
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =    5880.07 ms /    22 runs   (  267.28
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   16015.70 ms /   112 tokens
No. of rows:  51%| | 637/1258 [11:44:02<3:04:55, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.48 ms /    37 runs   (    0.36
ms per token,  2744.20 tokens per second)
llama_print_timings: prompt eval time =   12323.46 ms /   112 tokens (  110.03
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =    9459.40 ms /    36 runs   (  262.76
ms per token,     3.81 tokens per second)
llama_print_timings:       total time =   22001.92 ms /   148 tokens
No. of rows:  51%| | 638/1258 [11:44:24<3:17:28, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.46 ms /    17 runs   (    0.38
ms per token,  2633.62 tokens per second)
llama_print_timings: prompt eval time =    9755.20 ms /    87 tokens (  112.13
ms per token,     8.92 tokens per second)
```

```
llama_print_timings:        eval time =    3447.74 ms /    16 runs   (  215.48
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   13309.10 ms /   103 tokens
No. of rows:  51%| | 639/1258 [11:44:37<2:59:13, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.83 ms /    31 runs   (    0.38
ms per token,  2621.12 tokens per second)
llama_print_timings: prompt eval time =   12113.59 ms /   111 tokens (  109.13
ms per token,     9.16 tokens per second)
llama_print_timings:        eval time =    6499.11 ms /    30 runs   (  216.64
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   18798.57 ms /   141 tokens
No. of rows:  51%| | 640/1258 [11:44:56<3:03:20, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.72 ms /    48 runs   (    0.37
ms per token,  2709.11 tokens per second)
llama_print_timings: prompt eval time =   18358.48 ms /   158 tokens (  116.19
ms per token,     8.61 tokens per second)
llama_print_timings:        eval time =   12245.65 ms /    47 runs   (  260.55
ms per token,     3.84 tokens per second)
llama_print_timings:       total time =   30902.50 ms /   205 tokens
No. of rows:  51%| | 641/1258 [11:45:27<3:43:30, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.52 ms /    20 runs   (    0.38
ms per token,  2659.22 tokens per second)
llama_print_timings: prompt eval time =    8003.75 ms /    72 tokens (  111.16
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =    4410.08 ms /    19 runs   (  232.11
ms per token,     4.31 tokens per second)
llama_print_timings:       total time =   12534.50 ms /    91 tokens
No. of rows:  51%| | 642/1258 [11:45:40<3:14:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.12 ms /    34 runs   (    0.39
ms per token,  2591.07 tokens per second)
llama_print_timings: prompt eval time =   13414.98 ms /   120 tokens (  111.79
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    8688.07 ms /    33 runs   (  263.27
ms per token,     3.80 tokens per second)
llama_print_timings:       total time =   22305.93 ms /   153 tokens
No. of rows:  51%| | 643/1258 [11:46:02<3:24:46, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.10 ms /    32 runs   (    0.38
ms per token,  2643.75 tokens per second)
llama_print_timings: prompt eval time =   12715.46 ms /   116 tokens (  109.62
ms per token,     9.12 tokens per second)
llama_print_timings:        eval time =    6730.02 ms /    31 runs   (  217.10
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   19638.85 ms /   147 tokens
No. of rows:  51%| | 644/1258 [11:46:22<3:23:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.80 ms /    20 runs   (    0.44
ms per token,  2271.69 tokens per second)
llama_print_timings: prompt eval time =    8691.65 ms /    78 tokens (  111.43
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    4330.34 ms /    19 runs   (  227.91
ms per token,     4.39 tokens per second)
llama_print_timings:       total time =   13147.54 ms /    97 tokens
No. of rows:  51%| | 645/1258 [11:46:35<3:02:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.11 ms /    50 runs   (    0.38
ms per token,  2616.57 tokens per second)
llama_print_timings: prompt eval time =   17120.62 ms /   158 tokens (  108.36
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =   10661.42 ms /    49 runs   (  217.58
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   28089.83 ms /   207 tokens
No. of rows:  51%| | 646/1258 [11:47:03<3:33:32, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.49 ms /    20 runs   (    0.37
ms per token,  2670.23 tokens per second)
llama_print_timings: prompt eval time =    9790.18 ms /    88 tokens (  111.25
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    4079.82 ms /    19 runs   (  214.73
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   13994.66 ms /   107 tokens
No. of rows:  51%| | 647/1258 [11:47:17<3:11:58, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.04 ms /    16 runs   (    0.38
```

```
ms per token,  2647.25 tokens per second)
llama_print_timings: prompt eval time =   10278.35 ms /    82 tokens (  125.35
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =    3510.11 ms /    15 runs   (  234.01
ms per token,     4.27 tokens per second)
llama_print_timings:       total time =   13887.18 ms /    97 tokens
No. of rows:  52%| | 648/1258 [11:47:31<2:56:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.51 ms /    27 runs   (    0.39
ms per token,  2569.72 tokens per second)
llama_print_timings: prompt eval time =   12563.63 ms /   114 tokens (  110.21
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    5606.79 ms /    26 runs   (  215.65
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   18337.40 ms /   140 tokens
No. of rows:  52%| | 649/1258 [11:47:49<2:59:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.71 ms /    26 runs   (    0.37
ms per token,  2677.65 tokens per second)
llama_print_timings: prompt eval time =   12283.10 ms /    99 tokens (  124.07
ms per token,     8.06 tokens per second)
llama_print_timings:        eval time =    5377.24 ms /    25 runs   (  215.09
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   17815.16 ms /   124 tokens
No. of rows:  52%| | 650/1258 [11:48:07<2:59:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.59 ms /    32 runs   (    0.39
ms per token,  2541.90 tokens per second)
llama_print_timings: prompt eval time =   13059.90 ms /   108 tokens (  120.93
ms per token,     8.27 tokens per second)
llama_print_timings:        eval time =    6722.50 ms /    31 runs   (  216.85
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   19982.88 ms /   139 tokens
No. of rows:  52%| | 651/1258 [11:48:27<3:06:06, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.56 ms /    19 runs   (    0.50
ms per token,  1986.62 tokens per second)
llama_print_timings: prompt eval time =   10869.77 ms /    84 tokens (  129.40
ms per token,     7.73 tokens per second)
llama_print_timings:        eval time =    3907.86 ms /    18 runs   (  217.10
```

```
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   14894.30 ms /   102 tokens
No. of rows:  52%| | 652/1258 [11:48:42<2:55:12, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.29 ms /    24 runs   (    0.39
ms per token,  2583.70 tokens per second)
llama_print_timings: prompt eval time =    9755.39 ms /    89 tokens (  109.61
ms per token,     9.12 tokens per second)
llama_print_timings:        eval time =    4964.24 ms /    23 runs   (  215.84
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   14862.34 ms /   112 tokens
No. of rows:  52%| | 653/1258 [11:48:57<2:47:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.92 ms /    19 runs   (    0.36
ms per token,  2744.47 tokens per second)
llama_print_timings: prompt eval time =    8551.50 ms /    77 tokens (  111.06
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =    3856.66 ms /    18 runs   (  214.26
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   12519.74 ms /    95 tokens
No. of rows:  52%| | 654/1258 [11:49:09<2:34:50, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.54 ms /    21 runs   (    0.41
ms per token,  2459.30 tokens per second)
llama_print_timings: prompt eval time =    9412.40 ms /    82 tokens (  114.79
ms per token,     8.71 tokens per second)
llama_print_timings:        eval time =    5997.16 ms /    20 runs   (  299.86
ms per token,     3.33 tokens per second)
llama_print_timings:       total time =   15534.47 ms /   102 tokens
No. of rows:  52%| | 655/1258 [11:49:25<2:35:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.34 ms /    27 runs   (    0.38
ms per token,  2612.23 tokens per second)
llama_print_timings: prompt eval time =    9767.26 ms /    88 tokens (  110.99
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    5594.47 ms /    26 runs   (  215.17
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   15526.01 ms /   114 tokens
No. of rows:  52%| | 656/1258 [11:49:40<2:35:07, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.22 ms /    31 runs   (    0.39
ms per token,  2537.24 tokens per second)
llama_print_timings: prompt eval time =   12666.35 ms /   105 tokens (  120.63
ms per token,     8.29 tokens per second)
llama_print_timings:        eval time =    6639.70 ms /    30 runs   (  221.32
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   19496.98 ms /   135 tokens
No. of rows:  52%| | 657/1258 [11:50:00<2:47:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.11 ms /    28 runs   (    0.40
ms per token,  2520.03 tokens per second)
llama_print_timings: prompt eval time =    9872.27 ms /    90 tokens (  109.69
ms per token,     9.12 tokens per second)
llama_print_timings:        eval time =    5830.38 ms /    27 runs   (  215.94
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   15875.68 ms /   117 tokens
No. of rows:  52%| | 658/1258 [11:50:16<2:44:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.09 ms /    32 runs   (    0.41
ms per token,  2443.87 tokens per second)
llama_print_timings: prompt eval time =   13267.23 ms /   109 tokens (  121.72
ms per token,     8.22 tokens per second)
llama_print_timings:        eval time =    6893.69 ms /    31 runs   (  222.38
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   20358.99 ms /   140 tokens
No. of rows:  52%| | 659/1258 [11:50:36<2:55:52, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.68 ms /    31 runs   (    0.38
ms per token,  2654.34 tokens per second)
llama_print_timings: prompt eval time =   12864.76 ms /   117 tokens (  109.96
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =    6832.15 ms /    30 runs   (  227.74
ms per token,     4.39 tokens per second)
llama_print_timings:       total time =   19886.40 ms /   147 tokens
No. of rows:  52%| | 660/1258 [11:50:56<3:02:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.31 ms /    23 runs   (    0.40
ms per token,  2471.26 tokens per second)
```

```
llama_print_timings: prompt eval time =   10914.79 ms /    99 tokens (  110.25
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    4759.39 ms /    22 runs   (  216.34
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   15817.64 ms /   121 tokens
No. of rows:  53%| | 661/1258 [11:51:12<2:54:40, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.87 ms /    20 runs   (    0.44
ms per token,  2254.79 tokens per second)
llama_print_timings: prompt eval time =    9417.15 ms /    75 tokens (  125.56
ms per token,     7.96 tokens per second)
llama_print_timings:        eval time =    4057.93 ms /    19 runs   (  213.58
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   13598.21 ms /    94 tokens
No. of rows:  53%| | 662/1258 [11:51:25<2:42:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.39 ms /    27 runs   (    0.38
ms per token,  2598.90 tokens per second)
llama_print_timings: prompt eval time =    9982.93 ms /    92 tokens (  108.51
ms per token,     9.22 tokens per second)
llama_print_timings:        eval time =    5602.25 ms /    26 runs   (  215.47
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   15748.50 ms /   118 tokens
No. of rows:  53%| | 663/1258 [11:51:41<2:40:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.57 ms /    30 runs   (    0.42
ms per token,  2385.88 tokens per second)
llama_print_timings: prompt eval time =   13145.06 ms /   110 tokens (  119.50
ms per token,     8.37 tokens per second)
llama_print_timings:        eval time =    8294.25 ms /    29 runs   (  286.01
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   21635.42 ms /   139 tokens
No. of rows:  53%| | 664/1258 [11:52:03<2:56:28, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.73 ms /    28 runs   (    0.38
ms per token,  2610.48 tokens per second)
llama_print_timings: prompt eval time =   13176.60 ms /   121 tokens (  108.90
ms per token,     9.18 tokens per second)
llama_print_timings:        eval time =    7588.08 ms /    27 runs   (  281.04
ms per token,     3.56 tokens per second)
```

```
llama_print_timings:       total time =   20933.62 ms /   148 tokens
No. of rows:  53%| | 665/1258 [11:52:24<3:05:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.36 ms /    18 runs   (    0.41
ms per token,  2445.98 tokens per second)
llama_print_timings: prompt eval time =    8474.02 ms /    77 tokens (  110.05
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =    3870.91 ms /    17 runs   (  227.70
ms per token,     4.39 tokens per second)
llama_print_timings:       total time =   12463.43 ms /    94 tokens
No. of rows:  53%| | 666/1258 [11:52:36<2:46:28, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.61 ms /    26 runs   (    0.37
ms per token,  2705.23 tokens per second)
llama_print_timings: prompt eval time =   10017.49 ms /    91 tokens (  110.08
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    5344.44 ms /    25 runs   (  213.78
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   15517.35 ms /   116 tokens
No. of rows:  53%| | 667/1258 [11:52:52<2:42:11, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.58 ms /    28 runs   (    0.38
ms per token,  2646.00 tokens per second)
llama_print_timings: prompt eval time =   10012.94 ms /    90 tokens (  111.25
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    5812.99 ms /    27 runs   (  215.30
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   15991.17 ms /   117 tokens
No. of rows:  53%| | 668/1258 [11:53:08<2:40:33, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.19 ms /    22 runs   (    0.37
ms per token,  2685.87 tokens per second)
llama_print_timings: prompt eval time =   12027.47 ms /    99 tokens (  121.49
ms per token,     8.23 tokens per second)
llama_print_timings:        eval time =    4510.65 ms /    21 runs   (  214.79
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   16668.92 ms /   120 tokens
No. of rows:  53%| | 669/1258 [11:53:24<2:41:16, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.14 ms /    29 runs   (    0.38
ms per token,  2602.53 tokens per second)
llama_print_timings: prompt eval time =   12509.87 ms /   114 tokens (  109.74
ms per token,     9.11 tokens per second)
llama_print_timings:        eval time =    6086.46 ms /    28 runs   (  217.37
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   18771.28 ms /   142 tokens
No. of rows:  53%| | 670/1258 [11:53:43<2:47:54, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.84 ms /    27 runs   (    0.51
ms per token,  1950.30 tokens per second)
llama_print_timings: prompt eval time =   11478.33 ms /   106 tokens (  108.29
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =    5815.05 ms /    26 runs   (  223.66
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   17469.83 ms /   132 tokens
No. of rows:  53%| | 671/1258 [11:54:01<2:48:40, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.24 ms /    29 runs   (    0.39
ms per token,  2580.53 tokens per second)
llama_print_timings: prompt eval time =   11021.96 ms /   100 tokens (  110.22
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    6042.55 ms /    28 runs   (  215.81
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   17238.36 ms /   128 tokens
No. of rows:  53%| | 672/1258 [11:54:18<2:48:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.81 ms /    32 runs   (    0.37
ms per token,  2709.57 tokens per second)
llama_print_timings: prompt eval time =   11633.12 ms /   107 tokens (  108.72
ms per token,     9.20 tokens per second)
llama_print_timings:        eval time =    6857.20 ms /    31 runs   (  221.20
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   18683.23 ms /   138 tokens
No. of rows:  53%| | 673/1258 [11:54:37<2:52:20, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.70 ms /    31 runs   (    0.38
ms per token,  2650.71 tokens per second)
llama_print_timings: prompt eval time =   10332.83 ms /    94 tokens (  109.92
```

```
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    6423.74 ms /    30 runs   (  214.12
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   16947.13 ms /   124 tokens
No. of rows:  54%| | 674/1258 [11:54:54<2:49:54, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.19 ms /    29 runs   (    0.39
ms per token,  2591.37 tokens per second)
llama_print_timings: prompt eval time =   12971.65 ms /   116 tokens (  111.82
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =    6075.87 ms /    28 runs   (  217.00
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   19224.71 ms /   144 tokens
No. of rows:  54%| | 675/1258 [11:55:13<2:54:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.17 ms /    20 runs   (    0.36
ms per token,  2787.46 tokens per second)
llama_print_timings: prompt eval time =    9209.64 ms /    84 tokens (  109.64
ms per token,     9.12 tokens per second)
llama_print_timings:        eval time =    4059.09 ms /    19 runs   (  213.64
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   13387.11 ms /   103 tokens
No. of rows:  54%| | 676/1258 [11:55:26<2:41:08, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.84 ms /    20 runs   (    0.39
ms per token,  2550.70 tokens per second)
llama_print_timings: prompt eval time =   10774.03 ms /    85 tokens (  126.75
ms per token,     7.89 tokens per second)
llama_print_timings:        eval time =    4088.85 ms /    19 runs   (  215.20
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   14982.42 ms /   104 tokens
No. of rows:  54%| | 677/1258 [11:55:41<2:36:07, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.01 ms /    22 runs   (    0.36
ms per token,  2747.60 tokens per second)
llama_print_timings: prompt eval time =    9573.04 ms /    78 tokens (  122.73
ms per token,     8.15 tokens per second)
llama_print_timings:        eval time =    4603.78 ms /    21 runs   (  219.23
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   14305.37 ms /    99 tokens
```

```
No. of rows:  54%| | 678/1258 [11:55:55<2:30:38, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.81 ms /    24 runs   (    0.37
ms per token,  2725.11 tokens per second)
llama_print_timings: prompt eval time =   10216.74 ms /    92 tokens (  111.05
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =    4914.15 ms /    23 runs   (  213.66
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   15270.84 ms /   115 tokens
No. of rows:  54%| | 679/1258 [11:56:11<2:29:28, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.35 ms /    24 runs   (    0.39
ms per token,  2568.22 tokens per second)
llama_print_timings: prompt eval time =   11147.38 ms /   102 tokens (  109.29
ms per token,     9.15 tokens per second)
llama_print_timings:        eval time =    4994.28 ms /    23 runs   (  217.14
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   16289.10 ms /   125 tokens
No. of rows:  54%| | 680/1258 [11:56:27<2:31:33, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.81 ms /    26 runs   (    0.45
ms per token,  2201.15 tokens per second)
llama_print_timings: prompt eval time =   12127.77 ms /    98 tokens (  123.75
ms per token,     8.08 tokens per second)
llama_print_timings:        eval time =    5949.72 ms /    25 runs   (  237.99
ms per token,     4.20 tokens per second)
llama_print_timings:       total time =   18263.74 ms /   123 tokens
No. of rows:  54%| | 681/1258 [11:56:45<2:38:37, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.26 ms /    25 runs   (    0.41
ms per token,  2436.41 tokens per second)
llama_print_timings: prompt eval time =   11335.74 ms /    97 tokens (  116.86
ms per token,     8.56 tokens per second)
llama_print_timings:        eval time =    5488.35 ms /    24 runs   (  228.68
ms per token,     4.37 tokens per second)
llama_print_timings:       total time =   16987.14 ms /   121 tokens
No. of rows:  54%| | 682/1258 [11:57:02<2:39:47, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =      11.89 ms /    31 runs   (    0.38
ms per token,  2606.14 tokens per second)
llama_print_timings: prompt eval time =   12682.44 ms /   117 tokens (  108.40
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =    8266.76 ms /    30 runs   (  275.56
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   21138.09 ms /   147 tokens
No. of rows:  54%| | 683/1258 [11:57:23<2:52:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.35 ms /    27 runs   (    0.38
ms per token,  2608.44 tokens per second)
llama_print_timings: prompt eval time =   10853.94 ms /    98 tokens (  110.75
ms per token,     9.03 tokens per second)
llama_print_timings:        eval time =    5704.69 ms /    26 runs   (  219.41
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   16721.26 ms /   124 tokens
No. of rows:  54%| | 684/1258 [11:57:40<2:48:30, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.69 ms /    19 runs   (    0.40
ms per token,  2470.74 tokens per second)
llama_print_timings: prompt eval time =    9902.19 ms /    90 tokens (  110.02
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =    3842.12 ms /    18 runs   (  213.45
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   13859.26 ms /   108 tokens
No. of rows:  54%| | 685/1258 [11:57:54<2:37:29, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.28 ms /    32 runs   (    0.38
ms per token,  2606.50 tokens per second)
llama_print_timings: prompt eval time =   12375.47 ms /   100 tokens (  123.75
ms per token,     8.08 tokens per second)
llama_print_timings:        eval time =    8440.28 ms /    31 runs   (  272.27
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   21008.53 ms /   131 tokens
No. of rows:  55%| | 686/1258 [11:58:15<2:50:09, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.69 ms /    20 runs   (    0.38
ms per token,  2600.78 tokens per second)
llama_print_timings: prompt eval time =    9297.45 ms /    82 tokens (  113.38
ms per token,     8.82 tokens per second)
```

```
llama_print_timings:        eval time =    4097.74 ms /    19 runs   (  215.67
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   13515.28 ms /   101 tokens
No. of rows:  55%| | 687/1258 [11:58:29<2:37:30, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.04 ms /    21 runs   (    0.38
ms per token,  2611.94 tokens per second)
llama_print_timings: prompt eval time =    9181.41 ms /    84 tokens (  109.30
ms per token,     9.15 tokens per second)
llama_print_timings:        eval time =    4529.51 ms /    20 runs   (  226.48
ms per token,     4.42 tokens per second)
llama_print_timings:       total time =   13841.80 ms /   104 tokens
No. of rows:  55%| | 688/1258 [11:58:42<2:29:30, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.29 ms /    30 runs   (    0.38
ms per token,  2658.16 tokens per second)
llama_print_timings: prompt eval time =   11493.77 ms /   104 tokens (  110.52
ms per token,     9.05 tokens per second)
llama_print_timings:        eval time =    6306.21 ms /    29 runs   (  217.46
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   17980.35 ms /   133 tokens
No. of rows:  55%| | 689/1258 [11:59:00<2:35:40, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.02 ms /    22 runs   (    0.41
ms per token,  2440.11 tokens per second)
llama_print_timings: prompt eval time =   10076.17 ms /    91 tokens (  110.73
ms per token,     9.03 tokens per second)
llama_print_timings:        eval time =    6344.85 ms /    21 runs   (  302.14
ms per token,     3.31 tokens per second)
llama_print_timings:       total time =   16562.53 ms /   112 tokens
No. of rows:  55%| | 690/1258 [11:59:17<2:35:50, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.55 ms /    28 runs   (    0.38
ms per token,  2653.53 tokens per second)
llama_print_timings: prompt eval time =   11607.48 ms /   105 tokens (  110.55
ms per token,     9.05 tokens per second)
llama_print_timings:        eval time =    7473.39 ms /    27 runs   (  276.79
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   19246.41 ms /   132 tokens
No. of rows:  55%| | 691/1258 [11:59:36<2:43:28, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.92 ms /    29 runs   (    0.45
ms per token,  2244.93 tokens per second)
llama_print_timings: prompt eval time =   11462.51 ms /   103 tokens (  111.29
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    6071.58 ms /    28 runs   (  216.84
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   17708.02 ms /   131 tokens
No. of rows:  55%| | 692/1258 [11:59:54<2:44:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.04 ms /    34 runs   (    0.38
ms per token,  2607.16 tokens per second)
llama_print_timings: prompt eval time =   14278.05 ms /   119 tokens (  119.98
ms per token,     8.33 tokens per second)
llama_print_timings:        eval time =    7142.04 ms /    33 runs   (  216.43
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   21628.82 ms /   152 tokens
No. of rows:  55%| | 693/1258 [12:00:16<2:55:58, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.48 ms /    29 runs   (    0.36
ms per token,  2766.12 tokens per second)
llama_print_timings: prompt eval time =   11386.57 ms /    92 tokens (  123.77
ms per token,     8.08 tokens per second)
llama_print_timings:        eval time =    7099.26 ms /    28 runs   (  253.55
ms per token,     3.94 tokens per second)
llama_print_timings:       total time =   18661.59 ms /   120 tokens
No. of rows:  55%| | 694/1258 [12:00:34<2:55:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.86 ms /    28 runs   (    0.39
ms per token,  2578.51 tokens per second)
llama_print_timings: prompt eval time =    9675.76 ms /    85 tokens (  113.83
ms per token,     8.78 tokens per second)
llama_print_timings:        eval time =    5799.68 ms /    27 runs   (  214.80
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   15642.97 ms /   112 tokens
No. of rows:  55%| | 695/1258 [12:00:50<2:46:46, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.29 ms /    22 runs   (    0.38
```

```
ms per token,  2652.52 tokens per second)
llama_print_timings: prompt eval time =    8789.80 ms /    80 tokens (  109.87
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    4526.80 ms /    21 runs   (  215.56
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   13448.54 ms /   101 tokens
No. of rows:  55%| | 696/1258 [12:01:03<2:34:20, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.96 ms /    27 runs   (    0.37
ms per token,  2711.12 tokens per second)
llama_print_timings: prompt eval time =   11339.52 ms /   100 tokens (  113.40
ms per token,     8.82 tokens per second)
llama_print_timings:        eval time =    5638.01 ms /    26 runs   (  216.85
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   17135.37 ms /   126 tokens
No. of rows:  55%| | 697/1258 [12:01:21<2:35:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.75 ms /    36 runs   (    0.38
ms per token,  2618.94 tokens per second)
llama_print_timings: prompt eval time =   13397.76 ms /   112 tokens (  119.62
ms per token,     8.36 tokens per second)
llama_print_timings:        eval time =    9266.67 ms /    35 runs   (  264.76
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   22883.14 ms /   147 tokens
No. of rows:  55%| | 698/1258 [12:01:43<2:53:03, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.57 ms /    25 runs   (    0.42
ms per token,  2364.96 tokens per second)
llama_print_timings: prompt eval time =   11354.12 ms /   104 tokens (  109.17
ms per token,     9.16 tokens per second)
llama_print_timings:        eval time =    5338.05 ms /    24 runs   (  222.42
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   16848.01 ms /   128 tokens
No. of rows:  56%| | 699/1258 [12:02:00<2:47:59, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.78 ms /    31 runs   (    0.38
ms per token,  2632.25 tokens per second)
llama_print_timings: prompt eval time =   13943.25 ms /   116 tokens (  120.20
ms per token,     8.32 tokens per second)
llama_print_timings:        eval time =    6545.37 ms /    30 runs   (  218.18
```

```
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   20675.97 ms /   146 tokens
No. of rows:  56%| | 700/1258 [12:02:21<2:55:08, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.72 ms /    23 runs   (    0.38
ms per token,  2637.01 tokens per second)
llama_print_timings: prompt eval time =   12101.34 ms /    97 tokens (  124.76
ms per token,     8.02 tokens per second)
llama_print_timings:        eval time =    4782.58 ms /    22 runs   (  217.39
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   17020.36 ms /   119 tokens
No. of rows:  56%| | 701/1258 [12:02:38<2:49:46, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.76 ms /    28 runs   (    0.38
ms per token,  2603.44 tokens per second)
llama_print_timings: prompt eval time =   11297.01 ms /    90 tokens (  125.52
ms per token,     7.97 tokens per second)
llama_print_timings:        eval time =    5825.72 ms /    27 runs   (  215.77
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   17293.54 ms /   117 tokens
No. of rows:  56%| | 702/1258 [12:02:55<2:46:44, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.47 ms /    28 runs   (    0.37
ms per token,  2673.54 tokens per second)
llama_print_timings: prompt eval time =   10846.98 ms /    99 tokens (  109.57
ms per token,     9.13 tokens per second)
llama_print_timings:        eval time =    5818.25 ms /    27 runs   (  215.49
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   16832.74 ms /   126 tokens
No. of rows:  56%| | 703/1258 [12:03:12<2:43:14, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.26 ms /    22 runs   (    0.38
ms per token,  2663.44 tokens per second)
llama_print_timings: prompt eval time =    8505.71 ms /    77 tokens (  110.46
ms per token,     9.05 tokens per second)
llama_print_timings:        eval time =    4460.95 ms /    21 runs   (  212.43
ms per token,     4.71 tokens per second)
llama_print_timings:       total time =   13096.91 ms /    98 tokens
No. of rows:  56%| | 704/1258 [12:03:25<2:30:19, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.98 ms /    24 runs   (    0.37
ms per token,  2672.90 tokens per second)
llama_print_timings: prompt eval time =   10507.29 ms /    97 tokens (  108.32
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =    4925.78 ms /    23 runs   (  214.16
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   15578.10 ms /   120 tokens
No. of rows:  56%| | 705/1258 [12:03:41<2:28:07, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.86 ms /    24 runs   (    0.37
ms per token,  2708.19 tokens per second)
llama_print_timings: prompt eval time =   12310.94 ms /   100 tokens (  123.11
ms per token,     8.12 tokens per second)
llama_print_timings:        eval time =    4951.14 ms /    23 runs   (  215.27
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   17407.16 ms /   123 tokens
No. of rows:  56%| | 706/1258 [12:03:58<2:31:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.41 ms /    29 runs   (    0.39
ms per token,  2541.63 tokens per second)
llama_print_timings: prompt eval time =   11720.09 ms /    95 tokens (  123.37
ms per token,     8.11 tokens per second)
llama_print_timings:        eval time =    6402.97 ms /    28 runs   (  228.68
ms per token,     4.37 tokens per second)
llama_print_timings:       total time =   18314.46 ms /   123 tokens
No. of rows:  56%| | 707/1258 [12:04:17<2:36:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.24 ms /    24 runs   (    0.38
ms per token,  2597.68 tokens per second)
llama_print_timings: prompt eval time =   11502.66 ms /    94 tokens (  122.37
ms per token,     8.17 tokens per second)
llama_print_timings:        eval time =    4958.02 ms /    23 runs   (  215.57
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   16606.10 ms /   117 tokens
No. of rows:  56%| | 708/1258 [12:04:33<2:34:57, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.81 ms /    23 runs   (    0.38
ms per token,  2610.37 tokens per second)
```

```
llama_print_timings: prompt eval time =    9903.30 ms /    77 tokens (  128.61
ms per token,     7.78 tokens per second)
llama_print_timings:        eval time =    4788.93 ms /    22 runs   (  217.68
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   14831.25 ms /    99 tokens
No. of rows:  56%| | 709/1258 [12:04:48<2:28:59, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.57 ms /    24 runs   (    0.40
ms per token,  2508.62 tokens per second)
llama_print_timings: prompt eval time =    9475.43 ms /    84 tokens (  112.80
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    4974.84 ms /    23 runs   (  216.30
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   14596.91 ms /   107 tokens
No. of rows:  56%| | 710/1258 [12:05:03<2:24:09, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.06 ms /    22 runs   (    0.37
ms per token,  2730.21 tokens per second)
llama_print_timings: prompt eval time =    9454.76 ms /    85 tokens (  111.23
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    4531.75 ms /    21 runs   (  215.80
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   14116.80 ms /   106 tokens
No. of rows:  57%| | 711/1258 [12:05:17<2:19:20, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.02 ms /    28 runs   (    0.39
ms per token,  2541.07 tokens per second)
llama_print_timings: prompt eval time =   11064.09 ms /    89 tokens (  124.32
ms per token,     8.04 tokens per second)
llama_print_timings:        eval time =    7549.33 ms /    27 runs   (  279.60
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   18782.73 ms /   116 tokens
No. of rows:  57%| | 712/1258 [12:05:36<2:28:40, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.42 ms /    37 runs   (    0.44
ms per token,  2253.62 tokens per second)
llama_print_timings: prompt eval time =   11356.67 ms /   102 tokens (  111.34
ms per token,     8.98 tokens per second)
llama_print_timings:        eval time =    9419.07 ms /    36 runs   (  261.64
ms per token,     3.82 tokens per second)
```

```
llama_print_timings:       total time =   20997.63 ms /   138 tokens
No. of rows:  57%| | 713/1258 [12:05:57<2:41:05, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.12 ms /    22 runs   (    0.37
ms per token,  2709.03 tokens per second)
llama_print_timings: prompt eval time =    9562.64 ms /    84 tokens (  113.84
ms per token,     8.78 tokens per second)
llama_print_timings:        eval time =    5184.77 ms /    21 runs   (  246.89
ms per token,     4.05 tokens per second)
llama_print_timings:       total time =   14879.74 ms /   105 tokens
No. of rows:  57%| | 714/1258 [12:06:11<2:33:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.21 ms /    38 runs   (    0.40
ms per token,  2499.18 tokens per second)
llama_print_timings: prompt eval time =   13331.83 ms /   123 tokens (  108.39
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =    8272.66 ms /    37 runs   (  223.59
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   21838.45 ms /   160 tokens
No. of rows:  57%| | 715/1258 [12:06:33<2:46:14, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.96 ms /    24 runs   (    0.37
ms per token,  2677.97 tokens per second)
llama_print_timings: prompt eval time =   12305.45 ms /   100 tokens (  123.05
ms per token,     8.13 tokens per second)
llama_print_timings:        eval time =    4966.46 ms /    23 runs   (  215.93
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   17416.03 ms /   123 tokens
No. of rows:  57%| | 716/1258 [12:06:51<2:43:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.12 ms /    27 runs   (    0.37
ms per token,  2667.72 tokens per second)
llama_print_timings: prompt eval time =   12416.53 ms /   113 tokens (  109.88
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    5600.70 ms /    26 runs   (  215.41
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   18176.56 ms /   139 tokens
No. of rows:  57%| | 717/1258 [12:07:09<2:43:22, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.30 ms /    36 runs   (    0.40
ms per token,  2516.60 tokens per second)
llama_print_timings: prompt eval time =   13108.12 ms /   119 tokens (  110.15
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    7954.87 ms /    35 runs   (  227.28
ms per token,     4.40 tokens per second)
llama_print_timings:       total time =   21287.93 ms /   154 tokens
No. of rows:  57%| | 718/1258 [12:07:30<2:51:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.03 ms /    26 runs   (    0.39
ms per token,  2591.96 tokens per second)
llama_print_timings: prompt eval time =   13486.18 ms /   112 tokens (  120.41
ms per token,     8.30 tokens per second)
llama_print_timings:        eval time =    5411.76 ms /    25 runs   (  216.47
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   19052.00 ms /   137 tokens
No. of rows:  57%| | 719/1258 [12:07:49<2:51:18, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.66 ms /    20 runs   (    0.38
ms per token,  2611.99 tokens per second)
llama_print_timings: prompt eval time =   10906.41 ms /    85 tokens (  128.31
ms per token,     7.79 tokens per second)
llama_print_timings:        eval time =    4127.08 ms /    19 runs   (  217.21
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   15153.49 ms /   104 tokens
No. of rows:  57%| | 720/1258 [12:08:04<2:40:28, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.24 ms /    18 runs   (    0.40
ms per token,  2487.56 tokens per second)
llama_print_timings: prompt eval time =    9113.18 ms /    83 tokens (  109.80
ms per token,     9.11 tokens per second)
llama_print_timings:        eval time =    3630.07 ms /    17 runs   (  213.53
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   12856.69 ms /   100 tokens
No. of rows:  57%| | 721/1258 [12:08:17<2:26:39, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.45 ms /    20 runs   (    0.42
ms per token,  2367.42 tokens per second)
llama_print_timings: prompt eval time =    9451.69 ms /    85 tokens (  111.20
```

```
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    4103.19 ms /    19 runs   (  215.96
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   13685.94 ms /   104 tokens
No. of rows:  57%| | 722/1258 [12:08:31<2:19:10, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.64 ms /    50 runs   (    0.37
ms per token,  2682.55 tokens per second)
llama_print_timings: prompt eval time =   11299.99 ms /   102 tokens (  110.78
ms per token,     9.03 tokens per second)
llama_print_timings:        eval time =   10653.44 ms /    49 runs   (  217.42
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   22260.14 ms /   151 tokens
No. of rows:  57%| | 723/1258 [12:08:53<2:36:48, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.32 ms /    27 runs   (    0.38
ms per token,  2616.53 tokens per second)
llama_print_timings: prompt eval time =   10143.56 ms /    91 tokens (  111.47
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    5597.15 ms /    26 runs   (  215.28
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   15901.41 ms /   117 tokens
No. of rows:  58%| | 724/1258 [12:09:09<2:32:00, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.25 ms /    22 runs   (    0.38
ms per token,  2665.70 tokens per second)
llama_print_timings: prompt eval time =    9586.19 ms /    86 tokens (  111.47
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    4508.55 ms /    21 runs   (  214.69
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   14230.38 ms /   107 tokens
No. of rows:  58%| | 725/1258 [12:09:23<2:24:10, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.67 ms /    37 runs   (    0.37
ms per token,  2706.26 tokens per second)
llama_print_timings: prompt eval time =   16086.45 ms /   136 tokens (  118.28
ms per token,     8.45 tokens per second)
llama_print_timings:        eval time =    7797.06 ms /    36 runs   (  216.59
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   24106.53 ms /   172 tokens
```

```
No. of rows:  58%| | 726/1258 [12:09:47<2:44:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.62 ms /    22 runs   (    0.39
ms per token,  2551.32 tokens per second)
llama_print_timings: prompt eval time =   10302.17 ms /    81 tokens (  127.19
ms per token,     7.86 tokens per second)
llama_print_timings:        eval time =    4723.56 ms /    21 runs   (  224.93
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   15161.14 ms /   102 tokens
No. of rows:  58%| | 727/1258 [12:10:03<2:35:28, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.15 ms /    24 runs   (    0.38
ms per token,  2624.10 tokens per second)
llama_print_timings: prompt eval time =   10344.92 ms /    95 tokens (  108.89
ms per token,     9.18 tokens per second)
llama_print_timings:        eval time =    4946.40 ms /    23 runs   (  215.06
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   15434.25 ms /   118 tokens
No. of rows:  58%| | 728/1258 [12:10:18<2:29:30, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.76 ms /    28 runs   (    0.38
ms per token,  2602.23 tokens per second)
llama_print_timings: prompt eval time =   11146.42 ms /    90 tokens (  123.85
ms per token,     8.07 tokens per second)
llama_print_timings:        eval time =    7508.19 ms /    27 runs   (  278.08
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   18826.00 ms /   117 tokens
No. of rows:  58%| | 729/1258 [12:10:37<2:34:17, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.73 ms /    21 runs   (    0.37
ms per token,  2717.74 tokens per second)
llama_print_timings: prompt eval time =   10021.82 ms /    87 tokens (  115.19
ms per token,     8.68 tokens per second)
llama_print_timings:        eval time =    4397.30 ms /    20 runs   (  219.87
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   14545.30 ms /   107 tokens
No. of rows:  58%| | 730/1258 [12:10:51<2:26:13, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       7.81 ms /    21 runs   (    0.37
ms per token,  2688.86 tokens per second)
llama_print_timings: prompt eval time =    9760.97 ms /    88 tokens (  110.92
ms per token,     9.02 tokens per second)
llama_print_timings:        eval time =    4279.60 ms /    20 runs   (  213.98
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   14165.29 ms /   108 tokens
No. of rows:  58%| | 731/1258 [12:11:06<2:19:31, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.39 ms /    27 runs   (    0.38
ms per token,  2598.90 tokens per second)
llama_print_timings: prompt eval time =    8432.52 ms /    76 tokens (  110.95
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    6924.28 ms /    26 runs   (  266.32
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   15521.86 ms /   102 tokens
No. of rows:  58%| | 732/1258 [12:11:21<2:18:19, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.12 ms /    50 runs   (    0.38
ms per token,  2615.61 tokens per second)
llama_print_timings: prompt eval time =   15163.03 ms /   137 tokens (  110.68
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =   10604.75 ms /    49 runs   (  216.42
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   26069.66 ms /   186 tokens
No. of rows:  58%| | 733/1258 [12:11:47<2:45:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.21 ms /    36 runs   (    0.37
ms per token,  2724.38 tokens per second)
llama_print_timings: prompt eval time =   16559.12 ms /   139 tokens (  119.13
ms per token,     8.39 tokens per second)
llama_print_timings:        eval time =    7588.74 ms /    35 runs   (  216.82
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   24364.87 ms /   174 tokens
No. of rows:  58%| | 734/1258 [12:12:12<2:59:12, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.34 ms /    22 runs   (    0.38
ms per token,  2638.52 tokens per second)
llama_print_timings: prompt eval time =    9870.39 ms /    77 tokens (  128.19
ms per token,     7.80 tokens per second)
```

```
llama_print_timings:        eval time =    4538.96 ms /    21 runs   (  216.14
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   14542.29 ms /    98 tokens
No. of rows:  58%| | 735/1258 [12:12:26<2:43:12, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.03 ms /    29 runs   (    0.38
ms per token,  2628.48 tokens per second)
llama_print_timings: prompt eval time =   10276.08 ms /    86 tokens (  119.49
ms per token,     8.37 tokens per second)
llama_print_timings:        eval time =    7436.57 ms /    28 runs   (  265.59
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   17895.84 ms /   114 tokens
No. of rows:  59%| | 736/1258 [12:12:44<2:40:47, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.05 ms /    29 runs   (    0.38
ms per token,  2624.20 tokens per second)
llama_print_timings: prompt eval time =   12928.36 ms /   120 tokens (  107.74
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =    6081.97 ms /    28 runs   (  217.21
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   19188.47 ms /   148 tokens
No. of rows:  59%| | 737/1258 [12:13:03<2:42:19, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.86 ms /    34 runs   (    0.38
ms per token,  2643.45 tokens per second)
llama_print_timings: prompt eval time =   11646.01 ms /   104 tokens (  111.98
ms per token,     8.93 tokens per second)
llama_print_timings:        eval time =    8832.50 ms /    33 runs   (  267.65
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   20683.32 ms /   137 tokens
No. of rows:  59%| | 738/1258 [12:13:24<2:47:13, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.25 ms /    27 runs   (    0.38
ms per token,  2634.66 tokens per second)
llama_print_timings: prompt eval time =   11620.69 ms /   103 tokens (  112.82
ms per token,     8.86 tokens per second)
llama_print_timings:        eval time =    5598.08 ms /    26 runs   (  215.31
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   17381.20 ms /   129 tokens
No. of rows:  59%| | 739/1258 [12:13:41<2:41:55, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.21 ms /    33 runs   (    0.37
ms per token,  2702.92 tokens per second)
llama_print_timings: prompt eval time =   12729.72 ms /   106 tokens (  120.09
ms per token,     8.33 tokens per second)
llama_print_timings:        eval time =    7030.29 ms /    32 runs   (  219.70
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   19964.97 ms /   138 tokens
No. of rows:  59%| | 740/1258 [12:14:01<2:44:53, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.31 ms /    22 runs   (    0.38
ms per token,  2645.82 tokens per second)
llama_print_timings: prompt eval time =   10181.70 ms /    92 tokens (  110.67
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =    4513.28 ms /    21 runs   (  214.92
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   14829.43 ms /   113 tokens
No. of rows:  59%| | 741/1258 [12:14:16<2:33:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.81 ms /    32 runs   (    0.37
ms per token,  2709.80 tokens per second)
llama_print_timings: prompt eval time =   12106.79 ms /   110 tokens (  110.06
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =    6721.01 ms /    31 runs   (  216.81
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   19022.81 ms /   141 tokens
No. of rows:  59%| | 742/1258 [12:14:35<2:36:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.04 ms /    25 runs   (    0.40
ms per token,  2489.79 tokens per second)
llama_print_timings: prompt eval time =   10449.73 ms /    93 tokens (  112.36
ms per token,     8.90 tokens per second)
llama_print_timings:        eval time =    5266.41 ms /    24 runs   (  219.43
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   15873.33 ms /   117 tokens
No. of rows:  59%| | 743/1258 [12:14:51<2:30:07, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.81 ms /    29 runs   (    0.37
```

```
ms per token,  2683.20 tokens per second)
llama_print_timings: prompt eval time =    9953.76 ms /    86 tokens (  115.74
ms per token,     8.64 tokens per second)
llama_print_timings:        eval time =    6133.60 ms /    28 runs   (  219.06
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   16258.52 ms /   114 tokens
No. of rows:  59%| | 744/1258 [12:15:07<2:26:42, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.61 ms /    25 runs   (    0.38
ms per token,  2601.19 tokens per second)
llama_print_timings: prompt eval time =   10067.61 ms /    91 tokens (  110.63
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =    5210.53 ms /    24 runs   (  217.11
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   15430.62 ms /   115 tokens
No. of rows:  59%| | 745/1258 [12:15:23<2:22:05, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.09 ms /    24 runs   (    0.38
ms per token,  2640.85 tokens per second)
llama_print_timings: prompt eval time =   10795.50 ms /    98 tokens (  110.16
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    4976.48 ms /    23 runs   (  216.37
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   15915.84 ms /   121 tokens
No. of rows:  59%| | 746/1258 [12:15:39<2:20:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.95 ms /    34 runs   (    0.38
ms per token,  2625.28 tokens per second)
llama_print_timings: prompt eval time =   11756.01 ms /    96 tokens (  122.46
ms per token,     8.17 tokens per second)
llama_print_timings:        eval time =    8904.58 ms /    33 runs   (  269.84
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   20863.06 ms /   129 tokens
No. of rows:  59%| | 747/1258 [12:16:00<2:31:09, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.94 ms /    23 runs   (    0.39
ms per token,  2571.56 tokens per second)
llama_print_timings: prompt eval time =   10511.25 ms /    89 tokens (  118.10
ms per token,     8.47 tokens per second)
llama_print_timings:        eval time =    4754.68 ms /    22 runs   (  216.12
```

```
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   15408.52 ms /   111 tokens
No. of rows:  59%| | 748/1258 [12:16:15<2:24:55, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.52 ms /    20 runs   (    0.38
ms per token,  2660.28 tokens per second)
llama_print_timings: prompt eval time =   11333.82 ms /    92 tokens (  123.19
ms per token,     8.12 tokens per second)
llama_print_timings:        eval time =    4090.37 ms /    19 runs   (  215.28
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   15543.96 ms /   111 tokens
No. of rows:  60%| | 749/1258 [12:16:31<2:20:48, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.47 ms /    24 runs   (    0.39
ms per token,  2533.52 tokens per second)
llama_print_timings: prompt eval time =   11255.53 ms /    89 tokens (  126.47
ms per token,     7.91 tokens per second)
llama_print_timings:        eval time =    5143.29 ms /    23 runs   (  223.62
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   16556.00 ms /   112 tokens
No. of rows:  60%| | 750/1258 [12:16:47<2:20:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.70 ms /    21 runs   (    0.37
ms per token,  2727.27 tokens per second)
llama_print_timings: prompt eval time =   10287.37 ms /    92 tokens (  111.82
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =    4329.83 ms /    20 runs   (  216.49
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   14744.22 ms /   112 tokens
No. of rows:  60%| | 751/1258 [12:17:02<2:15:31, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.45 ms /    24 runs   (    0.39
ms per token,  2539.41 tokens per second)
llama_print_timings: prompt eval time =    8438.37 ms /    76 tokens (  111.03
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    4939.07 ms /    23 runs   (  214.74
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   13525.36 ms /    99 tokens
No. of rows:  60%| | 752/1258 [12:17:15<2:08:53, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.38 ms /    33 runs   (    0.38
ms per token,  2665.80 tokens per second)
llama_print_timings: prompt eval time =   12174.72 ms /    98 tokens (  124.23
ms per token,     8.05 tokens per second)
llama_print_timings:        eval time =    7472.12 ms /    32 runs   (  233.50
ms per token,     4.28 tokens per second)
llama_print_timings:       total time =   19849.52 ms /   130 tokens
No. of rows:  60%| | 753/1258 [12:17:35<2:20:12, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.12 ms /    29 runs   (    0.38
ms per token,  2607.91 tokens per second)
llama_print_timings: prompt eval time =   12486.00 ms /   114 tokens (  109.53
ms per token,     9.13 tokens per second)
llama_print_timings:        eval time =    6043.29 ms /    28 runs   (  215.83
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   18707.98 ms /   142 tokens
No. of rows:  60%| | 754/1258 [12:17:54<2:25:06, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.76 ms /    26 runs   (    0.38
ms per token,  2665.30 tokens per second)
llama_print_timings: prompt eval time =   12445.36 ms /   100 tokens (  124.45
ms per token,     8.04 tokens per second)
llama_print_timings:        eval time =    5415.46 ms /    25 runs   (  216.62
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   18014.99 ms /   125 tokens
No. of rows:  60%| | 755/1258 [12:18:12<2:26:42, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.66 ms /    50 runs   (    0.37
ms per token,  2680.10 tokens per second)
llama_print_timings: prompt eval time =   15420.99 ms /   131 tokens (  117.72
ms per token,     8.49 tokens per second)
llama_print_timings:        eval time =   12379.40 ms /    49 runs   (  252.64
ms per token,     3.96 tokens per second)
llama_print_timings:       total time =   28107.78 ms /   180 tokens
No. of rows:  60%| | 756/1258 [12:18:40<2:53:04, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.81 ms /    25 runs   (    0.39
ms per token,  2548.68 tokens per second)
```

```
llama_print_timings: prompt eval time =   10789.55 ms /    95 tokens (  113.57
ms per token,     8.80 tokens per second)
llama_print_timings:        eval time =    5280.80 ms /    24 runs   (  220.03
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   16227.17 ms /   119 tokens
No. of rows:  60%| | 757/1258 [12:18:56<2:41:33, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.46 ms /    50 runs   (    0.39
ms per token,  2569.24 tokens per second)
llama_print_timings: prompt eval time =   11328.70 ms /   104 tokens (  108.93
ms per token,     9.18 tokens per second)
llama_print_timings:        eval time =   10977.30 ms /    49 runs   (  224.03
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   22623.69 ms /   153 tokens
No. of rows:  60%| | 758/1258 [12:19:19<2:49:25, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.44 ms /    20 runs   (    0.37
ms per token,  2689.62 tokens per second)
llama_print_timings: prompt eval time =    9475.48 ms /    85 tokens (  111.48
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    4105.43 ms /    19 runs   (  216.08
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   13701.28 ms /   104 tokens
No. of rows:  60%| | 759/1258 [12:19:33<2:32:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.07 ms /    24 runs   (    0.38
ms per token,  2646.67 tokens per second)
llama_print_timings: prompt eval time =    9735.89 ms /    84 tokens (  115.90
ms per token,     8.63 tokens per second)
llama_print_timings:        eval time =    4955.63 ms /    23 runs   (  215.46
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14836.65 ms /   107 tokens
No. of rows:  60%| | 760/1258 [12:19:47<2:23:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.96 ms /    50 runs   (    0.40
ms per token,  2505.64 tokens per second)
llama_print_timings: prompt eval time =   17583.69 ms /   147 tokens (  119.62
ms per token,     8.36 tokens per second)
llama_print_timings:        eval time =   10706.71 ms /    49 runs   (  218.50
ms per token,     4.58 tokens per second)
```

```
llama_print_timings:       total time =   28597.25 ms /   196 tokens
No. of rows:  60%| | 761/1258 [12:20:16<2:51:23, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.20 ms /    24 runs   (    0.38
ms per token,  2609.83 tokens per second)
llama_print_timings: prompt eval time =   10097.00 ms /    80 tokens (  126.21
ms per token,     7.92 tokens per second)
llama_print_timings:        eval time =    4950.92 ms /    23 runs   (  215.26
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   15191.03 ms /   103 tokens
No. of rows:  61%| | 762/1258 [12:20:31<2:37:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.55 ms /    44 runs   (    0.38
ms per token,  2658.77 tokens per second)
llama_print_timings: prompt eval time =   12936.53 ms /   109 tokens (  118.68
ms per token,     8.43 tokens per second)
llama_print_timings:        eval time =    9583.11 ms /    43 runs   (  222.86
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   22798.02 ms /   152 tokens
No. of rows:  61%| | 763/1258 [12:20:54<2:46:25, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.33 ms /    29 runs   (    0.39
ms per token,  2560.48 tokens per second)
llama_print_timings: prompt eval time =   10844.19 ms /    98 tokens (  110.66
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =    6148.63 ms /    28 runs   (  219.59
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   17167.52 ms /   126 tokens
No. of rows:  61%| | 764/1258 [12:21:11<2:38:40, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.80 ms /    31 runs   (    0.38
ms per token,  2626.67 tokens per second)
llama_print_timings: prompt eval time =   12057.61 ms /   109 tokens (  110.62
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =    6477.87 ms /    30 runs   (  215.93
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   18720.71 ms /   139 tokens
No. of rows:  61%| | 765/1258 [12:21:30<2:37:00, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.15 ms /    31 runs   (    0.39
ms per token,  2550.81 tokens per second)
llama_print_timings: prompt eval time =   10928.58 ms /    88 tokens (  124.19
ms per token,     8.05 tokens per second)
llama_print_timings:        eval time =    6485.53 ms /    30 runs   (  216.18
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   17604.37 ms /   118 tokens
No. of rows:  61%| | 766/1258 [12:21:48<2:32:59, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.07 ms /    26 runs   (    0.39
ms per token,  2582.44 tokens per second)
llama_print_timings: prompt eval time =   10844.70 ms /    87 tokens (  124.65
ms per token,     8.02 tokens per second)
llama_print_timings:        eval time =    5537.22 ms /    25 runs   (  221.49
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   16544.42 ms /   112 tokens
No. of rows:  61%| | 767/1258 [12:22:04<2:27:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.51 ms /    35 runs   (    0.39
ms per token,  2590.67 tokens per second)
llama_print_timings: prompt eval time =   13383.03 ms /   110 tokens (  121.66
ms per token,     8.22 tokens per second)
llama_print_timings:        eval time =    7438.17 ms /    34 runs   (  218.77
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   21034.86 ms /   144 tokens
No. of rows:  61%| | 768/1258 [12:22:25<2:34:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.59 ms /    27 runs   (    0.39
ms per token,  2550.30 tokens per second)
llama_print_timings: prompt eval time =   11982.33 ms /    97 tokens (  123.53
ms per token,     8.10 tokens per second)
llama_print_timings:        eval time =    5625.68 ms /    26 runs   (  216.37
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   17772.83 ms /   123 tokens
No. of rows:  61%| | 769/1258 [12:22:43<2:31:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.95 ms /    20 runs   (    0.40
ms per token,  2514.46 tokens per second)
llama_print_timings: prompt eval time =   11904.27 ms /    96 tokens (  124.00
```

```
ms per token,     8.06 tokens per second)
llama_print_timings:        eval time =    4084.36 ms /    19 runs   (  214.97
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   16109.89 ms /   115 tokens
No. of rows:  61%| | 770/1258 [12:22:59<2:25:09, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.17 ms /    19 runs   (    0.38
ms per token,  2648.82 tokens per second)
llama_print_timings: prompt eval time =    9479.94 ms /    84 tokens (  112.86
ms per token,     8.86 tokens per second)
llama_print_timings:        eval time =    3866.27 ms /    18 runs   (  214.79
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   13458.09 ms /   102 tokens
No. of rows:  61%| | 771/1258 [12:23:13<2:14:11, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.70 ms /    22 runs   (    0.40
ms per token,  2527.86 tokens per second)
llama_print_timings: prompt eval time =    9549.99 ms /    86 tokens (  111.05
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    4526.21 ms /    21 runs   (  215.53
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14209.86 ms /   107 tokens
No. of rows:  61%| | 772/1258 [12:23:27<2:08:17, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.85 ms /    46 runs   (    0.37
ms per token,  2730.29 tokens per second)
llama_print_timings: prompt eval time =   17316.39 ms /   149 tokens (  116.22
ms per token,     8.60 tokens per second)
llama_print_timings:        eval time =    9840.34 ms /    45 runs   (  218.67
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   27434.52 ms /   194 tokens
No. of rows:  61%| | 773/1258 [12:23:54<2:36:09, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.73 ms /    27 runs   (    0.40
ms per token,  2516.78 tokens per second)
llama_print_timings: prompt eval time =    9605.21 ms /    85 tokens (  113.00
ms per token,     8.85 tokens per second)
llama_print_timings:        eval time =    5584.04 ms /    26 runs   (  214.77
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   15361.49 ms /   111 tokens
```

```
No. of rows:  62%| | 774/1258 [12:24:10<2:26:16, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.78 ms /    20 runs   (    0.39
ms per token,  2570.03 tokens per second)
llama_print_timings: prompt eval time =   10923.50 ms /    89 tokens (  122.74
ms per token,     8.15 tokens per second)
llama_print_timings:        eval time =    4101.65 ms /    19 runs   (  215.88
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   15144.87 ms /   108 tokens
No. of rows:  62%| | 775/1258 [12:24:25<2:18:47, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.58 ms /    20 runs   (    0.38
ms per token,  2637.13 tokens per second)
llama_print_timings: prompt eval time =    9559.98 ms /    85 tokens (  112.47
ms per token,     8.89 tokens per second)
llama_print_timings:        eval time =    4117.26 ms /    19 runs   (  216.70
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   13798.94 ms /   104 tokens
No. of rows:  62%| | 776/1258 [12:24:39<2:10:12, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.05 ms /    35 runs   (    0.37
ms per token,  2681.17 tokens per second)
llama_print_timings: prompt eval time =   13433.48 ms /   111 tokens (  121.02
ms per token,     8.26 tokens per second)
llama_print_timings:        eval time =    7374.46 ms /    34 runs   (  216.90
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   21020.36 ms /   145 tokens
No. of rows:  62%| | 777/1258 [12:25:00<2:21:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.10 ms /    37 runs   (    0.38
ms per token,  2623.74 tokens per second)
llama_print_timings: prompt eval time =   12551.63 ms /   113 tokens (  111.08
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =    8125.62 ms /    36 runs   (  225.71
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   20897.79 ms /   149 tokens
No. of rows:  62%| | 778/1258 [12:25:21<2:29:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       7.32 ms /    20 runs   (    0.37
ms per token,  2730.75 tokens per second)
llama_print_timings: prompt eval time =    9843.29 ms /    89 tokens (  110.60
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =    4098.04 ms /    19 runs   (  215.69
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14058.18 ms /   108 tokens
No. of rows:  62%| | 779/1258 [12:25:35<2:17:48, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.30 ms /    30 runs   (    0.38
ms per token,  2654.87 tokens per second)
llama_print_timings: prompt eval time =   11766.49 ms /    97 tokens (  121.30
ms per token,     8.24 tokens per second)
llama_print_timings:        eval time =    7783.49 ms /    29 runs   (  268.40
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   19731.54 ms /   126 tokens
No. of rows:  62%| | 780/1258 [12:25:54<2:23:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.13 ms /    34 runs   (    0.39
ms per token,  2589.10 tokens per second)
llama_print_timings: prompt eval time =   13581.01 ms /   124 tokens (  109.52
ms per token,     9.13 tokens per second)
llama_print_timings:        eval time =    7135.08 ms /    33 runs   (  216.21
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   20919.84 ms /   157 tokens
No. of rows:  62%| | 781/1258 [12:26:15<2:30:07, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.44 ms /    20 runs   (    0.37
ms per token,  2689.98 tokens per second)
llama_print_timings: prompt eval time =   10175.54 ms /    81 tokens (  125.62
ms per token,     7.96 tokens per second)
llama_print_timings:        eval time =    4395.70 ms /    19 runs   (  231.35
ms per token,     4.32 tokens per second)
llama_print_timings:       total time =   14691.54 ms /   100 tokens
No. of rows:  62%| | 782/1258 [12:26:30<2:19:50, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.68 ms /    28 runs   (    0.38
ms per token,  2622.46 tokens per second)
llama_print_timings: prompt eval time =   12162.33 ms /   111 tokens (  109.57
ms per token,     9.13 tokens per second)
```

```
llama_print_timings:        eval time =    5800.59 ms /    27 runs   (  214.84
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   18132.26 ms /   138 tokens
No. of rows:  62%| | 783/1258 [12:26:48<2:20:46, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.59 ms /    20 runs   (    0.38
ms per token,  2634.35 tokens per second)
llama_print_timings: prompt eval time =   10705.16 ms /    85 tokens (  125.94
ms per token,     7.94 tokens per second)
llama_print_timings:        eval time =    4091.18 ms /    19 runs   (  215.33
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14918.63 ms /   104 tokens
No. of rows:  62%| | 784/1258 [12:27:03<2:13:42, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.57 ms /    39 runs   (    0.40
ms per token,  2505.46 tokens per second)
llama_print_timings: prompt eval time =   13636.79 ms /   124 tokens (  109.97
ms per token,     9.09 tokens per second)
llama_print_timings:        eval time =   10063.92 ms /    38 runs   (  264.84
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   23943.03 ms /   162 tokens
No. of rows:  62%| | 785/1258 [12:27:27<2:30:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.36 ms /    33 runs   (    0.37
ms per token,  2670.34 tokens per second)
llama_print_timings: prompt eval time =   13786.52 ms /   127 tokens (  108.56
ms per token,     9.21 tokens per second)
llama_print_timings:        eval time =    6881.00 ms /    32 runs   (  215.03
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   20870.67 ms /   159 tokens
No. of rows:  62%| | 786/1258 [12:27:48<2:34:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.82 ms /    23 runs   (    0.38
ms per token,  2607.71 tokens per second)
llama_print_timings: prompt eval time =   11990.24 ms /    98 tokens (  122.35
ms per token,     8.17 tokens per second)
llama_print_timings:        eval time =    4895.92 ms /    22 runs   (  222.54
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   17029.90 ms /   120 tokens
No. of rows:  63%| | 787/1258 [12:28:05<2:27:45, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.22 ms /    32 runs   (    0.38
ms per token,  2618.87 tokens per second)
llama_print_timings: prompt eval time =   13250.42 ms /   110 tokens (  120.46
ms per token,     8.30 tokens per second)
llama_print_timings:        eval time =    6774.04 ms /    31 runs   (  218.52
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   20217.31 ms /   141 tokens
No. of rows:  63%| | 788/1258 [12:28:25<2:30:44, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.45 ms /    23 runs   (    0.37
ms per token,  2721.25 tokens per second)
llama_print_timings: prompt eval time =   11165.55 ms /    88 tokens (  126.88
ms per token,     7.88 tokens per second)
llama_print_timings:        eval time =    5938.98 ms /    22 runs   (  269.95
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   17244.25 ms /   110 tokens
No. of rows:  63%| | 789/1258 [12:28:42<2:25:43, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.13 ms /    35 runs   (    0.38
ms per token,  2665.04 tokens per second)
llama_print_timings: prompt eval time =   13640.27 ms /   125 tokens (  109.12
ms per token,     9.16 tokens per second)
llama_print_timings:        eval time =    7364.51 ms /    34 runs   (  216.60
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   21213.74 ms /   159 tokens
No. of rows:  63%| | 790/1258 [12:29:04<2:31:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.29 ms /    25 runs   (    0.37
ms per token,  2691.93 tokens per second)
llama_print_timings: prompt eval time =    9446.57 ms /    84 tokens (  112.46
ms per token,     8.89 tokens per second)
llama_print_timings:        eval time =    5143.33 ms /    24 runs   (  214.31
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   14743.98 ms /   108 tokens
No. of rows:  63%| | 791/1258 [12:29:18<2:20:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.22 ms /    26 runs   (    0.39
```

```
ms per token,  2543.28 tokens per second)
llama_print_timings: prompt eval time =   12444.87 ms /   101 tokens (  123.22
ms per token,     8.12 tokens per second)
llama_print_timings:        eval time =    5346.56 ms /    25 runs   (  213.86
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   17955.88 ms /   126 tokens
No. of rows:  63%| | 792/1258 [12:29:36<2:19:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.42 ms /    37 runs   (    0.39
ms per token,  2566.41 tokens per second)
llama_print_timings: prompt eval time =   14036.58 ms /   119 tokens (  117.95
ms per token,     8.48 tokens per second)
llama_print_timings:        eval time =    9536.35 ms /    36 runs   (  264.90
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   23802.91 ms /   155 tokens
No. of rows:  63%| | 793/1258 [12:30:00<2:32:59, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.99 ms /    21 runs   (    0.38
ms per token,  2627.63 tokens per second)
llama_print_timings: prompt eval time =    9487.30 ms /    85 tokens (  111.62
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =    4300.66 ms /    20 runs   (  215.03
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   13913.82 ms /   105 tokens
No. of rows:  63%| | 794/1258 [12:30:14<2:19:09, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.55 ms /    23 runs   (    0.37
ms per token,  2688.49 tokens per second)
llama_print_timings: prompt eval time =   10687.34 ms /    87 tokens (  122.84
ms per token,     8.14 tokens per second)
llama_print_timings:        eval time =    5311.64 ms /    22 runs   (  241.44
ms per token,     4.14 tokens per second)
llama_print_timings:       total time =   16136.58 ms /   109 tokens
No. of rows:  63%| | 795/1258 [12:30:30<2:14:34, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.75 ms /    23 runs   (    0.38
ms per token,  2630.07 tokens per second)
llama_print_timings: prompt eval time =    9998.11 ms /    90 tokens (  111.09
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =    4846.45 ms /    22 runs   (  220.29
```

```
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   14984.20 ms /   112 tokens
No. of rows:  63%| | 796/1258 [12:30:45<2:08:38, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.09 ms /    22 runs   (    0.37
ms per token,  2719.07 tokens per second)
llama_print_timings: prompt eval time =    9884.72 ms /    89 tokens (  111.06
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =    4616.38 ms /    21 runs   (  219.83
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   14634.06 ms /   110 tokens
No. of rows:  63%| | 797/1258 [12:31:00<2:03:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.97 ms /    32 runs   (    0.37
ms per token,  2673.13 tokens per second)
llama_print_timings: prompt eval time =   11919.31 ms /   110 tokens (  108.36
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =    7023.82 ms /    31 runs   (  226.57
ms per token,     4.41 tokens per second)
llama_print_timings:       total time =   19135.19 ms /   141 tokens
No. of rows:  63%| | 798/1258 [12:31:19<2:10:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.83 ms /    28 runs   (    0.39
ms per token,  2584.46 tokens per second)
llama_print_timings: prompt eval time =   10489.90 ms /    95 tokens (  110.42
ms per token,     9.06 tokens per second)
llama_print_timings:        eval time =    7598.68 ms /    27 runs   (  281.43
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   18261.72 ms /   122 tokens
No. of rows:  64%| | 799/1258 [12:31:37<2:12:57, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.90 ms /    22 runs   (    0.36
ms per token,  2784.11 tokens per second)
llama_print_timings: prompt eval time =   10501.58 ms /    95 tokens (  110.54
ms per token,     9.05 tokens per second)
llama_print_timings:        eval time =    6156.91 ms /    21 runs   (  293.19
ms per token,     3.41 tokens per second)
llama_print_timings:       total time =   16789.85 ms /   116 tokens
No. of rows:  64%| | 800/1258 [12:31:54<2:11:21, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.10 ms /    24 runs   (    0.38
ms per token,  2636.78 tokens per second)
llama_print_timings: prompt eval time =   10349.86 ms /    91 tokens (  113.73
ms per token,     8.79 tokens per second)
llama_print_timings:        eval time =    4968.13 ms /    23 runs   (  216.01
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   15462.43 ms /   114 tokens
No. of rows:  64%| | 801/1258 [12:32:09<2:07:05, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.49 ms /    42 runs   (    0.46
ms per token,  2154.84 tokens per second)
llama_print_timings: prompt eval time =   12868.82 ms /   116 tokens (  110.94
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    8895.11 ms /    41 runs   (  216.95
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   22018.99 ms /   157 tokens
No. of rows:  64%| | 802/1258 [12:32:31<2:18:58, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.47 ms /    19 runs   (    0.39
ms per token,  2544.87 tokens per second)
llama_print_timings: prompt eval time =    9499.96 ms /    85 tokens (  111.76
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    3856.38 ms /    18 runs   (  214.24
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   13470.55 ms /   103 tokens
No. of rows:  64%| | 803/1258 [12:32:45<2:07:45, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.23 ms /    24 runs   (    0.38
ms per token,  2601.06 tokens per second)
llama_print_timings: prompt eval time =   11033.06 ms /    88 tokens (  125.38
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =    5025.47 ms /    23 runs   (  218.50
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   16203.82 ms /   111 tokens
No. of rows:  64%| | 804/1258 [12:33:01<2:06:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.01 ms /    30 runs   (    0.37
ms per token,  2725.29 tokens per second)
```

```
llama_print_timings: prompt eval time =   10223.11 ms /    94 tokens (  108.76
ms per token,     9.19 tokens per second)
llama_print_timings:        eval time =    6207.92 ms /    29 runs   (  214.07
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   16607.42 ms /   123 tokens
No. of rows:  64%| | 805/1258 [12:33:18<2:05:39, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.92 ms /    29 runs   (    0.38
ms per token,  2655.92 tokens per second)
llama_print_timings: prompt eval time =    9744.94 ms /    88 tokens (  110.74
ms per token,     9.03 tokens per second)
llama_print_timings:        eval time =    6003.14 ms /    28 runs   (  214.40
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   15920.87 ms /   116 tokens
No. of rows:  64%| | 806/1258 [12:33:34<2:03:44, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.21 ms /    24 runs   (    0.38
ms per token,  2604.45 tokens per second)
llama_print_timings: prompt eval time =   12028.90 ms /    97 tokens (  124.01
ms per token,     8.06 tokens per second)
llama_print_timings:        eval time =    4916.73 ms /    23 runs   (  213.77
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   17091.25 ms /   120 tokens
No. of rows:  64%| | 807/1258 [12:33:51<2:04:59, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.53 ms /    43 runs   (    0.41
ms per token,  2452.66 tokens per second)
llama_print_timings: prompt eval time =   14627.00 ms /   123 tokens (  118.92
ms per token,     8.41 tokens per second)
llama_print_timings:        eval time =   10962.89 ms /    42 runs   (  261.02
ms per token,     3.83 tokens per second)
llama_print_timings:       total time =   25863.90 ms /   165 tokens
No. of rows:  64%| | 808/1258 [12:34:17<2:25:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.00 ms /    21 runs   (    0.38
ms per token,  2625.98 tokens per second)
llama_print_timings: prompt eval time =    9657.04 ms /    87 tokens (  111.00
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    4273.58 ms /    20 runs   (  213.68
ms per token,     4.68 tokens per second)
```

```
llama_print_timings:       total time =   14055.56 ms /   107 tokens
No. of rows:  64%| | 809/1258 [12:34:31<2:13:13, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.86 ms /    28 runs   (    0.39
ms per token,  2579.46 tokens per second)
llama_print_timings: prompt eval time =   11871.74 ms /   107 tokens (  110.95
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    5804.64 ms /    27 runs   (  214.99
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   17847.57 ms /   134 tokens
No. of rows:  64%| | 810/1258 [12:34:49<2:13:03, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.34 ms /    22 runs   (    0.38
ms per token,  2638.84 tokens per second)
llama_print_timings: prompt eval time =    9918.26 ms /    89 tokens (  111.44
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    4538.72 ms /    21 runs   (  216.13
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   14589.01 ms /   110 tokens
No. of rows:  64%| | 811/1258 [12:35:03<2:05:33, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.48 ms /    27 runs   (    0.39
ms per token,  2576.58 tokens per second)
llama_print_timings: prompt eval time =   12083.02 ms /   110 tokens (  109.85
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    5620.03 ms /    26 runs   (  216.16
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   17871.91 ms /   136 tokens
No. of rows:  65%| | 812/1258 [12:35:21<2:07:33, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.50 ms /    30 runs   (    0.38
ms per token,  2608.92 tokens per second)
llama_print_timings: prompt eval time =   10479.35 ms /    93 tokens (  112.68
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    6233.87 ms /    29 runs   (  214.96
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   16896.66 ms /   122 tokens
No. of rows:  65%| | 813/1258 [12:35:38<2:06:42, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.88 ms /    37 runs   (    0.46
ms per token,  2192.59 tokens per second)
llama_print_timings: prompt eval time =   13821.17 ms /   115 tokens (  120.18
ms per token,     8.32 tokens per second)
llama_print_timings:        eval time =    9508.66 ms /    36 runs   (  264.13
ms per token,     3.79 tokens per second)
llama_print_timings:       total time =   23556.73 ms /   151 tokens
No. of rows:  65%| | 814/1258 [12:36:02<2:20:47, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.63 ms /    50 runs   (    0.37
ms per token,  2683.99 tokens per second)
llama_print_timings: prompt eval time =   11329.20 ms /   102 tokens (  111.07
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =   10577.39 ms /    49 runs   (  215.87
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   22201.44 ms /   151 tokens
No. of rows:  65%| | 815/1258 [12:36:24<2:27:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.18 ms /    21 runs   (    0.39
ms per token,  2567.55 tokens per second)
llama_print_timings: prompt eval time =   11380.75 ms /    92 tokens (  123.70
ms per token,     8.08 tokens per second)
llama_print_timings:        eval time =    4284.82 ms /    20 runs   (  214.24
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   15791.32 ms /   112 tokens
No. of rows:  65%| | 816/1258 [12:36:40<2:17:57, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.31 ms /    42 runs   (    0.36
ms per token,  2742.59 tokens per second)
llama_print_timings: prompt eval time =   14084.28 ms /   126 tokens (  111.78
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    9109.05 ms /    41 runs   (  222.17
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   23448.62 ms /   167 tokens
No. of rows:  65%| | 817/1258 [12:37:03<2:28:04, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.17 ms /    18 runs   (    0.40
ms per token,  2509.41 tokens per second)
llama_print_timings: prompt eval time =    8785.61 ms /    79 tokens (  111.21
```

```
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    3659.74 ms /    17 runs   (  215.28
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   12556.82 ms /    96 tokens
No. of rows:  65%| | 818/1258 [12:37:16<2:11:03, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.45 ms /    26 runs   (    0.36
ms per token,  2752.20 tokens per second)
llama_print_timings: prompt eval time =   10393.13 ms /    93 tokens (  111.75
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    6718.71 ms /    25 runs   (  268.75
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   17264.59 ms /   118 tokens
No. of rows:  65%| | 819/1258 [12:37:33<2:09:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.49 ms /    30 runs   (    0.38
ms per token,  2610.28 tokens per second)
llama_print_timings: prompt eval time =   11093.64 ms /    99 tokens (  112.06
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    8053.36 ms /    29 runs   (  277.70
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   19332.40 ms /   128 tokens
No. of rows:  65%| | 820/1258 [12:37:52<2:12:44, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.89 ms /    37 runs   (    0.40
ms per token,  2484.39 tokens per second)
llama_print_timings: prompt eval time =   14207.23 ms /   129 tokens (  110.13
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    7759.03 ms /    36 runs   (  215.53
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   22193.65 ms /   165 tokens
No. of rows:  65%| | 821/1258 [12:38:14<2:21:13, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.71 ms /    26 runs   (    0.37
ms per token,  2676.55 tokens per second)
llama_print_timings: prompt eval time =   11274.18 ms /    91 tokens (  123.89
ms per token,     8.07 tokens per second)
llama_print_timings:        eval time =    6662.08 ms /    25 runs   (  266.48
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   18099.11 ms /   116 tokens
```

```
No. of rows:  65%| | 822/1258 [12:38:33<2:18:06, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.80 ms /    29 runs   (    0.37
ms per token,  2685.19 tokens per second)
llama_print_timings: prompt eval time =   11150.72 ms /    98 tokens (  113.78
ms per token,     8.79 tokens per second)
llama_print_timings:        eval time =    7749.24 ms /    28 runs   (  276.76
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   19074.30 ms /   126 tokens
No. of rows:  65%| | 823/1258 [12:38:52<2:17:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.69 ms /    28 runs   (    0.38
ms per token,  2619.76 tokens per second)
llama_print_timings: prompt eval time =   13357.79 ms /   118 tokens (  113.20
ms per token,     8.83 tokens per second)
llama_print_timings:        eval time =    5898.86 ms /    27 runs   (  218.48
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   19427.27 ms /   145 tokens
No. of rows:  66%| | 824/1258 [12:39:11<2:18:30, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.39 ms /    22 runs   (    0.38
ms per token,  2620.92 tokens per second)
llama_print_timings: prompt eval time =   10063.41 ms /    90 tokens (  111.82
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =    4577.50 ms /    21 runs   (  217.98
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   14774.44 ms /   111 tokens
No. of rows:  66%| | 825/1258 [12:39:26<2:08:45, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.51 ms /    20 runs   (    0.38
ms per token,  2663.47 tokens per second)
llama_print_timings: prompt eval time =    8705.71 ms /    79 tokens (  110.20
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    4070.93 ms /    19 runs   (  214.26
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   12897.09 ms /    98 tokens
No. of rows:  66%| | 826/1258 [12:39:39<1:57:47, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =      13.82 ms /    33 runs   (    0.42
ms per token,  2387.84 tokens per second)
llama_print_timings: prompt eval time =   12511.65 ms /   102 tokens (  122.66
ms per token,     8.15 tokens per second)
llama_print_timings:        eval time =    7529.90 ms /    32 runs   (  235.31
ms per token,     4.25 tokens per second)
llama_print_timings:       total time =   20269.69 ms /   134 tokens
No. of rows:  66%| | 827/1258 [12:39:59<2:05:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.82 ms /    25 runs   (    0.39
ms per token,  2546.08 tokens per second)
llama_print_timings: prompt eval time =   11343.46 ms /    89 tokens (  127.45
ms per token,     7.85 tokens per second)
llama_print_timings:        eval time =    6736.92 ms /    24 runs   (  280.70
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   18230.94 ms /   113 tokens
No. of rows:  66%| | 828/1258 [12:40:17<2:07:10, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.58 ms /    29 runs   (    0.40
ms per token,  2504.32 tokens per second)
llama_print_timings: prompt eval time =   12753.78 ms /   116 tokens (  109.95
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    7994.42 ms /    28 runs   (  285.52
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   20928.96 ms /   144 tokens
No. of rows:  66%| | 829/1258 [12:40:38<2:13:43, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.67 ms /    50 runs   (    0.39
ms per token,  2541.43 tokens per second)
llama_print_timings: prompt eval time =   16093.32 ms /   148 tokens (  108.74
ms per token,     9.20 tokens per second)
llama_print_timings:        eval time =   10894.21 ms /    49 runs   (  222.33
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   27295.05 ms /   197 tokens
No. of rows:  66%| | 830/1258 [12:41:05<2:31:48, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.04 ms /    21 runs   (    0.43
ms per token,  2322.24 tokens per second)
llama_print_timings: prompt eval time =    9356.53 ms /    82 tokens (  114.10
ms per token,     8.76 tokens per second)
```

```
llama_print_timings:        eval time =    4695.14 ms /    20 runs   (  234.76
ms per token,     4.26 tokens per second)
llama_print_timings:       total time =   14190.50 ms /   102 tokens
No. of rows:  66%| | 831/1258 [12:41:20<2:16:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.92 ms /    24 runs   (    0.37
ms per token,  2690.58 tokens per second)
llama_print_timings: prompt eval time =   11544.51 ms /   103 tokens (  112.08
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    5022.52 ms /    23 runs   (  218.37
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   16713.66 ms /   126 tokens
No. of rows:  66%| | 832/1258 [12:41:36<2:10:50, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.20 ms /    23 runs   (    0.40
ms per token,  2499.19 tokens per second)
llama_print_timings: prompt eval time =   10406.97 ms /    93 tokens (  111.90
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =    4787.52 ms /    22 runs   (  217.61
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   15337.04 ms /   115 tokens
No. of rows:  66%| | 833/1258 [12:41:52<2:03:58, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.04 ms /    21 runs   (    0.38
ms per token,  2611.94 tokens per second)
llama_print_timings: prompt eval time =    9624.27 ms /    86 tokens (  111.91
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =    4417.79 ms /    20 runs   (  220.89
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   14173.21 ms /   106 tokens
No. of rows:  66%| | 834/1258 [12:42:06<1:56:39, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.46 ms /    20 runs   (    0.37
ms per token,  2679.17 tokens per second)
llama_print_timings: prompt eval time =    9813.40 ms /    87 tokens (  112.80
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    4224.53 ms /    19 runs   (  222.34
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   14159.59 ms /   106 tokens
No. of rows:  66%| | 835/1258 [12:42:20<1:51:25, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.27 ms /    26 runs   (    0.39
ms per token,  2532.14 tokens per second)
llama_print_timings: prompt eval time =   12862.74 ms /   102 tokens (  126.11
ms per token,     7.93 tokens per second)
llama_print_timings:        eval time =    7505.75 ms /    25 runs   (  300.23
ms per token,     3.33 tokens per second)
llama_print_timings:       total time =   20533.27 ms /   127 tokens
No. of rows:  66%| | 836/1258 [12:42:41<2:01:09, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.85 ms /    28 runs   (    0.42
ms per token,  2362.27 tokens per second)
llama_print_timings: prompt eval time =   11258.84 ms /    95 tokens (  118.51
ms per token,     8.44 tokens per second)
llama_print_timings:        eval time =    6487.36 ms /    27 runs   (  240.27
ms per token,     4.16 tokens per second)
llama_print_timings:       total time =   17938.32 ms /   122 tokens
No. of rows:  67%| | 837/1258 [12:42:59<2:02:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.69 ms /    24 runs   (    0.40
ms per token,  2478.06 tokens per second)
llama_print_timings: prompt eval time =   10785.48 ms /    93 tokens (  115.97
ms per token,     8.62 tokens per second)
llama_print_timings:        eval time =    5156.18 ms /    23 runs   (  224.18
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   16093.89 ms /   116 tokens
No. of rows:  67%| | 838/1258 [12:43:15<1:59:17, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.38 ms /    29 runs   (    0.39
ms per token,  2548.11 tokens per second)
llama_print_timings: prompt eval time =   11833.83 ms /   100 tokens (  118.34
ms per token,     8.45 tokens per second)
llama_print_timings:        eval time =    6191.27 ms /    28 runs   (  221.12
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   18206.00 ms /   128 tokens
No. of rows:  67%| | 839/1258 [12:43:33<2:01:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.94 ms /    20 runs   (    0.40
```

```
ms per token,  2518.26 tokens per second)
llama_print_timings: prompt eval time =   10914.93 ms /    85 tokens (  128.41
ms per token,     7.79 tokens per second)
llama_print_timings:        eval time =    4213.72 ms /    19 runs   (  221.77
ms per token,     4.51 tokens per second)
llama_print_timings:       total time =   15256.83 ms /   104 tokens
No. of rows:  67%| | 840/1258 [12:43:48<1:56:42, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.54 ms /    32 runs   (    0.39
ms per token,  2551.63 tokens per second)
llama_print_timings: prompt eval time =   12314.15 ms /   107 tokens (  115.09
ms per token,     8.69 tokens per second)
llama_print_timings:        eval time =    8735.06 ms /    31 runs   (  281.78
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   21255.80 ms /   138 tokens
No. of rows:  67%| | 841/1258 [12:44:09<2:05:50, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.46 ms /    29 runs   (    0.40
ms per token,  2529.66 tokens per second)
llama_print_timings: prompt eval time =   12033.62 ms /   107 tokens (  112.46
ms per token,     8.89 tokens per second)
llama_print_timings:        eval time =    6192.92 ms /    28 runs   (  221.18
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   18409.03 ms /   135 tokens
No. of rows:  67%| | 842/1258 [12:44:28<2:06:10, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.18 ms /    22 runs   (    0.42
ms per token,  2397.56 tokens per second)
llama_print_timings: prompt eval time =   11967.70 ms /    93 tokens (  128.68
ms per token,     7.77 tokens per second)
llama_print_timings:        eval time =    6627.10 ms /    21 runs   (  315.58
ms per token,     3.17 tokens per second)
llama_print_timings:       total time =   18751.08 ms /   114 tokens
No. of rows:  67%| | 843/1258 [12:44:47<2:07:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.38 ms /    28 runs   (    0.41
ms per token,  2460.02 tokens per second)
llama_print_timings: prompt eval time =   11829.23 ms /   104 tokens (  113.74
ms per token,     8.79 tokens per second)
llama_print_timings:        eval time =    6093.03 ms /    27 runs   (  225.67
```

```
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   18098.72 ms /   131 tokens
No. of rows:  67%| | 844/1258 [12:45:05<2:06:11, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.52 ms /    27 runs   (    0.39
ms per token,  2566.30 tokens per second)
llama_print_timings: prompt eval time =   10543.11 ms /    95 tokens (  110.98
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    5710.26 ms /    26 runs   (  219.63
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   16420.36 ms /   121 tokens
No. of rows:  67%| | 845/1258 [12:45:21<2:02:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.19 ms /    29 runs   (    0.39
ms per token,  2592.06 tokens per second)
llama_print_timings: prompt eval time =   12569.10 ms /   114 tokens (  110.26
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    6077.72 ms /    28 runs   (  217.06
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   18824.43 ms /   142 tokens
No. of rows:  67%| | 846/1258 [12:45:40<2:04:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.79 ms /    25 runs   (    0.39
ms per token,  2553.89 tokens per second)
llama_print_timings: prompt eval time =   10212.23 ms /    82 tokens (  124.54
ms per token,     8.03 tokens per second)
llama_print_timings:        eval time =    5145.84 ms /    24 runs   (  214.41
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   15512.22 ms /   106 tokens
No. of rows:  67%| | 847/1258 [12:45:56<1:58:31, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.54 ms /    18 runs   (    0.53
ms per token,  1886.40 tokens per second)
llama_print_timings: prompt eval time =   11167.18 ms /    89 tokens (  125.47
ms per token,     7.97 tokens per second)
llama_print_timings:        eval time =    3709.77 ms /    17 runs   (  218.22
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   14989.92 ms /   106 tokens
No. of rows:  67%| | 848/1258 [12:46:10<1:53:30, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.17 ms /    24 runs   (    0.38
ms per token,  2618.37 tokens per second)
llama_print_timings: prompt eval time =   10743.83 ms /    88 tokens (  122.09
ms per token,     8.19 tokens per second)
llama_print_timings:        eval time =    4930.98 ms /    23 runs   (  214.39
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   15817.73 ms /   111 tokens
No. of rows:  67%| | 849/1258 [12:46:26<1:51:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.50 ms /    26 runs   (    0.40
ms per token,  2475.95 tokens per second)
llama_print_timings: prompt eval time =   10982.17 ms /    97 tokens (  113.22
ms per token,     8.83 tokens per second)
llama_print_timings:        eval time =    5714.07 ms /    25 runs   (  228.56
ms per token,     4.38 tokens per second)
llama_print_timings:       total time =   16855.63 ms /   122 tokens
No. of rows:  68%| | 850/1258 [12:46:43<1:52:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      23.25 ms /    34 runs   (    0.68
ms per token,  1462.05 tokens per second)
llama_print_timings: prompt eval time =   14591.40 ms /   113 tokens (  129.13
ms per token,     7.74 tokens per second)
llama_print_timings:        eval time =   10814.01 ms /    33 runs   (  327.70
ms per token,     3.05 tokens per second)
llama_print_timings:       total time =   25738.18 ms /   146 tokens
No. of rows:  68%| | 851/1258 [12:47:09<2:10:50, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.30 ms /    34 runs   (    0.51
ms per token,  1964.86 tokens per second)
llama_print_timings: prompt eval time =   10904.07 ms /    96 tokens (  113.58
ms per token,     8.80 tokens per second)
llama_print_timings:        eval time =    9383.97 ms /    33 runs   (  284.36
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   20569.51 ms /   129 tokens
No. of rows:  68%| | 852/1258 [12:47:30<2:13:08, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.82 ms /    21 runs   (    0.42
ms per token,  2381.49 tokens per second)
```

```
llama_print_timings: prompt eval time =    9670.22 ms /    84 tokens (  115.12
ms per token,     8.69 tokens per second)
llama_print_timings:        eval time =    5465.09 ms /    20 runs   (  273.25
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   15276.50 ms /   104 tokens
No. of rows:  68%| | 853/1258 [12:47:45<2:03:55, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.30 ms /    21 runs   (    0.68
ms per token,  1468.12 tokens per second)
llama_print_timings: prompt eval time =   10993.24 ms /    85 tokens (  129.33
ms per token,     7.73 tokens per second)
llama_print_timings:        eval time =    6332.25 ms /    20 runs   (  316.61
ms per token,     3.16 tokens per second)
llama_print_timings:       total time =   17518.01 ms /   105 tokens
No. of rows:  68%| | 854/1258 [12:48:02<2:01:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.29 ms /    22 runs   (    0.42
ms per token,  2368.14 tokens per second)
llama_print_timings: prompt eval time =   11448.72 ms /    89 tokens (  128.64
ms per token,     7.77 tokens per second)
llama_print_timings:        eval time =    5077.03 ms /    21 runs   (  241.76
ms per token,     4.14 tokens per second)
llama_print_timings:       total time =   16674.87 ms /   110 tokens
No. of rows:  68%| | 855/1258 [12:48:19<1:58:45, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.42 ms /    26 runs   (    0.48
ms per token,  2093.40 tokens per second)
llama_print_timings: prompt eval time =   10595.42 ms /    88 tokens (  120.40
ms per token,     8.31 tokens per second)
llama_print_timings:        eval time =    8192.02 ms /    25 runs   (  327.68
ms per token,     3.05 tokens per second)
llama_print_timings:       total time =   18974.66 ms /   113 tokens
No. of rows:  68%| | 856/1258 [12:48:38<2:01:05, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.50 ms /    22 runs   (    0.52
ms per token,  1913.21 tokens per second)
llama_print_timings: prompt eval time =    9368.56 ms /    78 tokens (  120.11
ms per token,     8.33 tokens per second)
llama_print_timings:        eval time =    7151.09 ms /    21 runs   (  340.53
ms per token,     2.94 tokens per second)
```

```
llama_print_timings:       total time =   16699.01 ms /    99 tokens
No. of rows:  68%| | 857/1258 [12:48:55<1:58:03, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.67 ms /    19 runs   (    0.46
ms per token,  2190.96 tokens per second)
llama_print_timings: prompt eval time =   12621.26 ms /    90 tokens (  140.24
ms per token,     7.13 tokens per second)
llama_print_timings:        eval time =    5042.35 ms /    18 runs   (  280.13
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   17805.47 ms /   108 tokens
No. of rows:  68%| | 858/1258 [12:49:13<1:58:03, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.01 ms /    20 runs   (    0.50
ms per token,  1998.20 tokens per second)
llama_print_timings: prompt eval time =   11723.09 ms /    89 tokens (  131.72
ms per token,     7.59 tokens per second)
llama_print_timings:        eval time =    5106.84 ms /    19 runs   (  268.78
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   17006.88 ms /   108 tokens
No. of rows:  68%| | 859/1258 [12:49:30<1:56:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.30 ms /    24 runs   (    0.47
ms per token,  2123.33 tokens per second)
llama_print_timings: prompt eval time =   10817.30 ms /    84 tokens (  128.78
ms per token,     7.77 tokens per second)
llama_print_timings:        eval time =    5508.26 ms /    23 runs   (  239.49
ms per token,     4.18 tokens per second)
llama_print_timings:       total time =   16498.99 ms /   107 tokens
No. of rows:  68%| | 860/1258 [12:49:46<1:54:07, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.30 ms /    28 runs   (    0.44
ms per token,  2276.79 tokens per second)
llama_print_timings: prompt eval time =   11028.36 ms /   103 tokens (  107.07
ms per token,     9.34 tokens per second)
llama_print_timings:        eval time =    6747.32 ms /    27 runs   (  249.90
ms per token,     4.00 tokens per second)
llama_print_timings:       total time =   17971.20 ms /   130 tokens
No. of rows:  68%| | 861/1258 [12:50:04<1:55:22, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.98 ms /    28 runs   (    0.43
ms per token,  2337.81 tokens per second)
llama_print_timings: prompt eval time =   11192.37 ms /    98 tokens (  114.21
ms per token,     8.76 tokens per second)
llama_print_timings:        eval time =    6122.56 ms /    27 runs   (  226.76
ms per token,     4.41 tokens per second)
llama_print_timings:       total time =   17504.05 ms /   125 tokens
No. of rows:  69%| | 862/1258 [12:50:22<1:55:13, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.86 ms /    27 runs   (    0.44
ms per token,  2277.52 tokens per second)
llama_print_timings: prompt eval time =   11982.03 ms /    96 tokens (  124.81
ms per token,     8.01 tokens per second)
llama_print_timings:        eval time =    7829.18 ms /    26 runs   (  301.12
ms per token,     3.32 tokens per second)
llama_print_timings:       total time =   20004.66 ms /   122 tokens
No. of rows:  69%| | 863/1258 [12:50:42<1:59:58, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.13 ms /    20 runs   (    0.51
ms per token,  1974.72 tokens per second)
llama_print_timings: prompt eval time =   11757.00 ms /    76 tokens (  154.70
ms per token,     6.46 tokens per second)
llama_print_timings:        eval time =    4938.75 ms /    19 runs   (  259.93
ms per token,     3.85 tokens per second)
llama_print_timings:       total time =   16848.27 ms /    95 tokens
No. of rows:  69%| | 864/1258 [12:50:58<1:57:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.20 ms /    28 runs   (    0.51
ms per token,  1971.41 tokens per second)
llama_print_timings: prompt eval time =   15413.92 ms /    96 tokens (  160.56
ms per token,     6.23 tokens per second)
llama_print_timings:        eval time =    7175.29 ms /    27 runs   (  265.75
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   22818.94 ms /   123 tokens
No. of rows:  69%| | 865/1258 [12:51:21<2:06:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.03 ms /    25 runs   (    0.48
ms per token,  2077.45 tokens per second)
llama_print_timings: prompt eval time =   12379.87 ms /    98 tokens (  126.33
```

```
ms per token,     7.92 tokens per second)
llama_print_timings:        eval time =    6144.57 ms /    24 runs   (  256.02
ms per token,     3.91 tokens per second)
llama_print_timings:       total time =   18711.92 ms /   122 tokens
No. of rows:  69%| | 866/1258 [12:51:40<2:05:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.52 ms /    27 runs   (    0.43
ms per token,  2342.94 tokens per second)
llama_print_timings: prompt eval time =   12022.18 ms /    94 tokens (  127.90
ms per token,     7.82 tokens per second)
llama_print_timings:        eval time =    7765.33 ms /    26 runs   (  298.67
ms per token,     3.35 tokens per second)
llama_print_timings:       total time =   19966.61 ms /   120 tokens
No. of rows:  69%| | 867/1258 [12:52:00<2:06:23, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      20.64 ms /    50 runs   (    0.41
ms per token,  2422.60 tokens per second)
llama_print_timings: prompt eval time =   11026.60 ms /    97 tokens (  113.68
ms per token,     8.80 tokens per second)
llama_print_timings:        eval time =   11362.51 ms /    49 runs   (  231.89
ms per token,     4.31 tokens per second)
llama_print_timings:       total time =   22728.33 ms /   146 tokens
No. of rows:  69%| | 868/1258 [12:52:23<2:12:34, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.28 ms /    32 runs   (    0.45
ms per token,  2240.90 tokens per second)
llama_print_timings: prompt eval time =   12995.98 ms /   105 tokens (  123.77
ms per token,     8.08 tokens per second)
llama_print_timings:        eval time =    9615.00 ms /    31 runs   (  310.16
ms per token,     3.22 tokens per second)
llama_print_timings:       total time =   22841.77 ms /   136 tokens
No. of rows:  69%| | 869/1258 [12:52:46<2:17:00, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.33 ms /    27 runs   (    0.46
ms per token,  2190.14 tokens per second)
llama_print_timings: prompt eval time =   13152.34 ms /    99 tokens (  132.85
ms per token,     7.53 tokens per second)
llama_print_timings:        eval time =    7712.64 ms /    26 runs   (  296.64
ms per token,     3.37 tokens per second)
llama_print_timings:       total time =   21045.46 ms /   125 tokens
```

```
No. of rows:  69%| | 870/1258 [12:53:07<2:16:30, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.75 ms /    29 runs   (    0.41
ms per token,  2467.25 tokens per second)
llama_print_timings: prompt eval time =   10579.09 ms /   106 tokens (   99.80
ms per token,    10.02 tokens per second)
llama_print_timings:        eval time =    6022.09 ms /    28 runs   (  215.07
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   16789.68 ms /   134 tokens
No. of rows:  69%| | 871/1258 [12:53:23<2:07:48, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.67 ms /    27 runs   (    0.40
ms per token,  2531.17 tokens per second)
llama_print_timings: prompt eval time =    9217.65 ms /    93 tokens (   99.11
ms per token,    10.09 tokens per second)
llama_print_timings:        eval time =    5544.38 ms /    26 runs   (  213.25
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   14933.45 ms /   119 tokens
No. of rows:  69%| | 872/1258 [12:53:38<1:58:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.89 ms /    22 runs   (    0.40
ms per token,  2474.69 tokens per second)
llama_print_timings: prompt eval time =    9565.21 ms /    95 tokens (  100.69
ms per token,     9.93 tokens per second)
llama_print_timings:        eval time =    4490.41 ms /    21 runs   (  213.83
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   14190.89 ms /   116 tokens
No. of rows:  69%| | 873/1258 [12:53:53<1:49:45, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.58 ms /    26 runs   (    0.41
ms per token,  2458.40 tokens per second)
llama_print_timings: prompt eval time =    9776.04 ms /    83 tokens (  117.78
ms per token,     8.49 tokens per second)
llama_print_timings:        eval time =    5812.10 ms /    25 runs   (  232.48
ms per token,     4.30 tokens per second)
llama_print_timings:       total time =   15759.47 ms /   108 tokens
No. of rows:  69%| | 874/1258 [12:54:08<1:46:54, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       9.43 ms /    23 runs   (    0.41
ms per token,  2437.99 tokens per second)
llama_print_timings: prompt eval time =    9256.75 ms /    95 tokens (   97.44
ms per token,    10.26 tokens per second)
llama_print_timings:        eval time =    4736.61 ms /    22 runs   (  215.30
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14140.53 ms /   117 tokens
No. of rows:  70%| | 875/1258 [12:54:22<1:41:44, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.09 ms /    29 runs   (    0.42
ms per token,  2399.27 tokens per second)
llama_print_timings: prompt eval time =    9849.94 ms /    97 tokens (  101.55
ms per token,     9.85 tokens per second)
llama_print_timings:        eval time =    6023.99 ms /    28 runs   (  215.14
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   16062.38 ms /   125 tokens
No. of rows:  70%| | 876/1258 [12:54:39<1:41:43, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.20 ms /    31 runs   (    0.46
ms per token,  2183.71 tokens per second)
llama_print_timings: prompt eval time =   11102.33 ms /   101 tokens (  109.92
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    6345.19 ms /    30 runs   (  211.51
ms per token,     4.73 tokens per second)
llama_print_timings:       total time =   17650.36 ms /   131 tokens
No. of rows:  70%| | 877/1258 [12:54:56<1:44:39, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.64 ms /    27 runs   (    0.39
ms per token,  2536.64 tokens per second)
llama_print_timings: prompt eval time =    8302.10 ms /    83 tokens (  100.03
ms per token,    10.00 tokens per second)
llama_print_timings:        eval time =    5604.12 ms /    26 runs   (  215.54
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14077.54 ms /   109 tokens
No. of rows:  70%| | 878/1258 [12:55:10<1:39:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.63 ms /    26 runs   (    0.41
ms per token,  2445.22 tokens per second)
llama_print_timings: prompt eval time =   10011.60 ms /   102 tokens (   98.15
ms per token,    10.19 tokens per second)
```

```
llama_print_timings:        eval time =    5288.00 ms /    25 runs   (  211.52
ms per token,     4.73 tokens per second)
llama_print_timings:       total time =   15463.89 ms /   127 tokens
No. of rows:  70%| | 879/1258 [12:55:26<1:39:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.25 ms /    28 runs   (    0.40
ms per token,  2488.00 tokens per second)
llama_print_timings: prompt eval time =   10377.09 ms /   108 tokens (   96.08
ms per token,    10.41 tokens per second)
llama_print_timings:        eval time =    5741.78 ms /    27 runs   (  212.66
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   16296.86 ms /   135 tokens
No. of rows:  70%| | 880/1258 [12:55:42<1:39:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.38 ms /    27 runs   (    0.42
ms per token,  2371.54 tokens per second)
llama_print_timings: prompt eval time =   11612.83 ms /   106 tokens (  109.55
ms per token,     9.13 tokens per second)
llama_print_timings:        eval time =    5524.98 ms /    26 runs   (  212.50
ms per token,     4.71 tokens per second)
llama_print_timings:       total time =   17315.45 ms /   132 tokens
No. of rows:  70%| | 881/1258 [12:55:59<1:42:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.88 ms /    27 runs   (    0.40
ms per token,  2480.48 tokens per second)
llama_print_timings: prompt eval time =    8751.23 ms /    88 tokens (   99.45
ms per token,    10.06 tokens per second)
llama_print_timings:        eval time =    5620.67 ms /    26 runs   (  216.18
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   14546.19 ms /   114 tokens
No. of rows:  70%| | 882/1258 [12:56:14<1:38:52, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.92 ms /    24 runs   (    0.41
ms per token,  2418.38 tokens per second)
llama_print_timings: prompt eval time =    9044.16 ms /    93 tokens (   97.25
ms per token,    10.28 tokens per second)
llama_print_timings:        eval time =    6563.09 ms /    23 runs   (  285.35
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   15758.38 ms /   116 tokens
No. of rows:  70%| | 883/1258 [12:56:30<1:38:35, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.37 ms /    30 runs   (    0.41
ms per token,  2425.22 tokens per second)
llama_print_timings: prompt eval time =   10535.73 ms /   106 tokens (   99.39
ms per token,    10.06 tokens per second)
llama_print_timings:        eval time =    7850.61 ms /    29 runs   (  270.71
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   18576.49 ms /   135 tokens
No. of rows:  70%| | 884/1258 [12:56:48<1:43:34, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.17 ms /    27 runs   (    0.45
ms per token,  2219.30 tokens per second)
llama_print_timings: prompt eval time =    9869.09 ms /   103 tokens (   95.82
ms per token,    10.44 tokens per second)
llama_print_timings:        eval time =    7393.46 ms /    26 runs   (  284.36
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   17448.76 ms /   129 tokens
No. of rows:  70%| | 885/1258 [12:57:06<1:44:52, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      20.31 ms /    43 runs   (    0.47
ms per token,  2117.50 tokens per second)
llama_print_timings: prompt eval time =   17029.70 ms /   150 tokens (  113.53
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =   10265.56 ms /    42 runs   (  244.42
ms per token,     4.09 tokens per second)
llama_print_timings:       total time =   27613.37 ms /   192 tokens
No. of rows:  70%| | 886/1258 [12:57:33<2:04:35, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.88 ms /    42 runs   (    0.43
ms per token,  2349.12 tokens per second)
llama_print_timings: prompt eval time =   15231.12 ms /   128 tokens (  118.99
ms per token,     8.40 tokens per second)
llama_print_timings:        eval time =   10958.27 ms /    41 runs   (  267.27
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   26468.49 ms /   169 tokens
No. of rows:  71%| | 887/1258 [12:58:00<2:16:05, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.43 ms /    37 runs   (    0.39
```

```
ms per token,  2564.64 tokens per second)
llama_print_timings: prompt eval time =   11846.51 ms /   122 tokens (   97.10
ms per token,    10.30 tokens per second)
llama_print_timings:        eval time =    7605.31 ms /    36 runs   (  211.26
ms per token,     4.73 tokens per second)
llama_print_timings:       total time =   19681.29 ms /   158 tokens
No. of rows:  71%| | 888/1258 [12:58:20<2:11:25, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.24 ms /    19 runs   (    0.38
ms per token,  2622.86 tokens per second)
llama_print_timings: prompt eval time =    7898.98 ms /    80 tokens (   98.74
ms per token,    10.13 tokens per second)
llama_print_timings:        eval time =    3777.06 ms /    18 runs   (  209.84
ms per token,     4.77 tokens per second)
llama_print_timings:       total time =   11793.91 ms /    98 tokens
No. of rows:  71%| | 889/1258 [12:58:31<1:53:31, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.48 ms /    20 runs   (    0.37
ms per token,  2675.23 tokens per second)
llama_print_timings: prompt eval time =    7883.45 ms /    79 tokens (   99.79
ms per token,    10.02 tokens per second)
llama_print_timings:        eval time =    5691.16 ms /    19 runs   (  299.53
ms per token,     3.34 tokens per second)
llama_print_timings:       total time =   13694.42 ms /    98 tokens
No. of rows:  71%| | 890/1258 [12:58:45<1:44:28, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.39 ms /    20 runs   (    0.57
ms per token,  1755.93 tokens per second)
llama_print_timings: prompt eval time =    8537.29 ms /    84 tokens (  101.63
ms per token,     9.84 tokens per second)
llama_print_timings:        eval time =    6413.89 ms /    19 runs   (  337.57
ms per token,     2.96 tokens per second)
llama_print_timings:       total time =   15134.41 ms /   103 tokens
No. of rows:  71%| | 891/1258 [12:59:00<1:40:42, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.13 ms /    24 runs   (    0.42
ms per token,  2369.20 tokens per second)
llama_print_timings: prompt eval time =   11198.18 ms /    93 tokens (  120.41
ms per token,     8.30 tokens per second)
llama_print_timings:        eval time =    5344.48 ms /    23 runs   (  232.37
```

```
ms per token,     4.30 tokens per second)
llama_print_timings:       total time =   16705.97 ms /   116 tokens
No. of rows:  71%| | 892/1258 [12:59:17<1:40:53, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.52 ms /    28 runs   (    0.38
ms per token,  2660.84 tokens per second)
llama_print_timings: prompt eval time =   10943.12 ms /    96 tokens (  113.99
ms per token,     8.77 tokens per second)
llama_print_timings:        eval time =    5939.71 ms /    27 runs   (  219.99
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   17055.78 ms /   123 tokens
No. of rows:  71%| | 893/1258 [12:59:34<1:41:34, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.34 ms /    19 runs   (    0.54
ms per token,  1837.70 tokens per second)
llama_print_timings: prompt eval time =    9723.36 ms /    83 tokens (  117.15
ms per token,     8.54 tokens per second)
llama_print_timings:        eval time =    5326.27 ms /    18 runs   (  295.90
ms per token,     3.38 tokens per second)
llama_print_timings:       total time =   15215.01 ms /   101 tokens
No. of rows:  71%| | 894/1258 [12:59:49<1:38:37, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.74 ms /    26 runs   (    0.37
ms per token,  2668.31 tokens per second)
llama_print_timings: prompt eval time =   12195.11 ms /   110 tokens (  110.86
ms per token,     9.02 tokens per second)
llama_print_timings:        eval time =    5476.01 ms /    25 runs   (  219.04
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   17827.58 ms /   135 tokens
No. of rows:  71%| | 895/1258 [13:00:07<1:41:13, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.59 ms /    31 runs   (    0.41
ms per token,  2462.66 tokens per second)
llama_print_timings: prompt eval time =   14154.55 ms /   112 tokens (  126.38
ms per token,     7.91 tokens per second)
llama_print_timings:        eval time =    8556.44 ms /    30 runs   (  285.21
ms per token,     3.51 tokens per second)
llama_print_timings:       total time =   22911.79 ms /   142 tokens
No. of rows:  71%| | 896/1258 [13:00:30<1:52:08, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.14 ms /    25 runs   (    0.41
ms per token,  2465.24 tokens per second)
llama_print_timings: prompt eval time =    9108.79 ms /    83 tokens (  109.74
ms per token,     9.11 tokens per second)
llama_print_timings:        eval time =    7089.53 ms /    24 runs   (  295.40
ms per token,     3.39 tokens per second)
llama_print_timings:       total time =   16357.45 ms /   107 tokens
No. of rows:  71%| | 897/1258 [13:00:46<1:47:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.30 ms /    23 runs   (    0.40
ms per token,  2472.59 tokens per second)
llama_print_timings: prompt eval time =    8855.43 ms /    79 tokens (  112.09
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    4845.76 ms /    22 runs   (  220.26
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   13847.41 ms /   101 tokens
No. of rows:  71%| | 898/1258 [13:01:00<1:40:12, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.11 ms /    18 runs   (    0.40
ms per token,  2530.93 tokens per second)
llama_print_timings: prompt eval time =    8672.38 ms /    80 tokens (  108.40
ms per token,     9.22 tokens per second)
llama_print_timings:        eval time =    4045.48 ms /    17 runs   (  237.97
ms per token,     4.20 tokens per second)
llama_print_timings:       total time =   12832.23 ms /    97 tokens
No. of rows:  71%| | 899/1258 [13:01:13<1:33:00, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.34 ms /    29 runs   (    0.49
ms per token,  2022.88 tokens per second)
llama_print_timings: prompt eval time =   10107.07 ms /    89 tokens (  113.56
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    7442.97 ms /    28 runs   (  265.82
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   17774.00 ms /   117 tokens
No. of rows:  72%| | 900/1258 [13:01:31<1:36:44, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.68 ms /    20 runs   (    0.38
ms per token,  2603.15 tokens per second)
```

```
llama_print_timings: prompt eval time =    8257.54 ms /    76 tokens (  108.65
ms per token,     9.20 tokens per second)
llama_print_timings:        eval time =    4079.28 ms /    19 runs   (  214.70
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   12461.57 ms /    95 tokens
No. of rows:  72%| | 901/1258 [13:01:43<1:29:47, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.58 ms /    30 runs   (    0.42
ms per token,  2385.69 tokens per second)
llama_print_timings: prompt eval time =   11674.91 ms /   110 tokens (  106.14
ms per token,     9.42 tokens per second)
llama_print_timings:        eval time =    6811.76 ms /    29 runs   (  234.89
ms per token,     4.26 tokens per second)
llama_print_timings:       total time =   18686.58 ms /   139 tokens
No. of rows:  72%| | 902/1258 [13:02:02<1:35:57, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.99 ms /    26 runs   (    0.50
ms per token,  2002.00 tokens per second)
llama_print_timings: prompt eval time =   14235.67 ms /   107 tokens (  133.04
ms per token,     7.52 tokens per second)
llama_print_timings:        eval time =    7953.04 ms /    25 runs   (  318.12
ms per token,     3.14 tokens per second)
llama_print_timings:       total time =   22381.73 ms /   132 tokens
No. of rows:  72%| | 903/1258 [13:02:24<1:46:43, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.38 ms /    20 runs   (    0.42
ms per token,  2388.06 tokens per second)
llama_print_timings: prompt eval time =   10321.20 ms /    88 tokens (  117.29
ms per token,     8.53 tokens per second)
llama_print_timings:        eval time =    4534.81 ms /    19 runs   (  238.67
ms per token,     4.19 tokens per second)
llama_print_timings:       total time =   14993.91 ms /   107 tokens
No. of rows:  72%| | 904/1258 [13:02:39<1:41:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.72 ms /    22 runs   (    0.81
ms per token,  1241.82 tokens per second)
llama_print_timings: prompt eval time =   12585.67 ms /    94 tokens (  133.89
ms per token,     7.47 tokens per second)
llama_print_timings:        eval time =    9766.11 ms /    21 runs   (  465.05
ms per token,     2.15 tokens per second)
```

```
llama_print_timings:       total time =   22596.88 ms /   115 tokens
No. of rows:  72%| | 905/1258 [13:03:02<1:50:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.00 ms /    28 runs   (    0.39
ms per token,  2546.61 tokens per second)
llama_print_timings: prompt eval time =   11355.65 ms /   100 tokens (  113.56
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    6643.54 ms /    27 runs   (  246.06
ms per token,     4.06 tokens per second)
llama_print_timings:       total time =   18182.83 ms /   127 tokens
No. of rows:  72%| | 906/1258 [13:03:20<1:49:06, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.60 ms /    30 runs   (    0.39
ms per token,  2586.88 tokens per second)
llama_print_timings: prompt eval time =   13423.91 ms /   113 tokens (  118.80
ms per token,     8.42 tokens per second)
llama_print_timings:        eval time =    6363.59 ms /    29 runs   (  219.43
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   19974.55 ms /   142 tokens
No. of rows:  72%| | 907/1258 [13:03:40<1:51:13, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.26 ms /    26 runs   (    0.39
ms per token,  2534.11 tokens per second)
llama_print_timings: prompt eval time =   10999.31 ms /   101 tokens (  108.90
ms per token,     9.18 tokens per second)
llama_print_timings:        eval time =    7289.73 ms /    25 runs   (  291.59
ms per token,     3.43 tokens per second)
llama_print_timings:       total time =   18450.46 ms /   126 tokens
No. of rows:  72%| | 908/1258 [13:03:59<1:49:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.08 ms /    19 runs   (    0.43
ms per token,  2352.36 tokens per second)
llama_print_timings: prompt eval time =    9069.03 ms /    81 tokens (  111.96
ms per token,     8.93 tokens per second)
llama_print_timings:        eval time =    4630.33 ms /    18 runs   (  257.24
ms per token,     3.89 tokens per second)
llama_print_timings:       total time =   13823.57 ms /    99 tokens
No. of rows:  72%| | 909/1258 [13:04:12<1:40:52, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.97 ms /    26 runs   (    0.38
ms per token,  2608.09 tokens per second)
llama_print_timings: prompt eval time =   10000.71 ms /    91 tokens (  109.90
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    5503.21 ms /    25 runs   (  220.13
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   15661.87 ms /   116 tokens
No. of rows:  72%| | 910/1258 [13:04:28<1:37:40, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.53 ms /    23 runs   (    0.41
ms per token,  2414.70 tokens per second)
llama_print_timings: prompt eval time =    9957.30 ms /    92 tokens (  108.23
ms per token,     9.24 tokens per second)
llama_print_timings:        eval time =    4962.07 ms /    22 runs   (  225.55
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   15067.69 ms /   114 tokens
No. of rows:  72%| | 911/1258 [13:04:43<1:34:18, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.45 ms /    20 runs   (    0.37
ms per token,  2686.01 tokens per second)
llama_print_timings: prompt eval time =    9014.32 ms /    82 tokens (  109.93
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    4132.16 ms /    19 runs   (  217.48
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   13267.37 ms /   101 tokens
No. of rows:  72%| | 912/1258 [13:04:56<1:28:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.03 ms /    19 runs   (    0.37
ms per token,  2703.86 tokens per second)
llama_print_timings: prompt eval time =    9922.03 ms /    79 tokens (  125.60
ms per token,     7.96 tokens per second)
llama_print_timings:        eval time =    4225.21 ms /    18 runs   (  234.73
ms per token,     4.26 tokens per second)
llama_print_timings:       total time =   14263.67 ms /    97 tokens
No. of rows:  73%| | 913/1258 [13:05:11<1:26:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.00 ms /    37 runs   (    0.46
ms per token,  2176.47 tokens per second)
llama_print_timings: prompt eval time =   13022.40 ms /   119 tokens (  109.43
```

```
ms per token,     9.14 tokens per second)
llama_print_timings:        eval time =    9521.03 ms /    36 runs   (  264.47
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   22810.55 ms /   155 tokens
No. of rows:  73%| | 914/1258 [13:05:34<1:39:41, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.06 ms /    25 runs   (    0.40
ms per token,  2484.10 tokens per second)
llama_print_timings: prompt eval time =    9914.34 ms /    87 tokens (  113.96
ms per token,     8.78 tokens per second)
llama_print_timings:        eval time =    5410.54 ms /    24 runs   (  225.44
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   15487.34 ms /   111 tokens
No. of rows:  73%| | 915/1258 [13:05:49<1:36:09, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.90 ms /    21 runs   (    0.47
ms per token,  2121.86 tokens per second)
llama_print_timings: prompt eval time =    8589.47 ms /    78 tokens (  110.12
ms per token,     9.08 tokens per second)
llama_print_timings:        eval time =    5127.59 ms /    20 runs   (  256.38
ms per token,     3.90 tokens per second)
llama_print_timings:       total time =   13873.34 ms /    98 tokens
No. of rows:  73%| | 916/1258 [13:06:03<1:30:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.19 ms /    18 runs   (    0.40
ms per token,  2504.87 tokens per second)
llama_print_timings: prompt eval time =   10668.12 ms /    72 tokens (  148.17
ms per token,     6.75 tokens per second)
llama_print_timings:        eval time =    4603.89 ms /    17 runs   (  270.82
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   15390.16 ms /    89 tokens
No. of rows:  73%| | 917/1258 [13:06:18<1:29:40, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.38 ms /    36 runs   (    0.40
ms per token,  2502.61 tokens per second)
llama_print_timings: prompt eval time =   13238.10 ms /   118 tokens (  112.19
ms per token,     8.91 tokens per second)
llama_print_timings:        eval time =    7855.49 ms /    35 runs   (  224.44
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   21322.92 ms /   153 tokens
```

```
No. of rows:  73%| | 918/1258 [13:06:40<1:38:50, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.21 ms /    24 runs   (    0.43
ms per token,  2350.18 tokens per second)
llama_print_timings: prompt eval time =    9412.37 ms /    84 tokens (  112.05
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    5578.77 ms /    23 runs   (  242.56
ms per token,     4.12 tokens per second)
llama_print_timings:       total time =   15161.14 ms /   107 tokens
No. of rows:  73%| | 919/1258 [13:06:55<1:34:42, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.52 ms /    30 runs   (    0.52
ms per token,  1932.99 tokens per second)
llama_print_timings: prompt eval time =   16563.16 ms /   121 tokens (  136.89
ms per token,     7.31 tokens per second)
llama_print_timings:        eval time =    9820.36 ms /    29 runs   (  338.63
ms per token,     2.95 tokens per second)
llama_print_timings:       total time =   26619.14 ms /   150 tokens
No. of rows:  73%| | 920/1258 [13:07:21<1:51:06, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      27.43 ms /    42 runs   (    0.65
ms per token,  1531.17 tokens per second)
llama_print_timings: prompt eval time =   14382.93 ms /   116 tokens (  123.99
ms per token,     8.07 tokens per second)
llama_print_timings:        eval time =   12875.02 ms /    41 runs   (  314.02
ms per token,     3.18 tokens per second)
llama_print_timings:       total time =   27645.31 ms /   157 tokens
No. of rows:  73%| | 921/1258 [13:07:49<2:04:08, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.62 ms /    19 runs   (    0.51
ms per token,  1976.08 tokens per second)
llama_print_timings: prompt eval time =   10404.94 ms /    79 tokens (  131.71
ms per token,     7.59 tokens per second)
llama_print_timings:        eval time =    5058.79 ms /    18 runs   (  281.04
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   15623.19 ms /    97 tokens
No. of rows:  73%| | 922/1258 [13:08:05<1:52:54, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       9.19 ms /    24 runs   (    0.38
ms per token,  2612.10 tokens per second)
llama_print_timings: prompt eval time =   10939.44 ms /    93 tokens (  117.63
ms per token,     8.50 tokens per second)
llama_print_timings:        eval time =    5167.16 ms /    23 runs   (  224.66
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   16259.33 ms /   116 tokens
No. of rows:  73%| | 923/1258 [13:08:21<1:46:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.49 ms /    23 runs   (    0.41
ms per token,  2424.11 tokens per second)
llama_print_timings: prompt eval time =    9329.80 ms /    81 tokens (  115.18
ms per token,     8.68 tokens per second)
llama_print_timings:        eval time =    5133.86 ms /    22 runs   (  233.36
ms per token,     4.29 tokens per second)
llama_print_timings:       total time =   14620.72 ms /   103 tokens
No. of rows:  73%| | 924/1258 [13:08:36<1:38:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.18 ms /    21 runs   (    0.39
ms per token,  2568.49 tokens per second)
llama_print_timings: prompt eval time =   10235.38 ms /    91 tokens (  112.48
ms per token,     8.89 tokens per second)
llama_print_timings:        eval time =    6159.24 ms /    20 runs   (  307.96
ms per token,     3.25 tokens per second)
llama_print_timings:       total time =   16525.80 ms /   111 tokens
No. of rows:  74%| | 925/1258 [13:08:52<1:36:13, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.30 ms /    25 runs   (    0.45
ms per token,  2212.78 tokens per second)
llama_print_timings: prompt eval time =   10357.01 ms /    81 tokens (  127.86
ms per token,     7.82 tokens per second)
llama_print_timings:        eval time =    5671.11 ms /    24 runs   (  236.30
ms per token,     4.23 tokens per second)
llama_print_timings:       total time =   16205.63 ms /   105 tokens
No. of rows:  74%| | 926/1258 [13:09:08<1:34:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.90 ms /    23 runs   (    0.47
ms per token,  2110.48 tokens per second)
llama_print_timings: prompt eval time =   12969.65 ms /   101 tokens (  128.41
ms per token,     7.79 tokens per second)
```

```
llama_print_timings:        eval time =    5517.68 ms /    22 runs   (  250.80
ms per token,     3.99 tokens per second)
llama_print_timings:       total time =   18658.19 ms /   123 tokens
No. of rows:  74%| | 927/1258 [13:09:27<1:36:33, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      22.07 ms /    39 runs   (    0.57
ms per token,  1766.78 tokens per second)
llama_print_timings: prompt eval time =   16400.56 ms /   125 tokens (  131.20
ms per token,     7.62 tokens per second)
llama_print_timings:        eval time =    9489.34 ms /    38 runs   (  249.72
ms per token,     4.00 tokens per second)
llama_print_timings:       total time =   26178.85 ms /   163 tokens
No. of rows:  74%| | 928/1258 [13:09:53<1:50:35, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.82 ms /    25 runs   (    0.43
ms per token,  2310.75 tokens per second)
llama_print_timings: prompt eval time =   11138.93 ms /    92 tokens (  121.08
ms per token,     8.26 tokens per second)
llama_print_timings:        eval time =    5714.69 ms /    24 runs   (  238.11
ms per token,     4.20 tokens per second)
llama_print_timings:       total time =   17020.51 ms /   116 tokens
No. of rows:  74%| | 929/1258 [13:10:10<1:45:12, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.95 ms /    30 runs   (    0.43
ms per token,  2317.14 tokens per second)
llama_print_timings: prompt eval time =   13167.03 ms /   106 tokens (  124.22
ms per token,     8.05 tokens per second)
llama_print_timings:        eval time =    8378.64 ms /    29 runs   (  288.92
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   21746.13 ms /   135 tokens
No. of rows:  74%| | 930/1258 [13:10:32<1:49:06, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.02 ms /    33 runs   (    0.42
ms per token,  2353.28 tokens per second)
llama_print_timings: prompt eval time =   14806.86 ms /   128 tokens (  115.68
ms per token,     8.64 tokens per second)
llama_print_timings:        eval time =    7834.24 ms /    32 runs   (  244.82
ms per token,     4.08 tokens per second)
llama_print_timings:       total time =   22862.81 ms /   160 tokens
No. of rows:  74%| | 931/1258 [13:10:55<1:53:30, 2Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.08 ms /    34 runs   (    0.38
ms per token,  2598.99 tokens per second)
llama_print_timings: prompt eval time =   12685.88 ms /    99 tokens (  128.14
ms per token,     7.80 tokens per second)
llama_print_timings:        eval time =    8755.45 ms /    33 runs   (  265.32
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   21651.56 ms /   132 tokens
No. of rows:  74%| | 932/1258 [13:11:17<1:54:31, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.18 ms /    33 runs   (    0.40
ms per token,  2504.55 tokens per second)
llama_print_timings: prompt eval time =   11363.94 ms /   106 tokens (  107.21
ms per token,     9.33 tokens per second)
llama_print_timings:        eval time =    7006.87 ms /    32 runs   (  218.96
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   18578.92 ms /   138 tokens
No. of rows:  74%| | 933/1258 [13:11:35<1:50:07, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.30 ms /    18 runs   (    0.41
ms per token,  2466.43 tokens per second)
llama_print_timings: prompt eval time =    8915.78 ms /    83 tokens (  107.42
ms per token,     9.31 tokens per second)
llama_print_timings:        eval time =    3648.44 ms /    17 runs   (  214.61
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   12676.99 ms /   100 tokens
No. of rows:  74%| | 934/1258 [13:11:48<1:37:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.33 ms /    16 runs   (    0.40
ms per token,  2526.85 tokens per second)
llama_print_timings: prompt eval time =    8371.61 ms /    79 tokens (  105.97
ms per token,     9.44 tokens per second)
llama_print_timings:        eval time =    3226.79 ms /    15 runs   (  215.12
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   11697.42 ms /    94 tokens
No. of rows:  74%| | 935/1258 [13:12:00<1:26:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.85 ms /    28 runs   (    0.46
```

```
ms per token,  2179.16 tokens per second)
llama_print_timings: prompt eval time =   10833.96 ms /    99 tokens (  109.43
ms per token,     9.14 tokens per second)
llama_print_timings:        eval time =    6502.78 ms /    27 runs   (  240.84
ms per token,     4.15 tokens per second)
llama_print_timings:       total time =   17536.53 ms /   126 tokens
No. of rows:  74%| | 936/1258 [13:12:17<1:28:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.05 ms /    31 runs   (    0.52
ms per token,  1930.86 tokens per second)
llama_print_timings: prompt eval time =   12175.93 ms /    99 tokens (  122.99
ms per token,     8.13 tokens per second)
llama_print_timings:        eval time =    9760.06 ms /    30 runs   (  325.34
ms per token,     3.07 tokens per second)
llama_print_timings:       total time =   22187.34 ms /   129 tokens
No. of rows:  74%| | 937/1258 [13:12:39<1:37:39, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.66 ms /    29 runs   (    0.44
ms per token,  2290.50 tokens per second)
llama_print_timings: prompt eval time =   12073.53 ms /    92 tokens (  131.23
ms per token,     7.62 tokens per second)
llama_print_timings:        eval time =    6333.01 ms /    28 runs   (  226.18
ms per token,     4.42 tokens per second)
llama_print_timings:       total time =   18600.36 ms /   120 tokens
No. of rows:  75%| | 938/1258 [13:12:58<1:37:55, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.25 ms /    32 runs   (    0.51
ms per token,  1968.99 tokens per second)
llama_print_timings: prompt eval time =   14387.79 ms /   102 tokens (  141.06
ms per token,     7.09 tokens per second)
llama_print_timings:        eval time =    8376.57 ms /    31 runs   (  270.21
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   23008.72 ms /   133 tokens
No. of rows:  75%| | 939/1258 [13:13:21<1:45:02, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.10 ms /    24 runs   (    0.55
ms per token,  1832.34 tokens per second)
llama_print_timings: prompt eval time =   14305.66 ms /   101 tokens (  141.64
ms per token,     7.06 tokens per second)
llama_print_timings:        eval time =    6246.55 ms /    23 runs   (  271.59
```

```
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   20765.21 ms /   124 tokens
No. of rows:  75%| | 940/1258 [13:13:42<1:46:19, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.78 ms /    34 runs   (    0.43
ms per token,  2300.09 tokens per second)
llama_print_timings: prompt eval time =   10426.43 ms /   100 tokens (  104.26
ms per token,     9.59 tokens per second)
llama_print_timings:        eval time =    7055.47 ms /    33 runs   (  213.80
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   17700.52 ms /   133 tokens
No. of rows:  75%| | 941/1258 [13:13:59<1:42:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.68 ms /    21 runs   (    0.46
ms per token,  2169.87 tokens per second)
llama_print_timings: prompt eval time =    8602.12 ms /    84 tokens (  102.41
ms per token,     9.77 tokens per second)
llama_print_timings:        eval time =    4446.21 ms /    20 runs   (  222.31
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   13189.33 ms /   104 tokens
No. of rows:  75%| | 942/1258 [13:14:13<1:32:12, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.56 ms /    22 runs   (    0.39
ms per token,  2569.49 tokens per second)
llama_print_timings: prompt eval time =    9026.15 ms /    91 tokens (   99.19
ms per token,    10.08 tokens per second)
llama_print_timings:        eval time =    4493.42 ms /    21 runs   (  213.97
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   13658.85 ms /   112 tokens
No. of rows:  75%| | 943/1258 [13:14:26<1:25:52, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.56 ms /    21 runs   (    0.41
ms per token,  2452.41 tokens per second)
llama_print_timings: prompt eval time =    9148.96 ms /    76 tokens (  120.38
ms per token,     8.31 tokens per second)
llama_print_timings:        eval time =    4280.35 ms /    20 runs   (  214.02
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   13564.94 ms /    96 tokens
No. of rows:  75%| | 944/1258 [13:14:40<1:21:13, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.59 ms /    26 runs   (    0.45
ms per token,  2243.51 tokens per second)
llama_print_timings: prompt eval time =    9167.74 ms /    89 tokens (  103.01
ms per token,     9.71 tokens per second)
llama_print_timings:        eval time =    5328.61 ms /    25 runs   (  213.14
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   14666.96 ms /   114 tokens
No. of rows:  75%| | 945/1258 [13:14:54<1:19:38, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.56 ms /    26 runs   (    0.41
ms per token,  2462.35 tokens per second)
llama_print_timings: prompt eval time =   11090.12 ms /    98 tokens (  113.16
ms per token,     8.84 tokens per second)
llama_print_timings:        eval time =    5338.96 ms /    25 runs   (  213.56
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   16596.52 ms /   123 tokens
No. of rows:  75%| | 946/1258 [13:15:11<1:21:28, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.33 ms /    23 runs   (    0.41
ms per token,  2465.17 tokens per second)
llama_print_timings: prompt eval time =   10387.64 ms /    90 tokens (  115.42
ms per token,     8.66 tokens per second)
llama_print_timings:        eval time =    5281.90 ms /    22 runs   (  240.09
ms per token,     4.17 tokens per second)
llama_print_timings:       total time =   15819.04 ms /   112 tokens
No. of rows:  75%| | 947/1258 [13:15:27<1:21:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.77 ms /    31 runs   (    0.41
ms per token,  2426.80 tokens per second)
llama_print_timings: prompt eval time =    9648.50 ms /   100 tokens (   96.48
ms per token,    10.36 tokens per second)
llama_print_timings:        eval time =    6488.93 ms /    30 runs   (  216.30
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   16337.71 ms /   130 tokens
No. of rows:  75%| | 948/1258 [13:15:43<1:22:10, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.74 ms /    24 runs   (    0.41
ms per token,  2464.82 tokens per second)
```

```
llama_print_timings: prompt eval time =    9052.63 ms /    92 tokens (   98.40
ms per token,    10.16 tokens per second)
llama_print_timings:        eval time =    4869.81 ms /    23 runs   (  211.73
ms per token,     4.72 tokens per second)
llama_print_timings:       total time =   14074.53 ms /   115 tokens
No. of rows:  75%| | 949/1258 [13:15:57<1:19:05, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.39 ms /    27 runs   (    0.38
ms per token,  2597.90 tokens per second)
llama_print_timings: prompt eval time =    9195.92 ms /    92 tokens (   99.96
ms per token,    10.00 tokens per second)
llama_print_timings:        eval time =    7292.15 ms /    26 runs   (  280.47
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   16661.14 ms /   118 tokens
No. of rows:  76%| | 950/1258 [13:16:14<1:20:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.65 ms /    21 runs   (    0.41
ms per token,  2428.31 tokens per second)
llama_print_timings: prompt eval time =    8000.90 ms /    81 tokens (   98.78
ms per token,    10.12 tokens per second)
llama_print_timings:        eval time =    4255.29 ms /    20 runs   (  212.76
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   12389.84 ms /   101 tokens
No. of rows:  76%| | 951/1258 [13:16:26<1:15:26, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.98 ms /    30 runs   (    0.40
ms per token,  2504.80 tokens per second)
llama_print_timings: prompt eval time =    8654.30 ms /    87 tokens (   99.47
ms per token,    10.05 tokens per second)
llama_print_timings:        eval time =    7969.82 ms /    29 runs   (  274.82
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   16811.05 ms /   116 tokens
No. of rows:  76%| | 952/1258 [13:16:43<1:18:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.12 ms /    23 runs   (    0.40
ms per token,  2521.38 tokens per second)
llama_print_timings: prompt eval time =    9813.55 ms /    99 tokens (   99.13
ms per token,    10.09 tokens per second)
llama_print_timings:        eval time =    4684.66 ms /    22 runs   (  212.94
ms per token,     4.70 tokens per second)
```

```
llama_print_timings:       total time =   14642.36 ms /   121 tokens
No. of rows:  76%| | 953/1258 [13:16:58<1:17:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.76 ms /    32 runs   (    0.43
ms per token,  2325.41 tokens per second)
llama_print_timings: prompt eval time =   13002.92 ms /   120 tokens (  108.36
ms per token,     9.23 tokens per second)
llama_print_timings:        eval time =    6602.04 ms /    31 runs   (  212.97
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   19812.58 ms /   151 tokens
No. of rows:  76%| | 954/1258 [13:17:18<1:23:51, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.06 ms /    31 runs   (    0.39
ms per token,  2569.84 tokens per second)
llama_print_timings: prompt eval time =   11780.08 ms /   118 tokens (   99.83
ms per token,    10.02 tokens per second)
llama_print_timings:        eval time =    6429.78 ms /    30 runs   (  214.33
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   18411.33 ms /   148 tokens
No. of rows:  76%| | 955/1258 [13:17:36<1:26:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.30 ms /    21 runs   (    0.40
ms per token,  2529.21 tokens per second)
llama_print_timings: prompt eval time =    8543.25 ms /    79 tokens (  108.14
ms per token,     9.25 tokens per second)
llama_print_timings:        eval time =    4384.65 ms /    20 runs   (  219.23
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   13060.89 ms /    99 tokens
No. of rows:  76%| | 956/1258 [13:17:49<1:20:01, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.55 ms /    20 runs   (    0.43
ms per token,  2339.45 tokens per second)
llama_print_timings: prompt eval time =    9852.97 ms /    92 tokens (  107.10
ms per token,     9.34 tokens per second)
llama_print_timings:        eval time =    4064.15 ms /    19 runs   (  213.90
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   14047.83 ms /   111 tokens
No. of rows:  76%| | 957/1258 [13:18:03<1:16:59, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.15 ms /    32 runs   (    0.41
ms per token,  2434.20 tokens per second)
llama_print_timings: prompt eval time =   13696.79 ms /   122 tokens (  112.27
ms per token,     8.91 tokens per second)
llama_print_timings:        eval time =    6649.50 ms /    31 runs   (  214.50
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   20550.00 ms /   153 tokens
No. of rows:  76%| | 958/1258 [13:18:24<1:24:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.89 ms /    23 runs   (    0.43
ms per token,  2325.82 tokens per second)
llama_print_timings: prompt eval time =    8482.39 ms /    81 tokens (  104.72
ms per token,     9.55 tokens per second)
llama_print_timings:        eval time =    4684.67 ms /    22 runs   (  212.94
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   13315.00 ms /   103 tokens
No. of rows:  76%| | 959/1258 [13:18:37<1:18:54, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.56 ms /    22 runs   (    0.39
ms per token,  2568.89 tokens per second)
llama_print_timings: prompt eval time =    8876.38 ms /    90 tokens (   98.63
ms per token,    10.14 tokens per second)
llama_print_timings:        eval time =    4450.12 ms /    21 runs   (  211.91
ms per token,     4.72 tokens per second)
llama_print_timings:       total time =   13463.46 ms /   111 tokens
No. of rows:  76%| | 960/1258 [13:18:51<1:15:07, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.55 ms /    21 runs   (    0.41
ms per token,  2456.72 tokens per second)
llama_print_timings: prompt eval time =   10185.78 ms /    88 tokens (  115.75
ms per token,     8.64 tokens per second)
llama_print_timings:        eval time =    4335.78 ms /    20 runs   (  216.79
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   14652.88 ms /   108 tokens
No. of rows:  76%| | 961/1258 [13:19:05<1:14:10, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.68 ms /    45 runs   (    0.42
ms per token,  2408.74 tokens per second)
llama_print_timings: prompt eval time =   12150.54 ms /   123 tokens (   98.78
```

```
ms per token,    10.12 tokens per second)
llama_print_timings:        eval time =    9424.09 ms /    44 runs   (  214.18
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   21867.10 ms /   167 tokens
No. of rows:  76%| | 962/1258 [13:19:27<1:24:07, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.81 ms /    23 runs   (    0.43
ms per token,  2345.02 tokens per second)
llama_print_timings: prompt eval time =    9883.05 ms /    84 tokens (  117.66
ms per token,     8.50 tokens per second)
llama_print_timings:        eval time =    4777.02 ms /    22 runs   (  217.14
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   14807.62 ms /   106 tokens
No. of rows:  77%| | 963/1258 [13:19:42<1:20:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.53 ms /    28 runs   (    0.41
ms per token,  2427.82 tokens per second)
llama_print_timings: prompt eval time =   10537.14 ms /   109 tokens (   96.67
ms per token,    10.34 tokens per second)
llama_print_timings:        eval time =    5716.18 ms /    27 runs   (  211.71
ms per token,     4.72 tokens per second)
llama_print_timings:       total time =   16432.65 ms /   136 tokens
No. of rows:  77%| | 964/1258 [13:19:58<1:20:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.11 ms /    23 runs   (    0.40
ms per token,  2524.14 tokens per second)
llama_print_timings: prompt eval time =    9353.85 ms /    82 tokens (  114.07
ms per token,     8.77 tokens per second)
llama_print_timings:        eval time =    5040.58 ms /    22 runs   (  229.12
ms per token,     4.36 tokens per second)
llama_print_timings:       total time =   14542.28 ms /   104 tokens
No. of rows:  77%| | 965/1258 [13:20:13<1:17:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.38 ms /    26 runs   (    0.40
ms per token,  2504.33 tokens per second)
llama_print_timings: prompt eval time =   10825.30 ms /   106 tokens (  102.13
ms per token,     9.79 tokens per second)
llama_print_timings:        eval time =    5340.83 ms /    25 runs   (  213.63
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   16329.98 ms /   131 tokens
```

```
No. of rows:  77%| | 966/1258 [13:20:29<1:17:49, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.58 ms /    23 runs   (    0.42
ms per token,  2400.08 tokens per second)
llama_print_timings: prompt eval time =    9136.80 ms /    93 tokens (   98.25
ms per token,    10.18 tokens per second)
llama_print_timings:        eval time =    4667.44 ms /    22 runs   (  212.16
ms per token,     4.71 tokens per second)
llama_print_timings:       total time =   13953.18 ms /   115 tokens
No. of rows:  77%| | 967/1258 [13:20:43<1:14:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.66 ms /    36 runs   (    0.41
ms per token,  2456.50 tokens per second)
llama_print_timings: prompt eval time =   12537.37 ms /   115 tokens (  109.02
ms per token,     9.17 tokens per second)
llama_print_timings:        eval time =    7478.28 ms /    35 runs   (  213.67
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   20248.13 ms /   150 tokens
No. of rows:  77%| | 968/1258 [13:21:03<1:21:25, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.64 ms /    28 runs   (    0.42
ms per token,  2405.50 tokens per second)
llama_print_timings: prompt eval time =   10806.17 ms /    98 tokens (  110.27
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    5734.09 ms /    27 runs   (  212.37
ms per token,     4.71 tokens per second)
llama_print_timings:       total time =   16718.49 ms /   125 tokens
No. of rows:  77%| | 969/1258 [13:21:20<1:20:57, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.01 ms /    20 runs   (    0.40
ms per token,  2497.81 tokens per second)
llama_print_timings: prompt eval time =    8126.33 ms /    83 tokens (   97.91
ms per token,    10.21 tokens per second)
llama_print_timings:        eval time =    4147.69 ms /    19 runs   (  218.30
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   12402.33 ms /   102 tokens
No. of rows:  77%| | 970/1258 [13:21:33<1:14:20, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =      14.72 ms /    35 runs   (    0.42
ms per token,  2377.72 tokens per second)
llama_print_timings: prompt eval time =   10043.45 ms /   102 tokens (   98.47
ms per token,    10.16 tokens per second)
llama_print_timings:        eval time =    7240.78 ms /    34 runs   (  212.96
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   17509.44 ms /   136 tokens
No. of rows:  77%| | 971/1258 [13:21:50<1:17:00, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.04 ms /    27 runs   (    0.48
ms per token,  2071.35 tokens per second)
llama_print_timings: prompt eval time =    8464.26 ms /    83 tokens (  101.98
ms per token,     9.81 tokens per second)
llama_print_timings:        eval time =    6395.66 ms /    26 runs   (  245.99
ms per token,     4.07 tokens per second)
llama_print_timings:       total time =   15035.58 ms /   109 tokens
No. of rows:  77%| | 972/1258 [13:22:05<1:15:13, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.88 ms /    29 runs   (    0.41
ms per token,  2441.28 tokens per second)
llama_print_timings: prompt eval time =    9952.25 ms /    97 tokens (  102.60
ms per token,     9.75 tokens per second)
llama_print_timings:        eval time =    6098.23 ms /    28 runs   (  217.79
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   16242.04 ms /   125 tokens
No. of rows:  77%| | 973/1258 [13:22:21<1:15:38, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.59 ms /    19 runs   (    0.45
ms per token,  2211.62 tokens per second)
llama_print_timings: prompt eval time =    8436.26 ms /    83 tokens (  101.64
ms per token,     9.84 tokens per second)
llama_print_timings:        eval time =    3859.29 ms /    18 runs   (  214.41
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   12418.78 ms /   101 tokens
No. of rows:  77%| | 974/1258 [13:22:34<1:10:24, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.22 ms /    26 runs   (    0.43
ms per token,  2316.88 tokens per second)
llama_print_timings: prompt eval time =   12372.83 ms /   117 tokens (  105.75
ms per token,     9.46 tokens per second)
```

```
llama_print_timings:        eval time =    5361.31 ms /    25 runs   (  214.45
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   17901.48 ms /   142 tokens
No. of rows:  78%| | 975/1258 [13:22:52<1:14:27, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.39 ms /    27 runs   (    0.42
ms per token,  2370.71 tokens per second)
llama_print_timings: prompt eval time =   10784.56 ms /    95 tokens (  113.52
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    5475.48 ms /    26 runs   (  210.60
ms per token,     4.75 tokens per second)
llama_print_timings:       total time =   16432.22 ms /   121 tokens
No. of rows:  78%| | 976/1258 [13:23:08<1:15:06, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.48 ms /    22 runs   (    0.39
ms per token,  2594.65 tokens per second)
llama_print_timings: prompt eval time =    9905.08 ms /    86 tokens (  115.18
ms per token,     8.68 tokens per second)
llama_print_timings:        eval time =    4439.10 ms /    21 runs   (  211.39
ms per token,     4.73 tokens per second)
llama_print_timings:       total time =   14480.83 ms /   107 tokens
No. of rows:  78%| | 977/1258 [13:23:23<1:12:45, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.15 ms /    16 runs   (    0.45
ms per token,  2238.08 tokens per second)
llama_print_timings: prompt eval time =    9701.35 ms /    99 tokens (   97.99
ms per token,    10.20 tokens per second)
llama_print_timings:        eval time =    3414.76 ms /    15 runs   (  227.65
ms per token,     4.39 tokens per second)
llama_print_timings:       total time =   13222.61 ms /   114 tokens
No. of rows:  78%| | 978/1258 [13:23:36<1:09:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.96 ms /    26 runs   (    0.38
ms per token,  2609.66 tokens per second)
llama_print_timings: prompt eval time =   10433.54 ms /    95 tokens (  109.83
ms per token,     9.11 tokens per second)
llama_print_timings:        eval time =    5306.28 ms /    25 runs   (  212.25
ms per token,     4.71 tokens per second)
llama_print_timings:       total time =   15902.82 ms /   120 tokens
No. of rows:  78%| | 979/1258 [13:23:52<1:10:30, 1Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.88 ms /    30 runs   (    0.40
ms per token,  2525.89 tokens per second)
llama_print_timings: prompt eval time =    9540.43 ms /    96 tokens (   99.38
ms per token,    10.06 tokens per second)
llama_print_timings:        eval time =    7822.61 ms /    29 runs   (  269.75
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   17552.91 ms /   125 tokens
No. of rows:  78%| | 980/1258 [13:24:09<1:13:35, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.92 ms /    20 runs   (    0.50
ms per token,  2016.94 tokens per second)
llama_print_timings: prompt eval time =    8997.51 ms /    85 tokens (  105.85
ms per token,     9.45 tokens per second)
llama_print_timings:        eval time =    5506.56 ms /    19 runs   (  289.82
ms per token,     3.45 tokens per second)
llama_print_timings:       total time =   14664.46 ms /   104 tokens
No. of rows:  78%| | 981/1258 [13:24:24<1:11:39, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.46 ms /    28 runs   (    0.45
ms per token,  2246.47 tokens per second)
llama_print_timings: prompt eval time =   12817.21 ms /   101 tokens (  126.90
ms per token,     7.88 tokens per second)
llama_print_timings:        eval time =    6510.08 ms /    27 runs   (  241.11
ms per token,     4.15 tokens per second)
llama_print_timings:       total time =   19520.38 ms /   128 tokens
No. of rows:  78%| | 982/1258 [13:24:44<1:16:56, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.58 ms /    18 runs   (    0.59
ms per token,  1700.84 tokens per second)
llama_print_timings: prompt eval time =    9566.17 ms /    75 tokens (  127.55
ms per token,     7.84 tokens per second)
llama_print_timings:        eval time =    4468.07 ms /    17 runs   (  262.83
ms per token,     3.80 tokens per second)
llama_print_timings:       total time =   14199.78 ms /    92 tokens
No. of rows:  78%| | 983/1258 [13:24:58<1:13:11, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.18 ms /    31 runs   (    0.43
```

```
ms per token,  2352.94 tokens per second)
llama_print_timings: prompt eval time =   10052.14 ms /    91 tokens (  110.46
ms per token,     9.05 tokens per second)
llama_print_timings:        eval time =    6761.87 ms /    30 runs   (  225.40
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   17018.72 ms /   121 tokens
No. of rows:  78%| | 984/1258 [13:25:15<1:14:22, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.56 ms /    23 runs   (    0.42
ms per token,  2406.36 tokens per second)
llama_print_timings: prompt eval time =   16804.43 ms /   110 tokens (  152.77
ms per token,     6.55 tokens per second)
llama_print_timings:        eval time =    5288.45 ms /    22 runs   (  240.38
ms per token,     4.16 tokens per second)
llama_print_timings:       total time =   22249.57 ms /   132 tokens
No. of rows:  78%| | 985/1258 [13:25:37<1:22:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.21 ms /    19 runs   (    0.43
ms per token,  2312.84 tokens per second)
llama_print_timings: prompt eval time =   10783.99 ms /    81 tokens (  133.14
ms per token,     7.51 tokens per second)
llama_print_timings:        eval time =    4371.25 ms /    18 runs   (  242.85
ms per token,     4.12 tokens per second)
llama_print_timings:       total time =   15289.48 ms /    99 tokens
No. of rows:  78%| | 986/1258 [13:25:52<1:18:11, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.37 ms /    24 runs   (    0.60
ms per token,  1669.80 tokens per second)
llama_print_timings: prompt eval time =   12009.94 ms /    88 tokens (  136.48
ms per token,     7.33 tokens per second)
llama_print_timings:        eval time =    8241.40 ms /    23 runs   (  358.32
ms per token,     2.79 tokens per second)
llama_print_timings:       total time =   20450.84 ms /   111 tokens
No. of rows:  78%| | 987/1258 [13:26:13<1:22:15, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.41 ms /    33 runs   (    0.44
ms per token,  2290.08 tokens per second)
llama_print_timings: prompt eval time =   14198.36 ms /   121 tokens (  117.34
ms per token,     8.52 tokens per second)
llama_print_timings:        eval time =    7569.49 ms /    32 runs   (  236.55
```

```
ms per token,     4.23 tokens per second)
llama_print_timings:       total time =   21998.78 ms /   153 tokens
No. of rows:  79%| | 988/1258 [13:26:35<1:27:04, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.04 ms /    20 runs   (    0.45
ms per token,  2211.41 tokens per second)
llama_print_timings: prompt eval time =   12165.03 ms /    78 tokens (  155.96
ms per token,     6.41 tokens per second)
llama_print_timings:        eval time =    4989.93 ms /    19 runs   (  262.63
ms per token,     3.81 tokens per second)
llama_print_timings:       total time =   17288.08 ms /    97 tokens
No. of rows:  79%| | 989/1258 [13:26:52<1:23:59, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.86 ms /    25 runs   (    0.47
ms per token,  2108.81 tokens per second)
llama_print_timings: prompt eval time =   10435.07 ms /    93 tokens (  112.21
ms per token,     8.91 tokens per second)
llama_print_timings:        eval time =    7337.32 ms /    24 runs   (  305.72
ms per token,     3.27 tokens per second)
llama_print_timings:       total time =   17942.73 ms /   117 tokens
No. of rows:  79%| | 990/1258 [13:27:10<1:22:37, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.42 ms /    26 runs   (    0.40
ms per token,  2495.20 tokens per second)
llama_print_timings: prompt eval time =   10001.38 ms /   101 tokens (   99.02
ms per token,    10.10 tokens per second)
llama_print_timings:        eval time =    5439.86 ms /    25 runs   (  217.59
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   15614.13 ms /   126 tokens
No. of rows:  79%| | 991/1258 [13:27:26<1:18:28, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.96 ms /    19 runs   (    0.42
ms per token,  2385.44 tokens per second)
llama_print_timings: prompt eval time =    8327.81 ms /    81 tokens (  102.81
ms per token,     9.73 tokens per second)
llama_print_timings:        eval time =    3924.18 ms /    18 runs   (  218.01
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   12375.69 ms /    99 tokens
No. of rows:  79%| | 992/1258 [13:27:38<1:11:11, 1Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.55 ms /    31 runs   (    0.44
ms per token,  2287.99 tokens per second)
llama_print_timings: prompt eval time =    9562.00 ms /    92 tokens (  103.93
ms per token,     9.62 tokens per second)
llama_print_timings:        eval time =    6548.35 ms /    30 runs   (  218.28
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   16307.63 ms /   122 tokens
No. of rows:  79%| | 993/1258 [13:27:54<1:11:16, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.66 ms /    27 runs   (    0.43
ms per token,  2315.61 tokens per second)
llama_print_timings: prompt eval time =   11236.70 ms /   105 tokens (  107.02
ms per token,     9.34 tokens per second)
llama_print_timings:        eval time =    5736.61 ms /    26 runs   (  220.64
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   17153.20 ms /   131 tokens
No. of rows:  79%| | 994/1258 [13:28:12<1:12:21, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.41 ms /    27 runs   (    0.42
ms per token,  2366.35 tokens per second)
llama_print_timings: prompt eval time =    8970.56 ms /    86 tokens (  104.31
ms per token,     9.59 tokens per second)
llama_print_timings:        eval time =    6879.90 ms /    26 runs   (  264.61
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   16023.76 ms /   112 tokens
No. of rows:  79%| | 995/1258 [13:28:28<1:11:32, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.92 ms /    26 runs   (    0.42
ms per token,  2381.82 tokens per second)
llama_print_timings: prompt eval time =    9612.23 ms /    97 tokens (   99.10
ms per token,    10.09 tokens per second)
llama_print_timings:        eval time =    5312.75 ms /    25 runs   (  212.51
ms per token,     4.71 tokens per second)
llama_print_timings:       total time =   15094.02 ms /   122 tokens
No. of rows:  79%| | 996/1258 [13:28:43<1:09:40, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.11 ms /    39 runs   (    0.44
ms per token,  2279.10 tokens per second)
```

```
llama_print_timings: prompt eval time =   14501.85 ms /   135 tokens (  107.42
ms per token,     9.31 tokens per second)
llama_print_timings:        eval time =   10136.12 ms /    38 runs   (  266.74
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   24901.30 ms /   173 tokens
No. of rows:  79%| | 997/1258 [13:29:08<1:21:05, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.99 ms /    32 runs   (    0.44
ms per token,  2287.35 tokens per second)
llama_print_timings: prompt eval time =   12174.58 ms /   124 tokens (   98.18
ms per token,    10.19 tokens per second)
llama_print_timings:        eval time =    6650.16 ms /    31 runs   (  214.52
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   19034.04 ms /   155 tokens
No. of rows:  79%| | 998/1258 [13:29:27<1:21:17, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.84 ms /    26 runs   (    0.42
ms per token,  2399.19 tokens per second)
llama_print_timings: prompt eval time =   10598.30 ms /   106 tokens (   99.98
ms per token,    10.00 tokens per second)
llama_print_timings:        eval time =    5380.15 ms /    25 runs   (  215.21
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   16147.40 ms /   131 tokens
No. of rows:  79%| | 999/1258 [13:29:43<1:17:36, 1Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.08 ms /    20 runs   (    0.40
ms per token,  2475.25 tokens per second)
llama_print_timings: prompt eval time =    9677.54 ms /    86 tokens (  112.53
ms per token,     8.89 tokens per second)
llama_print_timings:        eval time =    4367.76 ms /    19 runs   (  229.88
ms per token,     4.35 tokens per second)
llama_print_timings:       total time =   14176.00 ms /   105 tokens
No. of rows:  79%| | 1000/1258 [13:29:57<1:12:24, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.52 ms /    31 runs   (    0.40
ms per token,  2475.45 tokens per second)
llama_print_timings: prompt eval time =   10935.45 ms /   111 tokens (   98.52
ms per token,    10.15 tokens per second)
llama_print_timings:        eval time =    6425.41 ms /    30 runs   (  214.18
ms per token,     4.67 tokens per second)
```

```
llama_print_timings:       total time =   17557.59 ms /   141 tokens
No. of rows:  80%| | 1001/1258 [13:30:15<1:13:03, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.11 ms /    20 runs   (    0.41
ms per token,  2465.48 tokens per second)
llama_print_timings: prompt eval time =    9311.02 ms /    92 tokens (  101.21
ms per token,     9.88 tokens per second)
llama_print_timings:        eval time =    4055.05 ms /    19 runs   (  213.42
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   13494.13 ms /   111 tokens
No. of rows:  80%| | 1002/1258 [13:30:28<1:08:13, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.53 ms /    28 runs   (    0.41
ms per token,  2429.50 tokens per second)
llama_print_timings: prompt eval time =   10518.29 ms /    93 tokens (  113.10
ms per token,     8.84 tokens per second)
llama_print_timings:        eval time =    5837.84 ms /    27 runs   (  216.22
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   16533.17 ms /   120 tokens
No. of rows:  80%| | 1003/1258 [13:30:45<1:08:39, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.61 ms /    24 runs   (    0.40
ms per token,  2496.36 tokens per second)
llama_print_timings: prompt eval time =    8590.11 ms /    81 tokens (  106.05
ms per token,     9.43 tokens per second)
llama_print_timings:        eval time =    6681.24 ms /    23 runs   (  290.49
ms per token,     3.44 tokens per second)
llama_print_timings:       total time =   15423.41 ms /   104 tokens
No. of rows:  80%| | 1004/1258 [13:31:00<1:07:28, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.10 ms /    28 runs   (    0.40
ms per token,  2523.43 tokens per second)
llama_print_timings: prompt eval time =   10985.52 ms /   100 tokens (  109.86
ms per token,     9.10 tokens per second)
llama_print_timings:        eval time =    5725.46 ms /    27 runs   (  212.05
ms per token,     4.72 tokens per second)
llama_print_timings:       total time =   16890.02 ms /   127 tokens
No. of rows:  80%| | 1005/1258 [13:31:17<1:08:25, Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      20.59 ms /    50 runs   (    0.41
ms per token,  2428.95 tokens per second)
llama_print_timings: prompt eval time =   13217.03 ms /   133 tokens (   99.38
ms per token,    10.06 tokens per second)
llama_print_timings:        eval time =   12248.66 ms /    49 runs   (  249.97
ms per token,     4.00 tokens per second)
llama_print_timings:       total time =   25790.62 ms /   182 tokens
No. of rows:  80%| | 1006/1258 [13:31:43<1:20:12, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.82 ms /    27 runs   (    0.44
ms per token,  2284.07 tokens per second)
llama_print_timings: prompt eval time =    9594.32 ms /    95 tokens (  100.99
ms per token,     9.90 tokens per second)
llama_print_timings:        eval time =    5535.82 ms /    26 runs   (  212.92
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   15303.17 ms /   121 tokens
No. of rows:  80%| | 1007/1258 [13:31:58<1:15:08, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.91 ms /    26 runs   (    0.42
ms per token,  2382.48 tokens per second)
llama_print_timings: prompt eval time =    9118.16 ms /    87 tokens (  104.81
ms per token,     9.54 tokens per second)
llama_print_timings:        eval time =    5283.46 ms /    25 runs   (  211.34
ms per token,     4.73 tokens per second)
llama_print_timings:       total time =   14570.64 ms /   112 tokens
No. of rows:  80%| | 1008/1258 [13:32:13<1:10:36, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.01 ms /    26 runs   (    0.38
ms per token,  2597.92 tokens per second)
llama_print_timings: prompt eval time =   11944.34 ms /   103 tokens (  115.96
ms per token,     8.62 tokens per second)
llama_print_timings:        eval time =    5422.17 ms /    25 runs   (  216.89
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   17530.53 ms /   128 tokens
No. of rows:  80%| | 1009/1258 [13:32:30<1:11:03, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.64 ms /    31 runs   (    0.41
ms per token,  2451.76 tokens per second)
llama_print_timings: prompt eval time =   13010.68 ms /   116 tokens (  112.16
```

```
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    6393.82 ms /    30 runs   (  213.13
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   19604.09 ms /   146 tokens
No. of rows:  80%| | 1010/1258 [13:32:50<1:13:51, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.98 ms /    32 runs   (    0.41
ms per token,  2464.76 tokens per second)
llama_print_timings: prompt eval time =   10798.46 ms /   110 tokens (   98.17
ms per token,    10.19 tokens per second)
llama_print_timings:        eval time =    6599.50 ms /    31 runs   (  212.89
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   17604.88 ms /   141 tokens
No. of rows:  80%| | 1011/1258 [13:33:07<1:13:14, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.75 ms /    31 runs   (    0.41
ms per token,  2431.37 tokens per second)
llama_print_timings: prompt eval time =   13149.81 ms /   112 tokens (  117.41
ms per token,     8.52 tokens per second)
llama_print_timings:        eval time =    6487.42 ms /    30 runs   (  216.25
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   19833.76 ms /   142 tokens
No. of rows:  80%| | 1012/1258 [13:33:27<1:15:28, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.62 ms /    19 runs   (    0.40
ms per token,  2492.46 tokens per second)
llama_print_timings: prompt eval time =    9908.95 ms /    83 tokens (  119.39
ms per token,     8.38 tokens per second)
llama_print_timings:        eval time =    3814.15 ms /    18 runs   (  211.90
ms per token,     4.72 tokens per second)
llama_print_timings:       total time =   13848.27 ms /   101 tokens
No. of rows:  81%| | 1013/1258 [13:33:41<1:09:35, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       4.32 ms /    11 runs   (    0.39
ms per token,  2543.94 tokens per second)
llama_print_timings: prompt eval time =    7279.04 ms /    66 tokens (  110.29
ms per token,     9.07 tokens per second)
llama_print_timings:        eval time =    2096.44 ms /    10 runs   (  209.64
ms per token,     4.77 tokens per second)
llama_print_timings:       total time =    9446.94 ms /    76 tokens
```

```
No. of rows:  81%| | 1014/1258 [13:33:51<1:00:02, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.83 ms /    24 runs   (    0.41
ms per token,  2442.50 tokens per second)
llama_print_timings: prompt eval time =    9993.34 ms /    98 tokens (  101.97
ms per token,     9.81 tokens per second)
llama_print_timings:        eval time =    4894.03 ms /    23 runs   (  212.78
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   15045.54 ms /   121 tokens
No. of rows:  81%| | 1015/1258 [13:34:06<1:00:09, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.71 ms /    38 runs   (    0.41
ms per token,  2419.00 tokens per second)
llama_print_timings: prompt eval time =   12485.17 ms /   120 tokens (  104.04
ms per token,     9.61 tokens per second)
llama_print_timings:        eval time =    9629.01 ms /    37 runs   (  260.24
ms per token,     3.84 tokens per second)
llama_print_timings:       total time =   22362.06 ms /   157 tokens
No. of rows:  81%| | 1016/1258 [13:34:28<1:09:00, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.57 ms /    50 runs   (    0.39
ms per token,  2554.41 tokens per second)
llama_print_timings: prompt eval time =   14560.81 ms /   139 tokens (  104.75
ms per token,     9.55 tokens per second)
llama_print_timings:        eval time =   10578.51 ms /    49 runs   (  215.89
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   25454.34 ms /   188 tokens
No. of rows:  81%| | 1017/1258 [13:34:53<1:18:47, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.63 ms /    31 runs   (    0.41
ms per token,  2454.67 tokens per second)
llama_print_timings: prompt eval time =   10553.67 ms /   104 tokens (  101.48
ms per token,     9.85 tokens per second)
llama_print_timings:        eval time =    6364.32 ms /    30 runs   (  212.14
ms per token,     4.71 tokens per second)
llama_print_timings:       total time =   17118.03 ms /   134 tokens
No. of rows:  81%| | 1018/1258 [13:35:11<1:15:28, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       8.00 ms /    20 runs   (    0.40
ms per token,  2501.56 tokens per second)
llama_print_timings: prompt eval time =    9025.49 ms /    93 tokens (   97.05
ms per token,    10.30 tokens per second)
llama_print_timings:        eval time =    4029.69 ms /    19 runs   (  212.09
ms per token,     4.72 tokens per second)
llama_print_timings:       total time =   13182.35 ms /   112 tokens
No. of rows:  81%| | 1019/1258 [13:35:24<1:08:22, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.70 ms /    40 runs   (    0.44
ms per token,  2259.76 tokens per second)
llama_print_timings: prompt eval time =   13640.35 ms /   138 tokens (   98.84
ms per token,    10.12 tokens per second)
llama_print_timings:        eval time =    8373.52 ms /    39 runs   (  214.71
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   22273.65 ms /   177 tokens
No. of rows:  81%| | 1020/1258 [13:35:46<1:14:10, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.84 ms /    27 runs   (    0.40
ms per token,  2489.86 tokens per second)
llama_print_timings: prompt eval time =   10515.63 ms /   105 tokens (  100.15
ms per token,     9.99 tokens per second)
llama_print_timings:        eval time =    5534.23 ms /    26 runs   (  212.86
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   16228.93 ms /   131 tokens
No. of rows:  81%| | 1021/1258 [13:36:02<1:10:56, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.23 ms /    28 runs   (    0.40
ms per token,  2493.10 tokens per second)
llama_print_timings: prompt eval time =   10443.94 ms /   107 tokens (   97.61
ms per token,    10.25 tokens per second)
llama_print_timings:        eval time =    5905.41 ms /    27 runs   (  218.72
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   16532.72 ms /   134 tokens
No. of rows:  81%| | 1022/1258 [13:36:19<1:08:58, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.62 ms /    20 runs   (    0.43
ms per token,  2320.45 tokens per second)
llama_print_timings: prompt eval time =    9497.49 ms /    95 tokens (   99.97
ms per token,    10.00 tokens per second)
```

```
llama_print_timings:        eval time =    4104.31 ms /    19 runs   (  216.02
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   13732.35 ms /   114 tokens
No. of rows:  81%| | 1023/1258 [13:36:33<1:04:13, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.64 ms /    24 runs   (    0.40
ms per token,  2490.66 tokens per second)
llama_print_timings: prompt eval time =   10281.06 ms /    88 tokens (  116.83
ms per token,     8.56 tokens per second)
llama_print_timings:        eval time =    4917.71 ms /    23 runs   (  213.81
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   15350.38 ms /   111 tokens
No. of rows:  81%| | 1024/1258 [13:36:48<1:02:43, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.49 ms /    23 runs   (    0.41
ms per token,  2422.58 tokens per second)
llama_print_timings: prompt eval time =   12068.03 ms /   108 tokens (  111.74
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    4723.63 ms /    22 runs   (  214.71
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   16934.22 ms /   130 tokens
No. of rows:  81%| | 1025/1258 [13:37:05<1:03:27, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.78 ms /    26 runs   (    0.41
ms per token,  2412.55 tokens per second)
llama_print_timings: prompt eval time =   10491.33 ms /    92 tokens (  114.04
ms per token,     8.77 tokens per second)
llama_print_timings:        eval time =    5362.46 ms /    25 runs   (  214.50
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   16027.00 ms /   117 tokens
No. of rows:  82%| | 1026/1258 [13:37:21<1:02:50, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.98 ms /    24 runs   (    0.42
ms per token,  2405.29 tokens per second)
llama_print_timings: prompt eval time =    9403.47 ms /    93 tokens (  101.11
ms per token,     9.89 tokens per second)
llama_print_timings:        eval time =    4910.89 ms /    23 runs   (  213.52
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   14467.76 ms /   116 tokens
No. of rows:  82%| | 1027/1258 [13:37:35<1:00:30, Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.73 ms /    25 runs   (    0.39
ms per token,  2569.11 tokens per second)
llama_print_timings: prompt eval time =    9845.45 ms /   100 tokens (   98.45
ms per token,    10.16 tokens per second)
llama_print_timings:        eval time =    5090.85 ms /    24 runs   (  212.12
ms per token,     4.71 tokens per second)
llama_print_timings:       total time =   15094.34 ms /   124 tokens
No. of rows:  82%| | 1028/1258 [13:37:51<59:32, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.14 ms /    19 runs   (    0.43
ms per token,  2334.73 tokens per second)
llama_print_timings: prompt eval time =    8285.48 ms /    80 tokens (  103.57
ms per token,     9.66 tokens per second)
llama_print_timings:        eval time =    3850.29 ms /    18 runs   (  213.91
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   12262.77 ms /    98 tokens
No. of rows:  82%| | 1029/1258 [13:38:03<55:32, 14Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.18 ms /    29 runs   (    0.42
ms per token,  2380.17 tokens per second)
llama_print_timings: prompt eval time =    9245.66 ms /    87 tokens (  106.27
ms per token,     9.41 tokens per second)
llama_print_timings:        eval time =    6177.36 ms /    28 runs   (  220.62
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   15617.64 ms /   115 tokens
No. of rows:  82%| | 1030/1258 [13:38:18<56:31, 14Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.41 ms /    22 runs   (    0.43
ms per token,  2337.94 tokens per second)
llama_print_timings: prompt eval time =    9463.12 ms /    94 tokens (  100.67
ms per token,     9.93 tokens per second)
llama_print_timings:        eval time =    4480.19 ms /    21 runs   (  213.34
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   14086.07 ms /   115 tokens
No. of rows:  82%| | 1031/1258 [13:38:32<55:23, 14Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.21 ms /    26 runs   (    0.43
```

```
ms per token,  2318.53 tokens per second)
llama_print_timings: prompt eval time =    9779.25 ms /    97 tokens (  100.82
ms per token,     9.92 tokens per second)
llama_print_timings:        eval time =    5371.53 ms /    25 runs   (  214.86
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   15317.29 ms /   122 tokens
No. of rows:  82%| | 1032/1258 [13:38:48<55:55, 14Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.69 ms /    25 runs   (    0.43
ms per token,  2337.98 tokens per second)
llama_print_timings: prompt eval time =    8854.19 ms /    87 tokens (  101.77
ms per token,     9.83 tokens per second)
llama_print_timings:        eval time =    5135.11 ms /    24 runs   (  213.96
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   14151.20 ms /   111 tokens
No. of rows:  82%| | 1033/1258 [13:39:02<54:53, 14Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      20.36 ms /    45 runs   (    0.45
ms per token,  2209.78 tokens per second)
llama_print_timings: prompt eval time =   14232.90 ms /   135 tokens (  105.43
ms per token,     9.49 tokens per second)
llama_print_timings:        eval time =   11101.90 ms /    44 runs   (  252.32
ms per token,     3.96 tokens per second)
llama_print_timings:       total time =   25627.84 ms /   179 tokens
No. of rows:  82%| | 1034/1258 [13:39:28<1:06:58, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.91 ms /    30 runs   (    0.40
ms per token,  2518.26 tokens per second)
llama_print_timings: prompt eval time =   11260.77 ms /   111 tokens (  101.45
ms per token,     9.86 tokens per second)
llama_print_timings:        eval time =    7838.11 ms /    29 runs   (  270.28
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   19288.43 ms /   140 tokens
No. of rows:  82%| | 1035/1258 [13:39:47<1:08:11, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.87 ms /    33 runs   (    0.42
ms per token,  2380.09 tokens per second)
llama_print_timings: prompt eval time =    9961.06 ms /    99 tokens (  100.62
ms per token,     9.94 tokens per second)
llama_print_timings:        eval time =    6954.54 ms /    32 runs   (  217.33
```

```
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   17130.66 ms /   131 tokens
No. of rows:  82%| | 1036/1258 [13:40:04<1:06:32, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.20 ms /    37 runs   (    0.41
ms per token,  2434.21 tokens per second)
llama_print_timings: prompt eval time =   11333.17 ms /   115 tokens (   98.55
ms per token,    10.15 tokens per second)
llama_print_timings:        eval time =    7899.76 ms /    36 runs   (  219.44
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   19473.19 ms /   151 tokens
No. of rows:  82%| | 1037/1258 [13:40:24<1:07:53, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.78 ms /    24 runs   (    0.41
ms per token,  2455.24 tokens per second)
llama_print_timings: prompt eval time =    8160.38 ms /    81 tokens (  100.75
ms per token,     9.93 tokens per second)
llama_print_timings:        eval time =    4937.84 ms /    23 runs   (  214.69
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   13255.64 ms /   104 tokens
No. of rows:  83%| | 1038/1258 [13:40:37<1:01:54, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.39 ms /    29 runs   (    0.43
ms per token,  2339.65 tokens per second)
llama_print_timings: prompt eval time =    9040.60 ms /    87 tokens (  103.91
ms per token,     9.62 tokens per second)
llama_print_timings:        eval time =    6023.27 ms /    28 runs   (  215.12
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   15251.84 ms /   115 tokens
No. of rows:  83%| | 1039/1258 [13:40:52<59:50, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.60 ms /    32 runs   (    0.43
ms per token,  2352.60 tokens per second)
llama_print_timings: prompt eval time =   11408.38 ms /   104 tokens (  109.70
ms per token,     9.12 tokens per second)
llama_print_timings:        eval time =    6608.83 ms /    31 runs   (  213.19
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   18225.25 ms /   135 tokens
No. of rows:  83%| | 1040/1258 [13:41:10<1:01:34, Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.43 ms /    19 runs   (    0.39
ms per token,  2558.58 tokens per second)
llama_print_timings: prompt eval time =    9525.96 ms /    77 tokens (  123.71
ms per token,     8.08 tokens per second)
llama_print_timings:        eval time =    4154.64 ms /    18 runs   (  230.81
ms per token,     4.33 tokens per second)
llama_print_timings:       total time =   13800.80 ms /    95 tokens
No. of rows:  83%| | 1041/1258 [13:41:24<57:53, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.47 ms /    37 runs   (    0.42
ms per token,  2392.04 tokens per second)
llama_print_timings: prompt eval time =   12820.15 ms /   122 tokens (  105.08
ms per token,     9.52 tokens per second)
llama_print_timings:        eval time =    7794.56 ms /    36 runs   (  216.52
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   20855.82 ms /   158 tokens
No. of rows:  83%| | 1042/1258 [13:41:45<1:02:51, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.09 ms /    32 runs   (    0.41
ms per token,  2445.36 tokens per second)
llama_print_timings: prompt eval time =   10303.91 ms /   104 tokens (   99.08
ms per token,    10.09 tokens per second)
llama_print_timings:        eval time =    6851.87 ms /    31 runs   (  221.03
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   17365.04 ms /   135 tokens
No. of rows:  83%| | 1043/1258 [13:42:02<1:02:28, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.70 ms /    28 runs   (    0.38
ms per token,  2617.31 tokens per second)
llama_print_timings: prompt eval time =    9176.03 ms /    86 tokens (  106.70
ms per token,     9.37 tokens per second)
llama_print_timings:        eval time =    5900.16 ms /    27 runs   (  218.52
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   15253.53 ms /   113 tokens
No. of rows:  83%| | 1044/1258 [13:42:18<59:51, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.56 ms /    36 runs   (    0.43
ms per token,  2314.07 tokens per second)
```

```
llama_print_timings: prompt eval time =   14947.07 ms /   133 tokens (  112.38
ms per token,     8.90 tokens per second)
llama_print_timings:        eval time =    7634.56 ms /    35 runs   (  218.13
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   22816.24 ms /   168 tokens
No. of rows:  83%| | 1045/1258 [13:42:40<1:06:01, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      20.59 ms /    50 runs   (    0.41
ms per token,  2428.95 tokens per second)
llama_print_timings: prompt eval time =   12636.04 ms /   113 tokens (  111.82
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =   10699.68 ms /    49 runs   (  218.36
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   23662.96 ms /   162 tokens
No. of rows:  83%| | 1046/1258 [13:43:04<1:11:05, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.15 ms /    22 runs   (    0.42
ms per token,  2403.32 tokens per second)
llama_print_timings: prompt eval time =    8697.01 ms /    85 tokens (  102.32
ms per token,     9.77 tokens per second)
llama_print_timings:        eval time =    4485.67 ms /    21 runs   (  213.60
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   13323.13 ms /   106 tokens
No. of rows:  83%| | 1047/1258 [13:43:17<1:03:35, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.01 ms /    38 runs   (    0.40
ms per token,  2531.48 tokens per second)
llama_print_timings: prompt eval time =   13594.25 ms /   115 tokens (  118.21
ms per token,     8.46 tokens per second)
llama_print_timings:        eval time =    7964.92 ms /    37 runs   (  215.27
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   21803.86 ms /   152 tokens
No. of rows:  83%| | 1048/1258 [13:43:39<1:07:12, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.46 ms /    24 runs   (    0.52
ms per token,  1926.32 tokens per second)
llama_print_timings: prompt eval time =    9395.57 ms /    80 tokens (  117.44
ms per token,     8.51 tokens per second)
llama_print_timings:        eval time =    5243.02 ms /    23 runs   (  227.96
ms per token,     4.39 tokens per second)
```

```
llama_print_timings:       total time =   14797.77 ms /   103 tokens
No. of rows:  83%| | 1049/1258 [13:43:54<1:02:17, Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.13 ms /    22 runs   (    0.42
ms per token,  2409.37 tokens per second)
llama_print_timings: prompt eval time =    8992.41 ms /    87 tokens (  103.36
ms per token,     9.67 tokens per second)
llama_print_timings:        eval time =    4602.59 ms /    21 runs   (  219.17
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   13735.30 ms /   108 tokens
No. of rows:  83%| | 1050/1258 [13:44:08<57:41, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.61 ms /    23 runs   (    0.42
ms per token,  2394.09 tokens per second)
llama_print_timings: prompt eval time =    8973.28 ms /    90 tokens (   99.70
ms per token,    10.03 tokens per second)
llama_print_timings:        eval time =    4949.39 ms /    22 runs   (  224.97
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   14073.31 ms /   112 tokens
No. of rows:  84%| | 1051/1258 [13:44:22<54:45, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.19 ms /    28 runs   (    0.44
ms per token,  2297.53 tokens per second)
llama_print_timings: prompt eval time =   11020.86 ms /   104 tokens (  105.97
ms per token,     9.44 tokens per second)
llama_print_timings:        eval time =    5756.78 ms /    27 runs   (  213.21
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   16958.38 ms /   131 tokens
No. of rows:  84%| | 1052/1258 [13:44:39<55:37, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.50 ms /    20 runs   (    0.37
ms per token,  2667.02 tokens per second)
llama_print_timings: prompt eval time =    8426.85 ms /    83 tokens (  101.53
ms per token,     9.85 tokens per second)
llama_print_timings:        eval time =    4066.63 ms /    19 runs   (  214.03
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   12617.19 ms /   102 tokens
No. of rows:  84%| | 1053/1258 [13:44:51<51:41, 15Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.65 ms /    32 runs   (    0.43
ms per token,  2343.64 tokens per second)
llama_print_timings: prompt eval time =   10295.94 ms /   102 tokens (  100.94
ms per token,     9.91 tokens per second)
llama_print_timings:        eval time =    6616.31 ms /    31 runs   (  213.43
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   17116.82 ms /   133 tokens
No. of rows:  84%| | 1054/1258 [13:45:09<53:28, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.64 ms /    32 runs   (    0.43
ms per token,  2346.21 tokens per second)
llama_print_timings: prompt eval time =   10435.13 ms /    94 tokens (  111.01
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    6560.75 ms /    31 runs   (  211.64
ms per token,     4.73 tokens per second)
llama_print_timings:       total time =   17201.19 ms /   125 tokens
No. of rows:  84%| | 1055/1258 [13:45:26<54:42, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.11 ms /    30 runs   (    0.40
ms per token,  2476.68 tokens per second)
llama_print_timings: prompt eval time =   11292.62 ms /   100 tokens (  112.93
ms per token,     8.86 tokens per second)
llama_print_timings:        eval time =    6215.88 ms /    29 runs   (  214.34
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   17700.36 ms /   129 tokens
No. of rows:  84%| | 1056/1258 [13:45:43<55:59, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.48 ms /    33 runs   (    0.41
ms per token,  2448.07 tokens per second)
llama_print_timings: prompt eval time =   12465.85 ms /   125 tokens (   99.73
ms per token,    10.03 tokens per second)
llama_print_timings:        eval time =    7213.34 ms /    32 runs   (  225.42
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   19903.49 ms /   157 tokens
No. of rows:  84%| | 1057/1258 [13:46:03<59:00, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.65 ms /    19 runs   (    0.40
ms per token,  2485.28 tokens per second)
llama_print_timings: prompt eval time =    7070.57 ms /    71 tokens (   99.59
```

```
ms per token,    10.04 tokens per second)
llama_print_timings:        eval time =    3803.39 ms /    18 runs   (  211.30
ms per token,     4.73 tokens per second)
llama_print_timings:       total time =   10994.25 ms /    89 tokens
No. of rows:  84%| | 1058/1258 [13:46:14<52:06, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.44 ms /    26 runs   (    0.40
ms per token,  2489.71 tokens per second)
llama_print_timings: prompt eval time =   10512.69 ms /   101 tokens (  104.09
ms per token,     9.61 tokens per second)
llama_print_timings:        eval time =    5374.79 ms /    25 runs   (  214.99
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   16054.75 ms /   126 tokens
No. of rows:  84%| | 1059/1258 [13:46:30<52:16, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.83 ms /    22 runs   (    0.40
ms per token,  2492.64 tokens per second)
llama_print_timings: prompt eval time =    9169.51 ms /    88 tokens (  104.20
ms per token,     9.60 tokens per second)
llama_print_timings:        eval time =    4446.59 ms /    21 runs   (  211.74
ms per token,     4.72 tokens per second)
llama_print_timings:       total time =   13757.43 ms /   109 tokens
No. of rows:  84%| | 1060/1258 [13:46:44<50:02, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.80 ms /    22 runs   (    0.40
ms per token,  2500.57 tokens per second)
llama_print_timings: prompt eval time =    9689.19 ms /    83 tokens (  116.74
ms per token,     8.57 tokens per second)
llama_print_timings:        eval time =    4829.99 ms /    21 runs   (  230.00
ms per token,     4.35 tokens per second)
llama_print_timings:       total time =   14658.07 ms /   104 tokens
No. of rows:  84%| | 1061/1258 [13:46:59<49:17, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.66 ms /    35 runs   (    0.42
ms per token,  2387.61 tokens per second)
llama_print_timings: prompt eval time =   11465.66 ms /   116 tokens (   98.84
ms per token,    10.12 tokens per second)
llama_print_timings:        eval time =    7233.37 ms /    34 runs   (  212.75
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   18923.32 ms /   150 tokens
```

```
No. of rows:  84%| | 1062/1258 [13:47:18<52:52, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.59 ms /    37 runs   (    0.39
ms per token,  2535.81 tokens per second)
llama_print_timings: prompt eval time =   12906.15 ms /   114 tokens (  113.21
ms per token,     8.83 tokens per second)
llama_print_timings:        eval time =    7939.40 ms /    36 runs   (  220.54
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   21085.03 ms /   150 tokens
No. of rows:  84%| | 1063/1258 [13:47:39<57:23, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.74 ms /    21 runs   (    0.42
ms per token,  2401.92 tokens per second)
llama_print_timings: prompt eval time =    9255.24 ms /    81 tokens (  114.26
ms per token,     8.75 tokens per second)
llama_print_timings:        eval time =    4233.07 ms /    20 runs   (  211.65
ms per token,     4.72 tokens per second)
llama_print_timings:       total time =   13625.45 ms /   101 tokens
No. of rows:  85%| | 1064/1258 [13:47:53<53:11, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.45 ms /    33 runs   (    0.41
ms per token,  2454.08 tokens per second)
llama_print_timings: prompt eval time =   11854.85 ms /   116 tokens (  102.20
ms per token,     9.79 tokens per second)
llama_print_timings:        eval time =    6968.32 ms /    32 runs   (  217.76
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   19037.42 ms /   148 tokens
No. of rows:  85%| | 1065/1258 [13:48:12<55:25, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.21 ms /    38 runs   (    0.40
ms per token,  2497.54 tokens per second)
llama_print_timings: prompt eval time =   11746.44 ms /   115 tokens (  102.14
ms per token,     9.79 tokens per second)
llama_print_timings:        eval time =    7900.55 ms /    37 runs   (  213.53
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   19892.45 ms /   152 tokens
No. of rows:  85%| | 1066/1258 [13:48:31<57:41, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =      13.96 ms /    31 runs   (    0.45
ms per token,  2221.43 tokens per second)
llama_print_timings: prompt eval time =   11022.88 ms /   108 tokens (  102.06
ms per token,     9.80 tokens per second)
llama_print_timings:        eval time =    6407.76 ms /    30 runs   (  213.59
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   17628.59 ms /   138 tokens
No. of rows:  85%| | 1067/1258 [13:48:49<57:01, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.99 ms /    31 runs   (    0.42
ms per token,  2387.00 tokens per second)
llama_print_timings: prompt eval time =   10513.97 ms /   103 tokens (  102.08
ms per token,     9.80 tokens per second)
llama_print_timings:        eval time =    8262.38 ms /    30 runs   (  275.41
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   18976.88 ms /   133 tokens
No. of rows:  85%| | 1068/1258 [13:49:08<57:44, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.31 ms /    35 runs   (    0.44
ms per token,  2285.79 tokens per second)
llama_print_timings: prompt eval time =   10881.62 ms /   106 tokens (  102.66
ms per token,     9.74 tokens per second)
llama_print_timings:        eval time =    8982.16 ms /    34 runs   (  264.18
ms per token,     3.79 tokens per second)
llama_print_timings:       total time =   20088.33 ms /   140 tokens
No. of rows:  85%| | 1069/1258 [13:49:28<59:11, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.01 ms /    27 runs   (    0.41
ms per token,  2453.21 tokens per second)
llama_print_timings: prompt eval time =    9774.52 ms /    99 tokens (   98.73
ms per token,    10.13 tokens per second)
llama_print_timings:        eval time =    7004.10 ms /    26 runs   (  269.39
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   16952.13 ms /   125 tokens
No. of rows:  85%| | 1070/1258 [13:49:45<57:09, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.11 ms /    26 runs   (    0.39
ms per token,  2570.69 tokens per second)
llama_print_timings: prompt eval time =    9924.10 ms /    97 tokens (  102.31
ms per token,     9.77 tokens per second)
```

```
llama_print_timings:        eval time =    5389.99 ms /    25 runs   (  215.60
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   15475.49 ms /   122 tokens
No. of rows:  85%| | 1071/1258 [13:50:01<54:16, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.78 ms /    23 runs   (    0.38
ms per token,  2620.49 tokens per second)
llama_print_timings: prompt eval time =   10876.14 ms /   107 tokens (  101.65
ms per token,     9.84 tokens per second)
llama_print_timings:        eval time =    4772.02 ms /    22 runs   (  216.91
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   15788.55 ms /   129 tokens
No. of rows:  85%| | 1072/1258 [13:50:16<52:28, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.51 ms /    22 runs   (    0.39
ms per token,  2586.41 tokens per second)
llama_print_timings: prompt eval time =    9445.59 ms /    92 tokens (  102.67
ms per token,     9.74 tokens per second)
llama_print_timings:        eval time =    4444.70 ms /    21 runs   (  211.65
ms per token,     4.72 tokens per second)
llama_print_timings:       total time =   14032.46 ms /   113 tokens
No. of rows:  85%| | 1073/1258 [13:50:30<49:31, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.55 ms /    21 runs   (    0.45
ms per token,  2199.64 tokens per second)
llama_print_timings: prompt eval time =    9957.09 ms /    87 tokens (  114.45
ms per token,     8.74 tokens per second)
llama_print_timings:        eval time =    4222.39 ms /    20 runs   (  211.12
ms per token,     4.74 tokens per second)
llama_print_timings:       total time =   14316.40 ms /   107 tokens
No. of rows:  85%| | 1074/1258 [13:50:45<47:39, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.36 ms /    22 runs   (    0.43
ms per token,  2350.93 tokens per second)
llama_print_timings: prompt eval time =    8465.41 ms /    81 tokens (  104.51
ms per token,     9.57 tokens per second)
llama_print_timings:        eval time =    4469.15 ms /    21 runs   (  212.82
ms per token,     4.70 tokens per second)
llama_print_timings:       total time =   13075.39 ms /   102 tokens
No. of rows:  85%| | 1075/1258 [13:50:58<45:08, 14Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.06 ms /    25 runs   (    0.40
ms per token,  2483.85 tokens per second)
llama_print_timings: prompt eval time =   10713.95 ms /    95 tokens (  112.78
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    5068.92 ms /    24 runs   (  211.20
ms per token,     4.73 tokens per second)
llama_print_timings:       total time =   15945.35 ms /   119 tokens
No. of rows:  86%| | 1076/1258 [13:51:14<45:56, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.16 ms /    27 runs   (    0.38
ms per token,  2656.43 tokens per second)
llama_print_timings: prompt eval time =   12831.87 ms /   100 tokens (  128.32
ms per token,     7.79 tokens per second)
llama_print_timings:        eval time =    7390.54 ms /    26 runs   (  284.25
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   20385.92 ms /   126 tokens
No. of rows:  86%| | 1077/1258 [13:51:34<50:26, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.35 ms /    24 runs   (    0.39
ms per token,  2565.75 tokens per second)
llama_print_timings: prompt eval time =   10409.65 ms /    93 tokens (  111.93
ms per token,     8.93 tokens per second)
llama_print_timings:        eval time =    4997.03 ms /    23 runs   (  217.26
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   15556.80 ms /   116 tokens
No. of rows:  86%| | 1078/1258 [13:51:50<49:07, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.07 ms /    21 runs   (    0.38
ms per token,  2601.59 tokens per second)
llama_print_timings: prompt eval time =    9746.48 ms /    85 tokens (  114.66
ms per token,     8.72 tokens per second)
llama_print_timings:        eval time =    4463.12 ms /    20 runs   (  223.16
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   14338.23 ms /   105 tokens
No. of rows:  86%| | 1079/1258 [13:52:04<47:02, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.34 ms /    18 runs   (    0.41
```

```
ms per token,  2452.98 tokens per second)
llama_print_timings: prompt eval time =    9124.56 ms /    77 tokens (  118.50
ms per token,     8.44 tokens per second)
llama_print_timings:        eval time =    3677.46 ms /    17 runs   (  216.32
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   12916.24 ms /    94 tokens
No. of rows:  86%| | 1080/1258 [13:52:17<44:14, 14Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.88 ms /    22 runs   (    0.40
ms per token,  2476.64 tokens per second)
llama_print_timings: prompt eval time =   11294.93 ms /    97 tokens (  116.44
ms per token,     8.59 tokens per second)
llama_print_timings:        eval time =    4579.77 ms /    21 runs   (  218.08
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   16011.63 ms /   118 tokens
No. of rows:  86%| | 1081/1258 [13:52:33<44:58, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.01 ms /    26 runs   (    0.38
ms per token,  2598.18 tokens per second)
llama_print_timings: prompt eval time =   10351.39 ms /    92 tokens (  112.52
ms per token,     8.89 tokens per second)
llama_print_timings:        eval time =    5417.29 ms /    25 runs   (  216.69
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   15931.83 ms /   117 tokens
No. of rows:  86%| | 1082/1258 [13:52:49<45:20, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.18 ms /    30 runs   (    0.41
ms per token,  2462.25 tokens per second)
llama_print_timings: prompt eval time =   12326.89 ms /    98 tokens (  125.78
ms per token,     7.95 tokens per second)
llama_print_timings:        eval time =    6254.64 ms /    29 runs   (  215.68
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   18768.66 ms /   127 tokens
No. of rows:  86%| | 1083/1258 [13:53:08<47:58, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.66 ms /    25 runs   (    0.39
ms per token,  2588.80 tokens per second)
llama_print_timings: prompt eval time =    9992.82 ms /    87 tokens (  114.86
ms per token,     8.71 tokens per second)
llama_print_timings:        eval time =    5157.36 ms /    24 runs   (  214.89
```

```
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   15301.80 ms /   111 tokens
No. of rows:  86%| | 1084/1258 [13:53:23<46:42, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.05 ms /    29 runs   (    0.38
ms per token,  2623.48 tokens per second)
llama_print_timings: prompt eval time =   13502.90 ms /   110 tokens (  122.75
ms per token,     8.15 tokens per second)
llama_print_timings:        eval time =    8040.30 ms /    28 runs   (  287.15
ms per token,     3.48 tokens per second)
llama_print_timings:       total time =   21721.54 ms /   138 tokens
No. of rows:  86%| | 1085/1258 [13:53:45<51:18, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.59 ms /    20 runs   (    0.38
ms per token,  2633.66 tokens per second)
llama_print_timings: prompt eval time =    9582.03 ms /    82 tokens (  116.85
ms per token,     8.56 tokens per second)
llama_print_timings:        eval time =    4088.77 ms /    19 runs   (  215.20
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   13793.57 ms /   101 tokens
No. of rows:  86%| | 1086/1258 [13:53:59<47:34, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.91 ms /    50 runs   (    0.40
ms per token,  2510.92 tokens per second)
llama_print_timings: prompt eval time =   11715.20 ms /   105 tokens (  111.57
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =   10963.39 ms /    49 runs   (  223.74
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   22995.42 ms /   154 tokens
No. of rows:  86%| | 1087/1258 [13:54:22<52:46, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       4.26 ms /    11 runs   (    0.39
ms per token,  2584.59 tokens per second)
llama_print_timings: prompt eval time =    8101.45 ms /    70 tokens (  115.74
ms per token,     8.64 tokens per second)
llama_print_timings:        eval time =    2159.84 ms /    10 runs   (  215.98
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   10328.76 ms /    80 tokens
No. of rows:  86%| | 1088/1258 [13:54:32<45:30, 16Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.83 ms /    28 runs   (    0.39
ms per token,  2586.37 tokens per second)
llama_print_timings: prompt eval time =   11615.49 ms /   105 tokens (  110.62
ms per token,     9.04 tokens per second)
llama_print_timings:        eval time =    5871.20 ms /    27 runs   (  217.45
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   17658.51 ms /   132 tokens
No. of rows:  87%| | 1089/1258 [13:54:50<46:36, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.76 ms /    20 runs   (    0.39
ms per token,  2575.99 tokens per second)
llama_print_timings: prompt eval time =    8780.15 ms /    76 tokens (  115.53
ms per token,     8.66 tokens per second)
llama_print_timings:        eval time =    4220.86 ms /    19 runs   (  222.15
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   13125.12 ms /    95 tokens
No. of rows:  87%| | 1090/1258 [13:55:03<43:27, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.76 ms /    18 runs   (    0.38
ms per token,  2664.30 tokens per second)
llama_print_timings: prompt eval time =    9764.53 ms /    88 tokens (  110.96
ms per token,     9.01 tokens per second)
llama_print_timings:        eval time =    3631.93 ms /    17 runs   (  213.64
ms per token,     4.68 tokens per second)
llama_print_timings:       total time =   13507.22 ms /   105 tokens
No. of rows:  87%| | 1091/1258 [13:55:16<41:31, 14Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.01 ms /    18 runs   (    0.39
ms per token,  2567.76 tokens per second)
llama_print_timings: prompt eval time =    9358.95 ms /    82 tokens (  114.13
ms per token,     8.76 tokens per second)
llama_print_timings:        eval time =    3821.01 ms /    17 runs   (  224.77
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   13293.59 ms /    99 tokens
No. of rows:  87%| | 1092/1258 [13:55:30<39:55, 14Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.94 ms /    19 runs   (    0.37
ms per token,  2739.73 tokens per second)
```

```
llama_print_timings: prompt eval time =    9761.15 ms /    84 tokens (  116.20
ms per token,     8.61 tokens per second)
llama_print_timings:        eval time =    3873.93 ms /    18 runs   (  215.22
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   13747.39 ms /   102 tokens
No. of rows:  87%| | 1093/1258 [13:55:43<39:07, 14Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.85 ms /    20 runs   (    0.44
ms per token,  2260.14 tokens per second)
llama_print_timings: prompt eval time =    8878.04 ms /    78 tokens (  113.82
ms per token,     8.79 tokens per second)
llama_print_timings:        eval time =    4066.17 ms /    19 runs   (  214.01
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   13071.17 ms /    97 tokens
No. of rows:  87%| | 1094/1258 [13:55:56<37:57, 13Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.57 ms /    26 runs   (    0.41
ms per token,  2459.09 tokens per second)
llama_print_timings: prompt eval time =   10968.74 ms /    96 tokens (  114.26
ms per token,     8.75 tokens per second)
llama_print_timings:        eval time =    5472.18 ms /    25 runs   (  218.89
ms per token,     4.57 tokens per second)
llama_print_timings:       total time =   16603.52 ms /   121 tokens
No. of rows:  87%| | 1095/1258 [13:56:13<39:56, 14Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.56 ms /    49 runs   (    0.40
ms per token,  2505.37 tokens per second)
llama_print_timings: prompt eval time =   18644.00 ms /   157 tokens (  118.75
ms per token,     8.42 tokens per second)
llama_print_timings:        eval time =   10806.22 ms /    48 runs   (  225.13
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   29761.73 ms /   205 tokens
No. of rows:  87%| | 1096/1258 [13:56:43<51:53, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.23 ms /    26 runs   (    0.39
ms per token,  2541.54 tokens per second)
llama_print_timings: prompt eval time =   11553.42 ms /   102 tokens (  113.27
ms per token,     8.83 tokens per second)
llama_print_timings:        eval time =    5410.47 ms /    25 runs   (  216.42
ms per token,     4.62 tokens per second)
```

```
llama_print_timings:       total time =   17129.34 ms /   127 tokens
No. of rows:  87%| | 1097/1258 [13:57:00<49:54, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.69 ms /    28 runs   (    0.38
ms per token,  2618.29 tokens per second)
llama_print_timings: prompt eval time =   12156.91 ms /   100 tokens (  121.57
ms per token,     8.23 tokens per second)
llama_print_timings:        eval time =    5826.78 ms /    27 runs   (  215.81
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   18155.38 ms /   127 tokens
No. of rows:  87%| | 1098/1258 [13:57:18<49:14, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.18 ms /    22 runs   (    0.51
ms per token,  1968.68 tokens per second)
llama_print_timings: prompt eval time =   11348.96 ms /    88 tokens (  128.97
ms per token,     7.75 tokens per second)
llama_print_timings:        eval time =    4544.55 ms /    21 runs   (  216.41
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   16032.69 ms /   109 tokens
No. of rows:  87%| | 1099/1258 [13:57:34<47:00, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.88 ms /    18 runs   (    0.38
ms per token,  2618.18 tokens per second)
llama_print_timings: prompt eval time =    8683.67 ms /    76 tokens (  114.26
ms per token,     8.75 tokens per second)
llama_print_timings:        eval time =    3648.74 ms /    17 runs   (  214.63
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   12440.76 ms /    93 tokens
No. of rows:  87%| | 1100/1258 [13:57:47<42:32, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.75 ms /    34 runs   (    0.40
ms per token,  2473.09 tokens per second)
llama_print_timings: prompt eval time =   13292.81 ms /   108 tokens (  123.08
ms per token,     8.12 tokens per second)
llama_print_timings:        eval time =    7199.77 ms /    33 runs   (  218.17
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   20713.84 ms /   141 tokens
No. of rows:  88%| | 1101/1258 [13:58:07<45:50, 17Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.79 ms /    23 runs   (    0.38
ms per token,  2618.10 tokens per second)
llama_print_timings: prompt eval time =   12578.83 ms /   100 tokens (  125.79
ms per token,     7.95 tokens per second)
llama_print_timings:        eval time =    4924.65 ms /    22 runs   (  223.85
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   17643.66 ms /   122 tokens
No. of rows:  88%| | 1102/1258 [13:58:25<45:39, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.98 ms /    25 runs   (    0.40
ms per token,  2504.76 tokens per second)
llama_print_timings: prompt eval time =   11078.76 ms /    97 tokens (  114.21
ms per token,     8.76 tokens per second)
llama_print_timings:        eval time =    5188.77 ms /    24 runs   (  216.20
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   16420.55 ms /   121 tokens
No. of rows:  88%| | 1103/1258 [13:58:41<44:29, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.57 ms /    38 runs   (    0.38
ms per token,  2607.74 tokens per second)
llama_print_timings: prompt eval time =   13701.18 ms /   121 tokens (  113.23
ms per token,     8.83 tokens per second)
llama_print_timings:        eval time =   10126.36 ms /    37 runs   (  273.69
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   24066.26 ms /   158 tokens
No. of rows:  88%| | 1104/1258 [13:59:05<49:28, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.39 ms /    33 runs   (    0.38
ms per token,  2663.87 tokens per second)
llama_print_timings: prompt eval time =   14150.65 ms /   127 tokens (  111.42
ms per token,     8.97 tokens per second)
llama_print_timings:        eval time =    6912.79 ms /    32 runs   (  216.02
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   21264.62 ms /   159 tokens
No. of rows:  88%| | 1105/1258 [13:59:27<50:41, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.13 ms /    21 runs   (    0.39
ms per token,  2584.30 tokens per second)
llama_print_timings: prompt eval time =   10010.90 ms /    87 tokens (  115.07
```

```
ms per token,     8.69 tokens per second)
llama_print_timings:        eval time =    4280.09 ms /    20 runs   (  214.00
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   14418.84 ms /   107 tokens
No. of rows:  88%| | 1106/1258 [13:59:41<46:12, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.30 ms /    26 runs   (    0.40
ms per token,  2525.25 tokens per second)
llama_print_timings: prompt eval time =   13108.23 ms /   105 tokens (  124.84
ms per token,     8.01 tokens per second)
llama_print_timings:        eval time =    5373.65 ms /    25 runs   (  214.95
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   18643.16 ms /   130 tokens
No. of rows:  88%| | 1107/1258 [14:00:00<46:12, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.80 ms /    28 runs   (    0.39
ms per token,  2592.11 tokens per second)
llama_print_timings: prompt eval time =   13065.07 ms /   105 tokens (  124.43
ms per token,     8.04 tokens per second)
llama_print_timings:        eval time =    5829.48 ms /    27 runs   (  215.91
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   19067.12 ms /   132 tokens
No. of rows:  88%| | 1108/1258 [14:00:19<46:26, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.89 ms /    28 runs   (    0.39
ms per token,  2569.99 tokens per second)
llama_print_timings: prompt eval time =   15491.71 ms /   120 tokens (  129.10
ms per token,     7.75 tokens per second)
llama_print_timings:        eval time =    5852.80 ms /    27 runs   (  216.77
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   21514.06 ms /   147 tokens
No. of rows:  88%| | 1109/1258 [14:00:40<48:19, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.71 ms /    29 runs   (    0.37
ms per token,  2709.01 tokens per second)
llama_print_timings: prompt eval time =   11562.32 ms /   102 tokens (  113.36
ms per token,     8.82 tokens per second)
llama_print_timings:        eval time =    6051.87 ms /    28 runs   (  216.14
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   17790.47 ms /   130 tokens
```

```
No. of rows:  88%| | 1110/1258 [14:00:58<46:46, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.04 ms /    28 runs   (    0.39
ms per token,  2537.15 tokens per second)
llama_print_timings: prompt eval time =   10570.47 ms /    87 tokens (  121.50
ms per token,     8.23 tokens per second)
llama_print_timings:        eval time =    5826.25 ms /    27 runs   (  215.79
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   16571.58 ms /   114 tokens
No. of rows:  88%| | 1111/1258 [14:01:15<44:42, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.32 ms /    39 runs   (    0.39
ms per token,  2546.02 tokens per second)
llama_print_timings: prompt eval time =   13032.95 ms /   105 tokens (  124.12
ms per token,     8.06 tokens per second)
llama_print_timings:        eval time =    8243.56 ms /    38 runs   (  216.94
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   21523.99 ms /   143 tokens
No. of rows:  88%| | 1112/1258 [14:01:36<46:48, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.87 ms /    30 runs   (    0.40
ms per token,  2528.45 tokens per second)
llama_print_timings: prompt eval time =   14847.15 ms /   120 tokens (  123.73
ms per token,     8.08 tokens per second)
llama_print_timings:        eval time =    6245.17 ms /    29 runs   (  215.35
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   21281.64 ms /   149 tokens
No. of rows:  88%| | 1113/1258 [14:01:58<47:58, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.81 ms /    18 runs   (    0.38
ms per token,  2644.34 tokens per second)
llama_print_timings: prompt eval time =    9438.55 ms /    82 tokens (  115.10
ms per token,     8.69 tokens per second)
llama_print_timings:        eval time =    3651.45 ms /    17 runs   (  214.79
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   13197.99 ms /    99 tokens
No. of rows:  89%| | 1114/1258 [14:02:11<42:51, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       8.51 ms /    21 runs   (    0.41
ms per token,  2466.82 tokens per second)
llama_print_timings: prompt eval time =   10285.35 ms /    79 tokens (  130.19
ms per token,     7.68 tokens per second)
llama_print_timings:        eval time =    4675.82 ms /    20 runs   (  233.79
ms per token,     4.28 tokens per second)
llama_print_timings:       total time =   15101.69 ms /    99 tokens
No. of rows:  89%| | 1115/1258 [14:02:26<40:35, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.38 ms /    37 runs   (    0.39
ms per token,  2572.30 tokens per second)
llama_print_timings: prompt eval time =   11556.50 ms /   102 tokens (  113.30
ms per token,     8.83 tokens per second)
llama_print_timings:        eval time =    7814.35 ms /    36 runs   (  217.07
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   19599.68 ms /   138 tokens
No. of rows:  89%| | 1116/1258 [14:02:46<42:08, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.27 ms /    22 runs   (    0.38
ms per token,  2660.22 tokens per second)
llama_print_timings: prompt eval time =    9697.04 ms /    84 tokens (  115.44
ms per token,     8.66 tokens per second)
llama_print_timings:        eval time =    4492.18 ms /    21 runs   (  213.91
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   14322.62 ms /   105 tokens
No. of rows:  89%| | 1117/1258 [14:03:00<39:23, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.02 ms /    37 runs   (    0.41
ms per token,  2462.73 tokens per second)
llama_print_timings: prompt eval time =   15014.98 ms /   135 tokens (  111.22
ms per token,     8.99 tokens per second)
llama_print_timings:        eval time =    9611.95 ms /    36 runs   (  267.00
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   24864.54 ms /   171 tokens
No. of rows:  89%| | 1118/1258 [14:03:25<44:47, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.54 ms /    22 runs   (    0.39
ms per token,  2575.51 tokens per second)
llama_print_timings: prompt eval time =    9430.17 ms /    82 tokens (  115.00
ms per token,     8.70 tokens per second)
```

```
llama_print_timings:        eval time =    4528.93 ms /    21 runs   (  215.66
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14093.16 ms /   103 tokens
No. of rows:  89%| | 1119/1258 [14:03:39<40:55, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.15 ms /    26 runs   (    0.39
ms per token,  2561.83 tokens per second)
llama_print_timings: prompt eval time =   10644.65 ms /    96 tokens (  110.88
ms per token,     9.02 tokens per second)
llama_print_timings:        eval time =    5760.14 ms /    25 runs   (  230.41
ms per token,     4.34 tokens per second)
llama_print_timings:       total time =   16563.96 ms /   121 tokens
No. of rows:  89%| | 1120/1258 [14:03:55<39:52, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.29 ms /    27 runs   (    0.38
ms per token,  2623.40 tokens per second)
llama_print_timings: prompt eval time =    9844.57 ms /    86 tokens (  114.47
ms per token,     8.74 tokens per second)
llama_print_timings:        eval time =    5639.81 ms /    26 runs   (  216.92
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   15650.24 ms /   112 tokens
No. of rows:  89%| | 1121/1258 [14:04:11<38:25, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.67 ms /    28 runs   (    0.38
ms per token,  2623.69 tokens per second)
llama_print_timings: prompt eval time =   12525.08 ms /    99 tokens (  126.52
ms per token,     7.90 tokens per second)
llama_print_timings:        eval time =    6138.98 ms /    27 runs   (  227.37
ms per token,     4.40 tokens per second)
llama_print_timings:       total time =   18842.40 ms /   126 tokens
No. of rows:  89%| | 1122/1258 [14:04:30<39:31, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.17 ms /    21 runs   (    0.39
ms per token,  2570.69 tokens per second)
llama_print_timings: prompt eval time =    9596.16 ms /    84 tokens (  114.24
ms per token,     8.75 tokens per second)
llama_print_timings:        eval time =    4312.21 ms /    20 runs   (  215.61
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14037.52 ms /   104 tokens
No. of rows:  89%| | 1123/1258 [14:04:44<36:56, 16Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.16 ms /    21 runs   (    0.39
ms per token,  2572.90 tokens per second)
llama_print_timings: prompt eval time =    9475.11 ms /    82 tokens (  115.55
ms per token,     8.65 tokens per second)
llama_print_timings:        eval time =    4322.56 ms /    20 runs   (  216.13
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   13927.28 ms /   102 tokens
No. of rows:  89%| | 1124/1258 [14:04:58<35:00, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.15 ms /    27 runs   (    0.38
ms per token,  2659.84 tokens per second)
llama_print_timings: prompt eval time =   10215.01 ms /    89 tokens (  114.78
ms per token,     8.71 tokens per second)
llama_print_timings:        eval time =    5598.66 ms /    26 runs   (  215.33
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   15976.81 ms /   115 tokens
No. of rows:  89%| | 1125/1258 [14:05:14<34:57, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.23 ms /    24 runs   (    0.38
ms per token,  2600.22 tokens per second)
llama_print_timings: prompt eval time =   11533.36 ms /    92 tokens (  125.36
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =    5080.73 ms /    23 runs   (  220.90
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   16759.23 ms /   115 tokens
No. of rows:  90%| | 1126/1258 [14:05:31<35:20, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      16.26 ms /    41 runs   (    0.40
ms per token,  2521.22 tokens per second)
llama_print_timings: prompt eval time =   12767.88 ms /   102 tokens (  125.18
ms per token,     7.99 tokens per second)
llama_print_timings:        eval time =    8667.53 ms /    40 runs   (  216.69
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   21692.72 ms /   142 tokens
No. of rows:  90%| | 1127/1258 [14:05:52<38:46, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.00 ms /    20 runs   (    0.40
```

```
ms per token,  2501.56 tokens per second)
llama_print_timings: prompt eval time =   11562.52 ms /    92 tokens (  125.68
ms per token,     7.96 tokens per second)
llama_print_timings:        eval time =    4114.03 ms /    19 runs   (  216.53
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   15798.19 ms /   111 tokens
No. of rows:  90%| | 1128/1258 [14:06:08<37:12, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.97 ms /    26 runs   (    0.38
ms per token,  2606.52 tokens per second)
llama_print_timings: prompt eval time =   11758.01 ms /    95 tokens (  123.77
ms per token,     8.08 tokens per second)
llama_print_timings:        eval time =    5576.24 ms /    25 runs   (  223.05
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   17501.18 ms /   120 tokens
No. of rows:  90%| | 1129/1258 [14:06:26<37:08, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.30 ms /    35 runs   (    0.41
ms per token,  2447.55 tokens per second)
llama_print_timings: prompt eval time =   13230.23 ms /   106 tokens (  124.81
ms per token,     8.01 tokens per second)
llama_print_timings:        eval time =    7526.31 ms /    34 runs   (  221.36
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   20978.25 ms /   140 tokens
No. of rows:  90%| | 1130/1258 [14:06:47<39:13, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.47 ms /    29 runs   (    0.40
ms per token,  2527.67 tokens per second)
llama_print_timings: prompt eval time =   11912.48 ms /   106 tokens (  112.38
ms per token,     8.90 tokens per second)
llama_print_timings:        eval time =    6114.27 ms /    28 runs   (  218.37
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   18209.30 ms /   134 tokens
No. of rows:  90%| | 1131/1258 [14:07:05<38:48, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.65 ms /    26 runs   (    0.41
ms per token,  2441.31 tokens per second)
llama_print_timings: prompt eval time =   10203.53 ms /    91 tokens (  112.13
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    5450.42 ms /    25 runs   (  218.02
```

```
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   15818.61 ms /   116 tokens
No. of rows:  90%| | 1132/1258 [14:07:21<36:55, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.04 ms /    20 runs   (    0.40
ms per token,  2487.87 tokens per second)
llama_print_timings: prompt eval time =    9882.83 ms /    83 tokens (  119.07
ms per token,     8.40 tokens per second)
llama_print_timings:        eval time =    4108.86 ms /    19 runs   (  216.26
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   14120.27 ms /   102 tokens
No. of rows:  90%| | 1133/1258 [14:07:35<34:28, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.83 ms /    26 runs   (    0.38
ms per token,  2645.77 tokens per second)
llama_print_timings: prompt eval time =   13175.41 ms /   107 tokens (  123.13
ms per token,     8.12 tokens per second)
llama_print_timings:        eval time =    5777.95 ms /    25 runs   (  231.12
ms per token,     4.33 tokens per second)
llama_print_timings:       total time =   19120.07 ms /   132 tokens
No. of rows:  90%| | 1134/1258 [14:07:54<35:47, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.04 ms /    27 runs   (    0.37
ms per token,  2690.05 tokens per second)
llama_print_timings: prompt eval time =   10895.48 ms /    90 tokens (  121.06
ms per token,     8.26 tokens per second)
llama_print_timings:        eval time =    7354.66 ms /    26 runs   (  282.87
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   18415.69 ms /   116 tokens
No. of rows:  90%| | 1135/1258 [14:08:12<36:11, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.16 ms /    20 runs   (    0.41
ms per token,  2451.88 tokens per second)
llama_print_timings: prompt eval time =    9625.98 ms /    84 tokens (  114.59
ms per token,     8.73 tokens per second)
llama_print_timings:        eval time =    4274.77 ms /    19 runs   (  224.99
ms per token,     4.44 tokens per second)
llama_print_timings:       total time =   14025.50 ms /   103 tokens
No. of rows:  90%| | 1136/1258 [14:08:26<33:41, 16Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.04 ms /    30 runs   (    0.40
ms per token,  2492.52 tokens per second)
llama_print_timings: prompt eval time =   12792.35 ms /   112 tokens (  114.22
ms per token,     8.76 tokens per second)
llama_print_timings:        eval time =    6321.56 ms /    29 runs   (  217.98
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   19300.73 ms /   141 tokens
No. of rows:  90%| | 1137/1258 [14:08:46<35:04, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.91 ms /    25 runs   (    0.40
ms per token,  2522.45 tokens per second)
llama_print_timings: prompt eval time =   10871.14 ms /    96 tokens (  113.24
ms per token,     8.83 tokens per second)
llama_print_timings:        eval time =    5212.62 ms /    24 runs   (  217.19
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   16240.02 ms /   120 tokens
No. of rows:  90%| | 1138/1258 [14:09:02<34:05, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.88 ms /    29 runs   (    0.38
ms per token,  2664.71 tokens per second)
llama_print_timings: prompt eval time =   13034.40 ms /   108 tokens (  120.69
ms per token,     8.29 tokens per second)
llama_print_timings:        eval time =    7776.44 ms /    28 runs   (  277.73
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   20989.26 ms /   136 tokens
No. of rows:  91%| | 1139/1258 [14:09:23<36:09, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      22.36 ms /    50 runs   (    0.45
ms per token,  2236.44 tokens per second)
llama_print_timings: prompt eval time =   16294.95 ms /   146 tokens (  111.61
ms per token,     8.96 tokens per second)
llama_print_timings:        eval time =   10619.82 ms /    49 runs   (  216.73
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   27232.10 ms /   195 tokens
No. of rows:  91%| | 1140/1258 [14:09:50<41:10, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.58 ms /    38 runs   (    0.38
ms per token,  2607.03 tokens per second)
```

```
llama_print_timings: prompt eval time =   12832.13 ms /   112 tokens (  114.57
ms per token,     8.73 tokens per second)
llama_print_timings:        eval time =    7991.75 ms /    37 runs   (  215.99
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   21056.98 ms /   149 tokens
No. of rows:  91%| | 1141/1258 [14:10:11<40:53, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.26 ms /    27 runs   (    0.42
ms per token,  2398.72 tokens per second)
llama_print_timings: prompt eval time =   12336.13 ms /    97 tokens (  127.18
ms per token,     7.86 tokens per second)
llama_print_timings:        eval time =    6223.33 ms /    26 runs   (  239.36
ms per token,     4.18 tokens per second)
llama_print_timings:       total time =   18742.95 ms /   123 tokens
No. of rows:  91%| | 1142/1258 [14:10:30<39:15, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.59 ms /    21 runs   (    0.41
ms per token,  2444.13 tokens per second)
llama_print_timings: prompt eval time =   11694.39 ms /    92 tokens (  127.11
ms per token,     7.87 tokens per second)
llama_print_timings:        eval time =    4419.89 ms /    20 runs   (  220.99
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   16246.73 ms /   112 tokens
No. of rows:  91%| | 1143/1258 [14:10:46<36:35, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.17 ms /    23 runs   (    0.40
ms per token,  2509.27 tokens per second)
llama_print_timings: prompt eval time =   10193.25 ms /    88 tokens (  115.83
ms per token,     8.63 tokens per second)
llama_print_timings:        eval time =    4805.94 ms /    22 runs   (  218.45
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   15139.29 ms /   110 tokens
No. of rows:  91%| | 1144/1258 [14:11:01<34:01, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.68 ms /    22 runs   (    0.39
ms per token,  2533.69 tokens per second)
llama_print_timings: prompt eval time =   10075.17 ms /    88 tokens (  114.49
ms per token,     8.73 tokens per second)
llama_print_timings:        eval time =    4531.07 ms /    21 runs   (  215.77
ms per token,     4.63 tokens per second)
```

```
llama_print_timings:       total time =   14739.90 ms /   109 tokens
No. of rows:  91%| | 1145/1258 [14:11:16<31:56, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.95 ms /    11 runs   (    0.63
ms per token,  1582.05 tokens per second)
llama_print_timings: prompt eval time =    8871.80 ms /    66 tokens (  134.42
ms per token,     7.44 tokens per second)
llama_print_timings:        eval time =    2144.93 ms /    10 runs   (  214.49
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   11086.74 ms /    76 tokens
No. of rows:  91%| | 1146/1258 [14:11:27<28:22, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.35 ms /    32 runs   (    0.39
ms per token,  2592.14 tokens per second)
llama_print_timings: prompt eval time =   15474.59 ms /   138 tokens (  112.13
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    6721.68 ms /    31 runs   (  216.83
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   22396.08 ms /   169 tokens
No. of rows:  91%| | 1147/1258 [14:11:50<32:07, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.17 ms /    21 runs   (    0.39
ms per token,  2569.44 tokens per second)
llama_print_timings: prompt eval time =    9677.05 ms /    85 tokens (  113.85
ms per token,     8.78 tokens per second)
llama_print_timings:        eval time =    4393.06 ms /    20 runs   (  219.65
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   14199.12 ms /   105 tokens
No. of rows:  91%| | 1148/1258 [14:12:04<30:05, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.01 ms /    23 runs   (    0.39
ms per token,  2552.44 tokens per second)
llama_print_timings: prompt eval time =   11541.76 ms /    90 tokens (  128.24
ms per token,     7.80 tokens per second)
llama_print_timings:        eval time =    4754.89 ms /    22 runs   (  216.13
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   16439.86 ms /   112 tokens
No. of rows:  91%| | 1149/1258 [14:12:20<29:50, 16Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.06 ms /    22 runs   (    0.41
ms per token,  2427.18 tokens per second)
llama_print_timings: prompt eval time =   11383.97 ms /    83 tokens (  137.16
ms per token,     7.29 tokens per second)
llama_print_timings:        eval time =    4878.63 ms /    21 runs   (  232.32
ms per token,     4.30 tokens per second)
llama_print_timings:       total time =   16412.77 ms /   104 tokens
No. of rows:  91%| | 1150/1258 [14:12:37<29:33, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.98 ms /    39 runs   (    0.46
ms per token,  2168.84 tokens per second)
llama_print_timings: prompt eval time =   13555.80 ms /   119 tokens (  113.91
ms per token,     8.78 tokens per second)
llama_print_timings:        eval time =    8267.15 ms /    38 runs   (  217.56
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   22070.36 ms /   157 tokens
No. of rows:  91%| | 1151/1258 [14:12:59<32:18, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.64 ms /    24 runs   (    0.40
ms per token,  2490.14 tokens per second)
llama_print_timings: prompt eval time =   11481.08 ms /   101 tokens (  113.67
ms per token,     8.80 tokens per second)
llama_print_timings:        eval time =    4966.08 ms /    23 runs   (  215.92
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   16595.59 ms /   124 tokens
No. of rows:  92%| | 1152/1258 [14:13:15<31:12, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.00 ms /    25 runs   (    0.40
ms per token,  2499.25 tokens per second)
llama_print_timings: prompt eval time =    9649.12 ms /    83 tokens (  116.25
ms per token,     8.60 tokens per second)
llama_print_timings:        eval time =    6835.53 ms /    24 runs   (  284.81
ms per token,     3.51 tokens per second)
llama_print_timings:       total time =   16648.19 ms /   107 tokens
No. of rows:  92%| | 1153/1258 [14:13:32<30:22, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.98 ms /    29 runs   (    0.38
ms per token,  2641.41 tokens per second)
llama_print_timings: prompt eval time =   10094.89 ms /    88 tokens (  114.71
```

```
ms per token,     8.72 tokens per second)
llama_print_timings:        eval time =    7763.78 ms /    28 runs   (  277.28
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   18041.24 ms /   116 tokens
No. of rows:  92%| | 1154/1258 [14:13:50<30:27, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.61 ms /    47 runs   (    0.40
ms per token,  2525.52 tokens per second)
llama_print_timings: prompt eval time =   14258.75 ms /   125 tokens (  114.07
ms per token,     8.77 tokens per second)
llama_print_timings:        eval time =    9956.85 ms /    46 runs   (  216.45
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   24511.19 ms /   171 tokens
No. of rows:  92%| | 1155/1258 [14:14:15<33:44, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.37 ms /    39 runs   (    0.39
ms per token,  2537.25 tokens per second)
llama_print_timings: prompt eval time =   15483.00 ms /   125 tokens (  123.86
ms per token,     8.07 tokens per second)
llama_print_timings:        eval time =    8545.37 ms /    38 runs   (  224.88
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   24283.04 ms /   163 tokens
No. of rows:  92%| | 1156/1258 [14:14:39<35:46, 21Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.50 ms /    21 runs   (    0.40
ms per token,  2471.75 tokens per second)
llama_print_timings: prompt eval time =    9024.84 ms /    78 tokens (  115.70
ms per token,     8.64 tokens per second)
llama_print_timings:        eval time =    4296.89 ms /    20 runs   (  214.84
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   13448.55 ms /    98 tokens
No. of rows:  92%| | 1157/1258 [14:14:52<31:35, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.67 ms /    39 runs   (    0.38
ms per token,  2658.49 tokens per second)
llama_print_timings: prompt eval time =   13226.65 ms /   115 tokens (  115.01
ms per token,     8.69 tokens per second)
llama_print_timings:        eval time =    8272.86 ms /    38 runs   (  217.71
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   21740.68 ms /   153 tokens
```

```
No. of rows:  92%| | 1158/1258 [14:15:14<32:46, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.52 ms /    30 runs   (    0.38
ms per token,  2604.85 tokens per second)
llama_print_timings: prompt eval time =   11699.21 ms /   101 tokens (  115.83
ms per token,     8.63 tokens per second)
llama_print_timings:        eval time =    6278.25 ms /    29 runs   (  216.49
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   18165.95 ms /   130 tokens
No. of rows:  92%| | 1159/1258 [14:15:32<31:42, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.00 ms /    28 runs   (    0.39
ms per token,  2544.99 tokens per second)
llama_print_timings: prompt eval time =   12461.15 ms /   111 tokens (  112.26
ms per token,     8.91 tokens per second)
llama_print_timings:        eval time =    7749.78 ms /    27 runs   (  287.03
ms per token,     3.48 tokens per second)
llama_print_timings:       total time =   20384.89 ms /   138 tokens
No. of rows:  92%| | 1160/1258 [14:15:53<31:57, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.93 ms /    21 runs   (    0.38
ms per token,  2648.17 tokens per second)
llama_print_timings: prompt eval time =    8835.40 ms /    76 tokens (  116.26
ms per token,     8.60 tokens per second)
llama_print_timings:        eval time =    4314.56 ms /    20 runs   (  215.73
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   13279.22 ms /    96 tokens
No. of rows:  92%| | 1161/1258 [14:16:06<28:35, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.55 ms /    22 runs   (    0.39
ms per token,  2571.90 tokens per second)
llama_print_timings: prompt eval time =    9897.59 ms /    86 tokens (  115.09
ms per token,     8.69 tokens per second)
llama_print_timings:        eval time =    4666.91 ms /    21 runs   (  222.23
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   14702.19 ms /   107 tokens
No. of rows:  92%| | 1162/1258 [14:16:21<26:52, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =      10.61 ms /    26 runs   (    0.41
ms per token,  2449.59 tokens per second)
llama_print_timings: prompt eval time =   10927.12 ms /    95 tokens (  115.02
ms per token,     8.69 tokens per second)
llama_print_timings:        eval time =    5419.39 ms /    25 runs   (  216.78
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   16510.67 ms /   120 tokens
No. of rows:  92%| | 1163/1258 [14:16:37<26:27, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.02 ms /    22 runs   (    0.36
ms per token,  2743.14 tokens per second)
llama_print_timings: prompt eval time =   11150.81 ms /    94 tokens (  118.63
ms per token,     8.43 tokens per second)
llama_print_timings:        eval time =    4509.68 ms /    21 runs   (  214.75
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   15794.75 ms /   115 tokens
No. of rows:  93%| | 1164/1258 [14:16:53<25:45, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.47 ms /    25 runs   (    0.38
ms per token,  2639.36 tokens per second)
llama_print_timings: prompt eval time =    9539.26 ms /    83 tokens (  114.93
ms per token,     8.70 tokens per second)
llama_print_timings:        eval time =    5163.75 ms /    24 runs   (  215.16
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   14857.46 ms /   107 tokens
No. of rows:  93%| | 1165/1258 [14:17:08<24:45, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.06 ms /    23 runs   (    0.39
ms per token,  2538.91 tokens per second)
llama_print_timings: prompt eval time =   11583.60 ms /   102 tokens (  113.56
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    5207.80 ms /    22 runs   (  236.72
ms per token,     4.22 tokens per second)
llama_print_timings:       total time =   16939.71 ms /   124 tokens
No. of rows:  93%| | 1166/1258 [14:17:25<24:56, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.47 ms /    20 runs   (    0.37
ms per token,  2678.09 tokens per second)
llama_print_timings: prompt eval time =    9736.90 ms /    86 tokens (  113.22
ms per token,     8.83 tokens per second)
```

```
llama_print_timings:        eval time =    4104.89 ms /    19 runs   (  216.05
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   13963.80 ms /   105 tokens
No. of rows:  93%| | 1167/1258 [14:17:39<23:37, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.81 ms /    25 runs   (    0.39
ms per token,  2547.64 tokens per second)
llama_print_timings: prompt eval time =   12346.63 ms /    99 tokens (  124.71
ms per token,     8.02 tokens per second)
llama_print_timings:        eval time =    5201.64 ms /    24 runs   (  216.73
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   17700.84 ms /   123 tokens
No. of rows:  93%| | 1168/1258 [14:17:57<24:19, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.89 ms /    47 runs   (    0.40
ms per token,  2487.43 tokens per second)
llama_print_timings: prompt eval time =   16421.05 ms /   137 tokens (  119.86
ms per token,     8.34 tokens per second)
llama_print_timings:        eval time =   11668.38 ms /    46 runs   (  253.66
ms per token,     3.94 tokens per second)
llama_print_timings:       total time =   28388.44 ms /   183 tokens
No. of rows:  93%| | 1169/1258 [14:18:25<29:28, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.06 ms /    29 runs   (    0.42
ms per token,  2404.24 tokens per second)
llama_print_timings: prompt eval time =   12183.94 ms /   107 tokens (  113.87
ms per token,     8.78 tokens per second)
llama_print_timings:        eval time =    8118.71 ms /    28 runs   (  289.95
ms per token,     3.45 tokens per second)
llama_print_timings:       total time =   20487.68 ms /   135 tokens
No. of rows:  93%| | 1170/1258 [14:18:45<29:24, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.45 ms /    25 runs   (    0.42
ms per token,  2393.26 tokens per second)
llama_print_timings: prompt eval time =   10977.66 ms /    97 tokens (  113.17
ms per token,     8.84 tokens per second)
llama_print_timings:        eval time =    5362.58 ms /    24 runs   (  223.44
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   16499.99 ms /   121 tokens
No. of rows:  93%| | 1171/1258 [14:19:02<27:32, 18Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.71 ms /    28 runs   (    0.38
ms per token,  2615.36 tokens per second)
llama_print_timings: prompt eval time =   10295.31 ms /    89 tokens (  115.68
ms per token,     8.64 tokens per second)
llama_print_timings:        eval time =    5863.33 ms /    27 runs   (  217.16
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   16326.69 ms /   116 tokens
No. of rows:  93%| | 1172/1258 [14:19:18<26:04, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.99 ms /    21 runs   (    0.38
ms per token,  2629.93 tokens per second)
llama_print_timings: prompt eval time =   10336.08 ms /    80 tokens (  129.20
ms per token,     7.74 tokens per second)
llama_print_timings:        eval time =    4300.02 ms /    20 runs   (  215.00
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   14764.02 ms /   100 tokens
No. of rows:  93%| | 1173/1258 [14:19:33<24:19, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.16 ms /    36 runs   (    0.39
ms per token,  2542.19 tokens per second)
llama_print_timings: prompt eval time =   13480.31 ms /   117 tokens (  115.22
ms per token,     8.68 tokens per second)
llama_print_timings:        eval time =    7570.32 ms /    35 runs   (  216.29
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   21276.88 ms /   152 tokens
No. of rows:  93%| | 1174/1258 [14:19:54<25:46, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.80 ms /    36 runs   (    0.38
ms per token,  2609.26 tokens per second)
llama_print_timings: prompt eval time =   13124.71 ms /   104 tokens (  126.20
ms per token,     7.92 tokens per second)
llama_print_timings:        eval time =    7555.43 ms /    35 runs   (  215.87
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   20903.54 ms /   139 tokens
No. of rows:  93%| | 1175/1258 [14:20:15<26:29, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.28 ms /    20 runs   (    0.41
```

```
ms per token,  2415.75 tokens per second)
llama_print_timings: prompt eval time =   11080.34 ms /    86 tokens (  128.84
ms per token,     7.76 tokens per second)
llama_print_timings:        eval time =    4269.32 ms /    19 runs   (  224.70
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   15480.81 ms /   105 tokens
No. of rows:  93%| | 1176/1258 [14:20:31<24:40, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.04 ms /    23 runs   (    0.44
ms per token,  2290.38 tokens per second)
llama_print_timings: prompt eval time =    9385.90 ms /    81 tokens (  115.88
ms per token,     8.63 tokens per second)
llama_print_timings:        eval time =    4943.45 ms /    22 runs   (  224.70
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   14482.85 ms /   103 tokens
No. of rows:  94%| | 1177/1258 [14:20:45<22:55, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.45 ms /    19 runs   (    0.39
ms per token,  2550.68 tokens per second)
llama_print_timings: prompt eval time =    9600.21 ms /    83 tokens (  115.67
ms per token,     8.65 tokens per second)
llama_print_timings:        eval time =    3944.15 ms /    18 runs   (  219.12
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   13660.96 ms /   101 tokens
No. of rows:  94%| | 1178/1258 [14:20:59<21:19, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.51 ms /    25 runs   (    0.38
ms per token,  2629.64 tokens per second)
llama_print_timings: prompt eval time =    9942.37 ms /    85 tokens (  116.97
ms per token,     8.55 tokens per second)
llama_print_timings:        eval time =    5152.40 ms /    24 runs   (  214.68
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   15250.80 ms /   109 tokens
No. of rows:  94%| | 1179/1258 [14:21:14<20:45, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.78 ms /    22 runs   (    0.40
ms per token,  2504.84 tokens per second)
llama_print_timings: prompt eval time =   10680.47 ms /    94 tokens (  113.62
ms per token,     8.80 tokens per second)
llama_print_timings:        eval time =    4556.03 ms /    21 runs   (  216.95
```

```
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   15375.04 ms /   115 tokens
No. of rows:  94%| | 1180/1258 [14:21:30<20:20, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.04 ms /    26 runs   (    0.50
ms per token,  1993.10 tokens per second)
llama_print_timings: prompt eval time =   10560.11 ms /    93 tokens (  113.55
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    5502.73 ms /    25 runs   (  220.11
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   16228.66 ms /   118 tokens
No. of rows:  94%| | 1181/1258 [14:21:46<20:18, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.60 ms /    31 runs   (    0.37
ms per token,  2671.72 tokens per second)
llama_print_timings: prompt eval time =   12162.59 ms /    97 tokens (  125.39
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =    8158.63 ms /    30 runs   (  271.95
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   20514.66 ms /   127 tokens
No. of rows:  94%| | 1182/1258 [14:22:06<21:49, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.32 ms /    23 runs   (    0.41
ms per token,  2467.28 tokens per second)
llama_print_timings: prompt eval time =   11997.24 ms /   107 tokens (  112.12
ms per token,     8.92 tokens per second)
llama_print_timings:        eval time =    4833.07 ms /    22 runs   (  219.69
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   16977.37 ms /   129 tokens
No. of rows:  94%| | 1183/1258 [14:22:23<21:27, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.49 ms /    20 runs   (    0.37
ms per token,  2671.30 tokens per second)
llama_print_timings: prompt eval time =   10081.02 ms /    86 tokens (  117.22
ms per token,     8.53 tokens per second)
llama_print_timings:        eval time =    4147.55 ms /    19 runs   (  218.29
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   14350.64 ms /   105 tokens
No. of rows:  94%| | 1184/1258 [14:22:38<20:07, 16Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.09 ms /    24 runs   (    0.38
ms per token,  2640.26 tokens per second)
llama_print_timings: prompt eval time =   11104.30 ms /    94 tokens (  118.13
ms per token,     8.47 tokens per second)
llama_print_timings:        eval time =    4970.05 ms /    23 runs   (  216.09
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   16221.97 ms /   117 tokens
No. of rows:  94%| | 1185/1258 [14:22:54<19:49, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.06 ms /    29 runs   (    0.38
ms per token,  2622.06 tokens per second)
llama_print_timings: prompt eval time =   11618.24 ms /   103 tokens (  112.80
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    7744.34 ms /    28 runs   (  276.58
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   19542.12 ms /   131 tokens
No. of rows:  94%| | 1186/1258 [14:23:13<20:43, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.57 ms /    29 runs   (    0.40
ms per token,  2506.27 tokens per second)
llama_print_timings: prompt eval time =   10731.03 ms /    95 tokens (  112.96
ms per token,     8.85 tokens per second)
llama_print_timings:        eval time =    6023.94 ms /    28 runs   (  215.14
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   16938.47 ms /   123 tokens
No. of rows:  94%| | 1187/1258 [14:23:30<20:19, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.83 ms /    36 runs   (    0.38
ms per token,  2603.98 tokens per second)
llama_print_timings: prompt eval time =   13528.85 ms /   108 tokens (  125.27
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =    7533.15 ms /    35 runs   (  215.23
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   21285.71 ms /   143 tokens
No. of rows:  94%| | 1188/1258 [14:23:52<21:28, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.06 ms /    33 runs   (    0.40
ms per token,  2526.03 tokens per second)
```

```
llama_print_timings: prompt eval time =   13487.38 ms /   110 tokens (  122.61
ms per token,     8.16 tokens per second)
llama_print_timings:        eval time =    8702.76 ms /    32 runs   (  271.96
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   22389.72 ms /   142 tokens
No. of rows:  95%| | 1189/1258 [14:24:14<22:32, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      14.87 ms /    30 runs   (    0.50
ms per token,  2017.48 tokens per second)
llama_print_timings: prompt eval time =   11038.00 ms /    97 tokens (  113.79
ms per token,     8.79 tokens per second)
llama_print_timings:        eval time =    6475.51 ms /    29 runs   (  223.29
ms per token,     4.48 tokens per second)
llama_print_timings:       total time =   17707.73 ms /   126 tokens
No. of rows:  95%| | 1190/1258 [14:24:32<21:34, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.88 ms /    26 runs   (    0.42
ms per token,  2389.93 tokens per second)
llama_print_timings: prompt eval time =   11322.34 ms /    89 tokens (  127.22
ms per token,     7.86 tokens per second)
llama_print_timings:        eval time =    5602.79 ms /    25 runs   (  224.11
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   17091.88 ms /   114 tokens
No. of rows:  95%| | 1191/1258 [14:24:49<20:36, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.47 ms /    26 runs   (    0.40
ms per token,  2483.52 tokens per second)
llama_print_timings: prompt eval time =   10978.06 ms /    96 tokens (  114.35
ms per token,     8.74 tokens per second)
llama_print_timings:        eval time =    5409.20 ms /    25 runs   (  216.37
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   16547.10 ms /   121 tokens
No. of rows:  95%| | 1192/1258 [14:25:05<19:40, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.54 ms /    22 runs   (    0.39
ms per token,  2574.61 tokens per second)
llama_print_timings: prompt eval time =   10204.08 ms /    90 tokens (  113.38
ms per token,     8.82 tokens per second)
llama_print_timings:        eval time =    4494.84 ms /    21 runs   (  214.04
ms per token,     4.67 tokens per second)
```

```
llama_print_timings:       total time =   14832.27 ms /   111 tokens
No. of rows:  95%| | 1193/1258 [14:25:20<18:23, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.35 ms /    26 runs   (    0.40
ms per token,  2512.32 tokens per second)
llama_print_timings: prompt eval time =    9610.94 ms /    83 tokens (  115.79
ms per token,     8.64 tokens per second)
llama_print_timings:        eval time =    5409.60 ms /    25 runs   (  216.38
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   15184.52 ms /   108 tokens
No. of rows:  95%| | 1194/1258 [14:25:35<17:31, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.41 ms /    31 runs   (    0.37
ms per token,  2716.68 tokens per second)
llama_print_timings: prompt eval time =   11778.95 ms /    94 tokens (  125.31
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =    6466.72 ms /    30 runs   (  215.56
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   18434.52 ms /   124 tokens
No. of rows:  95%| | 1195/1258 [14:25:54<17:53, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.30 ms /    34 runs   (    0.39
ms per token,  2556.97 tokens per second)
llama_print_timings: prompt eval time =   13796.21 ms /   110 tokens (  125.42
ms per token,     7.97 tokens per second)
llama_print_timings:        eval time =    7101.93 ms /    33 runs   (  215.21
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   21107.30 ms /   143 tokens
No. of rows:  95%| | 1196/1258 [14:26:15<18:52, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.76 ms /    23 runs   (    0.38
ms per token,  2625.57 tokens per second)
llama_print_timings: prompt eval time =   11395.86 ms /    98 tokens (  116.28
ms per token,     8.60 tokens per second)
llama_print_timings:        eval time =    4900.68 ms /    22 runs   (  222.76
ms per token,     4.49 tokens per second)
llama_print_timings:       total time =   16438.48 ms /   120 tokens
No. of rows:  95%| | 1197/1258 [14:26:31<18:00, 17Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.74 ms /    25 runs   (    0.39
ms per token,  2567.79 tokens per second)
llama_print_timings: prompt eval time =   10299.01 ms /    90 tokens (  114.43
ms per token,     8.74 tokens per second)
llama_print_timings:        eval time =    5161.96 ms /    24 runs   (  215.08
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   15613.71 ms /   114 tokens
No. of rows:  95%| | 1198/1258 [14:26:47<17:05, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.15 ms /    27 runs   (    0.38
ms per token,  2660.36 tokens per second)
llama_print_timings: prompt eval time =   12005.15 ms /    93 tokens (  129.09
ms per token,     7.75 tokens per second)
llama_print_timings:        eval time =    5653.39 ms /    26 runs   (  217.44
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   17825.06 ms /   119 tokens
No. of rows:  95%| | 1199/1258 [14:27:05<17:01, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.50 ms /    26 runs   (    0.40
ms per token,  2475.25 tokens per second)
llama_print_timings: prompt eval time =   11275.24 ms /   100 tokens (  112.75
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    5365.26 ms /    25 runs   (  214.61
ms per token,     4.66 tokens per second)
llama_print_timings:       total time =   16803.40 ms /   125 tokens
No. of rows:  95%| | 1200/1258 [14:27:22<16:35, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.12 ms /    27 runs   (    0.37
ms per token,  2667.19 tokens per second)
llama_print_timings: prompt eval time =   10168.81 ms /    88 tokens (  115.55
ms per token,     8.65 tokens per second)
llama_print_timings:        eval time =    5619.41 ms /    26 runs   (  216.13
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   15960.69 ms /   114 tokens
No. of rows:  95%| | 1201/1258 [14:27:38<15:57, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.20 ms /    22 runs   (    0.37
ms per token,  2682.27 tokens per second)
llama_print_timings: prompt eval time =   10560.42 ms /    92 tokens (  114.79
```

```
ms per token,     8.71 tokens per second)
llama_print_timings:        eval time =    4580.80 ms /    21 runs   (  218.13
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   15278.55 ms /   113 tokens
No. of rows:  96%| | 1202/1258 [14:27:53<15:15, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.40 ms /    50 runs   (    0.39
ms per token,  2577.05 tokens per second)
llama_print_timings: prompt eval time =   13732.54 ms /   113 tokens (  121.53
ms per token,     8.23 tokens per second)
llama_print_timings:        eval time =   10540.16 ms /    49 runs   (  215.11
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   24578.30 ms /   162 tokens
No. of rows:  96%| | 1203/1258 [14:28:18<17:15, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.32 ms /    27 runs   (    0.38
ms per token,  2615.27 tokens per second)
llama_print_timings: prompt eval time =   11504.81 ms /   100 tokens (  115.05
ms per token,     8.69 tokens per second)
llama_print_timings:        eval time =    5657.98 ms /    26 runs   (  217.61
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   17328.97 ms /   126 tokens
No. of rows:  96%| | 1204/1258 [14:28:35<16:32, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.08 ms /    31 runs   (    0.39
ms per token,  2566.01 tokens per second)
llama_print_timings: prompt eval time =   11602.71 ms /   103 tokens (  112.65
ms per token,     8.88 tokens per second)
llama_print_timings:        eval time =    6716.71 ms /    30 runs   (  223.89
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   18514.16 ms /   133 tokens
No. of rows:  96%| | 1205/1258 [14:28:53<16:16, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.93 ms /    35 runs   (    0.37
ms per token,  2706.67 tokens per second)
llama_print_timings: prompt eval time =   11320.80 ms /    98 tokens (  115.52
ms per token,     8.66 tokens per second)
llama_print_timings:        eval time =    7480.05 ms /    34 runs   (  220.00
ms per token,     4.55 tokens per second)
llama_print_timings:       total time =   19012.52 ms /   132 tokens
```

```
No. of rows:  96%| | 1206/1258 [14:29:12<16:07, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.91 ms /    28 runs   (    0.39
ms per token,  2566.45 tokens per second)
llama_print_timings: prompt eval time =   12293.09 ms /   110 tokens (  111.76
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    7602.09 ms /    27 runs   (  281.56
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   20067.41 ms /   137 tokens
No. of rows:  96%| | 1207/1258 [14:29:32<16:11, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.90 ms /    23 runs   (    0.39
ms per token,  2583.40 tokens per second)
llama_print_timings: prompt eval time =    9762.96 ms /    86 tokens (  113.52
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    4742.69 ms /    22 runs   (  215.58
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14647.77 ms /   108 tokens
No. of rows:  96%| | 1208/1258 [14:29:47<14:46, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      22.30 ms /    50 runs   (    0.45
ms per token,  2242.45 tokens per second)
llama_print_timings: prompt eval time =   15280.47 ms /   123 tokens (  124.23
ms per token,     8.05 tokens per second)
llama_print_timings:        eval time =   10585.16 ms /    49 runs   (  216.02
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   26177.55 ms /   172 tokens
No. of rows:  96%| | 1209/1258 [14:30:13<16:32, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.20 ms /    24 runs   (    0.38
ms per token,  2608.41 tokens per second)
llama_print_timings: prompt eval time =   12122.32 ms /    95 tokens (  127.60
ms per token,     7.84 tokens per second)
llama_print_timings:        eval time =    5154.67 ms /    23 runs   (  224.12
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   17426.77 ms /   118 tokens
No. of rows:  96%| | 1210/1258 [14:30:31<15:31, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
```

```
llama_print_timings:      sample time =       9.06 ms /    24 runs   (    0.38
ms per token,  2649.59 tokens per second)
llama_print_timings: prompt eval time =   10115.40 ms /    88 tokens (  114.95
ms per token,     8.70 tokens per second)
llama_print_timings:        eval time =    4984.46 ms /    23 runs   (  216.72
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   15246.17 ms /   111 tokens
No. of rows:  96%| | 1211/1258 [14:30:46<14:13, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.05 ms /    22 runs   (    0.41
ms per token,  2429.60 tokens per second)
llama_print_timings: prompt eval time =   13414.46 ms /   107 tokens (  125.37
ms per token,     7.98 tokens per second)
llama_print_timings:        eval time =    4791.22 ms /    21 runs   (  228.15
ms per token,     4.38 tokens per second)
llama_print_timings:       total time =   18348.48 ms /   128 tokens
No. of rows:  96%| | 1212/1258 [14:31:04<13:58, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.73 ms /    31 runs   (    0.38
ms per token,  2642.57 tokens per second)
llama_print_timings: prompt eval time =   11843.05 ms /   105 tokens (  112.79
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    6612.02 ms /    30 runs   (  220.40
ms per token,     4.54 tokens per second)
llama_print_timings:       total time =   18650.33 ms /   135 tokens
No. of rows:  96%| | 1213/1258 [14:31:23<13:46, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.74 ms /    30 runs   (    0.39
ms per token,  2554.71 tokens per second)
llama_print_timings: prompt eval time =   11782.27 ms /   100 tokens (  117.82
ms per token,     8.49 tokens per second)
llama_print_timings:        eval time =    6263.57 ms /    29 runs   (  215.99
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   18231.60 ms /   129 tokens
No. of rows:  97%| | 1214/1258 [14:31:41<13:25, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.65 ms /    20 runs   (    0.38
ms per token,  2614.04 tokens per second)
llama_print_timings: prompt eval time =    9682.22 ms /    84 tokens (  115.26
ms per token,     8.68 tokens per second)
```

```
llama_print_timings:        eval time =    5813.97 ms /    19 runs   (  306.00
ms per token,     3.27 tokens per second)
llama_print_timings:       total time =   15620.70 ms /   103 tokens
No. of rows:  97%| | 1215/1258 [14:31:57<12:33, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.84 ms /    29 runs   (    0.37
ms per token,  2674.29 tokens per second)
llama_print_timings: prompt eval time =   12261.25 ms /   108 tokens (  113.53
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    6041.00 ms /    28 runs   (  215.75
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   18482.15 ms /   136 tokens
No. of rows:  97%| | 1216/1258 [14:32:15<12:27, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.82 ms /    18 runs   (    0.38
ms per token,  2638.91 tokens per second)
llama_print_timings: prompt eval time =    9384.81 ms /    81 tokens (  115.86
ms per token,     8.63 tokens per second)
llama_print_timings:        eval time =    3627.83 ms /    17 runs   (  213.40
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   13120.65 ms /    98 tokens
No. of rows:  97%| | 1217/1258 [14:32:29<11:12, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.50 ms /    27 runs   (    0.39
ms per token,  2571.18 tokens per second)
llama_print_timings: prompt eval time =   11227.19 ms /    99 tokens (  113.41
ms per token,     8.82 tokens per second)
llama_print_timings:        eval time =    7302.84 ms /    26 runs   (  280.88
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   18693.89 ms /   125 tokens
No. of rows:  97%| | 1218/1258 [14:32:47<11:23, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.60 ms /    22 runs   (    0.39
ms per token,  2558.73 tokens per second)
llama_print_timings: prompt eval time =   10840.50 ms /    94 tokens (  115.32
ms per token,     8.67 tokens per second)
llama_print_timings:        eval time =    4571.40 ms /    21 runs   (  217.69
ms per token,     4.59 tokens per second)
llama_print_timings:       total time =   15546.52 ms /   115 tokens
No. of rows:  97%| | 1219/1258 [14:33:03<10:48, 16Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.05 ms /    21 runs   (    0.38
ms per token,  2607.08 tokens per second)
llama_print_timings: prompt eval time =    9458.70 ms /    82 tokens (  115.35
ms per token,     8.67 tokens per second)
llama_print_timings:        eval time =    4267.70 ms /    20 runs   (  213.38
ms per token,     4.69 tokens per second)
llama_print_timings:       total time =   13855.71 ms /   102 tokens
No. of rows:  97%| | 1220/1258 [14:33:17<10:00, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      19.74 ms /    50 runs   (    0.39
ms per token,  2533.06 tokens per second)
llama_print_timings: prompt eval time =   13629.28 ms /   121 tokens (  112.64
ms per token,     8.88 tokens per second)
llama_print_timings:        eval time =   10636.38 ms /    49 runs   (  217.07
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   24578.78 ms /   170 tokens
No. of rows:  97%| | 1221/1258 [14:33:41<11:22, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       6.87 ms /    18 runs   (    0.38
ms per token,  2622.00 tokens per second)
llama_print_timings: prompt eval time =    8818.39 ms /    76 tokens (  116.03
ms per token,     8.62 tokens per second)
llama_print_timings:        eval time =    3653.23 ms /    17 runs   (  214.90
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   12581.58 ms /    93 tokens
No. of rows:  97%| | 1222/1258 [14:33:54<10:00, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.64 ms /    23 runs   (    0.38
ms per token,  2662.04 tokens per second)
llama_print_timings: prompt eval time =    9914.97 ms /    87 tokens (  113.97
ms per token,     8.77 tokens per second)
llama_print_timings:        eval time =    4707.82 ms /    22 runs   (  213.99
ms per token,     4.67 tokens per second)
llama_print_timings:       total time =   14764.38 ms /   109 tokens
No. of rows:  97%| | 1223/1258 [14:34:09<09:23, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.06 ms /    21 runs   (    0.38
```

```
ms per token,  2605.46 tokens per second)
llama_print_timings: prompt eval time =    9603.42 ms /    82 tokens (  117.11
ms per token,     8.54 tokens per second)
llama_print_timings:        eval time =    4298.05 ms /    20 runs   (  214.90
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   14030.47 ms /   102 tokens
No. of rows:  97%| | 1224/1258 [14:34:23<08:46, 15Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.37 ms /    29 runs   (    0.39
ms per token,  2551.24 tokens per second)
llama_print_timings: prompt eval time =   14225.67 ms /   126 tokens (  112.90
ms per token,     8.86 tokens per second)
llama_print_timings:        eval time =    6186.92 ms /    28 runs   (  220.96
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   20597.53 ms /   154 tokens
No. of rows:  97%| | 1225/1258 [14:34:43<09:21, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.10 ms /    28 runs   (    0.40
ms per token,  2522.52 tokens per second)
llama_print_timings: prompt eval time =   11386.02 ms /    89 tokens (  127.93
ms per token,     7.82 tokens per second)
llama_print_timings:        eval time =    5916.67 ms /    27 runs   (  219.14
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   17478.47 ms /   116 tokens
No. of rows:  97%| | 1226/1258 [14:35:01<09:09, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.24 ms /    25 runs   (    0.41
ms per token,  2442.36 tokens per second)
llama_print_timings: prompt eval time =   10601.30 ms /    92 tokens (  115.23
ms per token,     8.68 tokens per second)
llama_print_timings:        eval time =    5171.12 ms /    24 runs   (  215.46
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   15928.19 ms /   116 tokens
No. of rows:  98%| | 1227/1258 [14:35:17<08:40, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.83 ms /    23 runs   (    0.38
ms per token,  2605.64 tokens per second)
llama_print_timings: prompt eval time =   11620.25 ms /   104 tokens (  111.73
ms per token,     8.95 tokens per second)
llama_print_timings:        eval time =    4757.02 ms /    22 runs   (  216.23
```

```
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   16518.24 ms /   126 tokens
No. of rows:  98%| | 1228/1258 [14:35:33<08:21, 16Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      17.68 ms /    44 runs   (    0.40
ms per token,  2488.55 tokens per second)
llama_print_timings: prompt eval time =   19270.51 ms /   173 tokens (  111.39
ms per token,     8.98 tokens per second)
llama_print_timings:        eval time =    9331.08 ms /    43 runs   (  217.00
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   28882.04 ms /   216 tokens
No. of rows:  98%| | 1229/1258 [14:36:02<09:50, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.95 ms /    41 runs   (    0.39
ms per token,  2571.02 tokens per second)
llama_print_timings: prompt eval time =   15439.34 ms /   129 tokens (  119.68
ms per token,     8.36 tokens per second)
llama_print_timings:        eval time =    8679.82 ms /    40 runs   (  217.00
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   24374.03 ms /   169 tokens
No. of rows:  98%| | 1230/1258 [14:36:26<10:03, 21Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.13 ms /    27 runs   (    0.41
ms per token,  2425.22 tokens per second)
llama_print_timings: prompt eval time =   13379.74 ms /   106 tokens (  126.22
ms per token,     7.92 tokens per second)
llama_print_timings:        eval time =    6026.36 ms /    26 runs   (  231.78
ms per token,     4.31 tokens per second)
llama_print_timings:       total time =   19573.36 ms /   132 tokens
No. of rows:  98%| | 1231/1258 [14:36:46<09:26, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.01 ms /    18 runs   (    0.39
ms per token,  2568.49 tokens per second)
llama_print_timings: prompt eval time =    9623.27 ms /    81 tokens (  118.81
ms per token,     8.42 tokens per second)
llama_print_timings:        eval time =    3726.15 ms /    17 runs   (  219.19
ms per token,     4.56 tokens per second)
llama_print_timings:       total time =   13461.16 ms /    98 tokens
No. of rows:  98%| | 1232/1258 [14:36:59<08:06, 18Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.69 ms /    29 runs   (    0.40
ms per token,  2480.75 tokens per second)
llama_print_timings: prompt eval time =   11674.02 ms /   101 tokens (  115.58
ms per token,     8.65 tokens per second)
llama_print_timings:        eval time =    6072.55 ms /    28 runs   (  216.88
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   17925.69 ms /   129 tokens
No. of rows:  98%| | 1233/1258 [14:37:17<07:42, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.13 ms /    25 runs   (    0.41
ms per token,  2467.19 tokens per second)
llama_print_timings: prompt eval time =    9862.91 ms /    85 tokens (  116.03
ms per token,     8.62 tokens per second)
llama_print_timings:        eval time =    5200.18 ms /    24 runs   (  216.67
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   15221.51 ms /   109 tokens
No. of rows:  98%| | 1234/1258 [14:37:33<07:00, 17Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.11 ms /    35 runs   (    0.37
ms per token,  2668.90 tokens per second)
llama_print_timings: prompt eval time =   13760.75 ms /   120 tokens (  114.67
ms per token,     8.72 tokens per second)
llama_print_timings:        eval time =    7424.03 ms /    34 runs   (  218.35
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   21404.34 ms /   154 tokens
No. of rows:  98%| | 1235/1258 [14:37:54<07:09, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.04 ms /    39 runs   (    0.39
ms per token,  2593.09 tokens per second)
llama_print_timings: prompt eval time =   14656.86 ms /   120 tokens (  122.14
ms per token,     8.19 tokens per second)
llama_print_timings:        eval time =    9895.72 ms /    38 runs   (  260.41
ms per token,     3.84 tokens per second)
llama_print_timings:       total time =   24794.70 ms /   158 tokens
No. of rows:  98%| | 1236/1258 [14:38:19<07:31, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.30 ms /    32 runs   (    0.38
ms per token,  2600.57 tokens per second)
```

```
llama_print_timings: prompt eval time =   15102.11 ms /   134 tokens (  112.70
ms per token,     8.87 tokens per second)
llama_print_timings:        eval time =    6745.22 ms /    31 runs   (  217.59
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   22043.30 ms /   165 tokens
No. of rows:  98%| | 1237/1258 [14:38:41<07:20, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.87 ms /    27 runs   (    0.40
ms per token,  2483.44 tokens per second)
llama_print_timings: prompt eval time =   12209.93 ms /    96 tokens (  127.19
ms per token,     7.86 tokens per second)
llama_print_timings:        eval time =    5901.73 ms /    26 runs   (  226.99
ms per token,     4.41 tokens per second)
llama_print_timings:       total time =   18292.85 ms /   122 tokens
No. of rows:  98%| | 1238/1258 [14:38:59<06:43, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      18.99 ms /    50 runs   (    0.38
ms per token,  2632.96 tokens per second)
llama_print_timings: prompt eval time =   14069.42 ms /   125 tokens (  112.56
ms per token,     8.88 tokens per second)
llama_print_timings:        eval time =   10645.95 ms /    49 runs   (  217.26
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   25023.96 ms /   174 tokens
No. of rows:  98%| | 1239/1258 [14:39:24<06:51, 21Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       8.32 ms /    22 runs   (    0.38
ms per token,  2644.23 tokens per second)
llama_print_timings: prompt eval time =    9854.75 ms /    85 tokens (  115.94
ms per token,     8.63 tokens per second)
llama_print_timings:        eval time =    4634.36 ms /    21 runs   (  220.68
ms per token,     4.53 tokens per second)
llama_print_timings:       total time =   14624.36 ms /   106 tokens
No. of rows:  99%| | 1240/1258 [14:39:39<05:51, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       7.93 ms /    21 runs   (    0.38
ms per token,  2647.17 tokens per second)
llama_print_timings: prompt eval time =   10892.88 ms /    96 tokens (  113.47
ms per token,     8.81 tokens per second)
llama_print_timings:        eval time =    6015.36 ms /    20 runs   (  300.77
ms per token,     3.32 tokens per second)
```

```
llama_print_timings:       total time =   17039.06 ms /   116 tokens
No. of rows:  99%| | 1241/1258 [14:39:56<05:19, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.54 ms /    34 runs   (    0.40
ms per token,  2511.08 tokens per second)
llama_print_timings: prompt eval time =   12949.08 ms /   113 tokens (  114.59
ms per token,     8.73 tokens per second)
llama_print_timings:        eval time =    7102.30 ms /    33 runs   (  215.22
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   20264.89 ms /   146 tokens
No. of rows:  99%| | 1242/1258 [14:40:16<05:07, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.88 ms /    27 runs   (    0.44
ms per token,  2271.77 tokens per second)
llama_print_timings: prompt eval time =   12749.19 ms /   114 tokens (  111.84
ms per token,     8.94 tokens per second)
llama_print_timings:        eval time =    5828.31 ms /    26 runs   (  224.17
ms per token,     4.46 tokens per second)
llama_print_timings:       total time =   18755.48 ms /   140 tokens
No. of rows:  99%| | 1243/1258 [14:40:35<04:46, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      15.58 ms /    39 runs   (    0.40
ms per token,  2504.01 tokens per second)
llama_print_timings: prompt eval time =   14865.08 ms /   120 tokens (  123.88
ms per token,     8.07 tokens per second)
llama_print_timings:        eval time =    8445.46 ms /    38 runs   (  222.25
ms per token,     4.50 tokens per second)
llama_print_timings:       total time =   23556.84 ms /   158 tokens
No. of rows:  99%| | 1244/1258 [14:40:59<04:46, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.11 ms /    23 runs   (    0.40
ms per token,  2525.81 tokens per second)
llama_print_timings: prompt eval time =   10517.64 ms /    91 tokens (  115.58
ms per token,     8.65 tokens per second)
llama_print_timings:        eval time =    4741.26 ms /    22 runs   (  215.51
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   15403.22 ms /   113 tokens
No. of rows:  99%| | 1245/1258 [14:41:14<04:06, 18Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.89 ms /    28 runs   (    0.39
ms per token,  2571.40 tokens per second)
llama_print_timings: prompt eval time =   11228.89 ms /    98 tokens (  114.58
ms per token,     8.73 tokens per second)
llama_print_timings:        eval time =    5803.40 ms /    27 runs   (  214.94
ms per token,     4.65 tokens per second)
llama_print_timings:       total time =   17204.69 ms /   125 tokens
No. of rows:  99%| | 1246/1258 [14:41:31<03:40, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.53 ms /    29 runs   (    0.40
ms per token,  2514.96 tokens per second)
llama_print_timings: prompt eval time =   13066.68 ms /   115 tokens (  113.62
ms per token,     8.80 tokens per second)
llama_print_timings:        eval time =    6047.75 ms /    28 runs   (  215.99
ms per token,     4.63 tokens per second)
llama_print_timings:       total time =   19296.59 ms /   143 tokens
No. of rows:  99%| | 1247/1258 [14:41:50<03:25, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.13 ms /    26 runs   (    0.39
ms per token,  2567.65 tokens per second)
llama_print_timings: prompt eval time =   12155.83 ms /    99 tokens (  122.79
ms per token,     8.14 tokens per second)
llama_print_timings:        eval time =    5427.45 ms /    25 runs   (  217.10
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   17745.05 ms /   124 tokens
No. of rows:  99%| | 1248/1258 [14:42:08<03:04, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.50 ms /    32 runs   (    0.39
ms per token,  2559.80 tokens per second)
llama_print_timings: prompt eval time =   14110.96 ms /   113 tokens (  124.88
ms per token,     8.01 tokens per second)
llama_print_timings:        eval time =    8342.86 ms /    31 runs   (  269.12
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   22650.04 ms /   144 tokens
No. of rows:  99%| | 1249/1258 [14:42:31<02:57, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      11.49 ms /    30 runs   (    0.38
ms per token,  2610.51 tokens per second)
llama_print_timings: prompt eval time =   12252.83 ms /   105 tokens (  116.69
```

```
ms per token,     8.57 tokens per second)
llama_print_timings:        eval time =    6932.54 ms /    29 runs   (  239.05
ms per token,     4.18 tokens per second)
llama_print_timings:       total time =   19365.13 ms /   134 tokens
No. of rows:  99%| | 1250/1258 [14:42:50<02:36, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.91 ms /    21 runs   (    0.52
ms per token,  1925.55 tokens per second)
llama_print_timings: prompt eval time =   10101.81 ms /    83 tokens (  121.71
ms per token,     8.22 tokens per second)
llama_print_timings:        eval time =    4327.57 ms /    20 runs   (  216.38
ms per token,     4.62 tokens per second)
llama_print_timings:       total time =   14562.70 ms /   103 tokens
No. of rows:  99%| | 1251/1258 [14:43:05<02:06, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.49 ms /    33 runs   (    0.38
ms per token,  2642.75 tokens per second)
llama_print_timings: prompt eval time =   12778.04 ms /   113 tokens (  113.08
ms per token,     8.84 tokens per second)
llama_print_timings:        eval time =    6937.99 ms /    32 runs   (  216.81
ms per token,     4.61 tokens per second)
llama_print_timings:       total time =   19917.24 ms /   145 tokens
No. of rows: 100%| | 1252/1258 [14:43:25<01:51, 18Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      13.05 ms /    32 runs   (    0.41
ms per token,  2452.30 tokens per second)
llama_print_timings: prompt eval time =   13443.77 ms /   119 tokens (  112.97
ms per token,     8.85 tokens per second)
llama_print_timings:        eval time =    6746.03 ms /    31 runs   (  217.61
ms per token,     4.60 tokens per second)
llama_print_timings:       total time =   20388.36 ms /   150 tokens
No. of rows: 100%| | 1253/1258 [14:43:45<01:35, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.84 ms /    33 runs   (    0.39
ms per token,  2569.89 tokens per second)
llama_print_timings: prompt eval time =   14447.77 ms /   117 tokens (  123.49
ms per token,     8.10 tokens per second)
llama_print_timings:        eval time =    6981.65 ms /    32 runs   (  218.18
ms per token,     4.58 tokens per second)
llama_print_timings:       total time =   21635.81 ms /   149 tokens
```

```
No. of rows: 100%| | 1254/1258 [14:44:07<01:19, 19Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      20.83 ms /    50 runs   (    0.42
ms per token,  2400.50 tokens per second)
llama_print_timings: prompt eval time =   18179.22 ms /   155 tokens (  117.29
ms per token,     8.53 tokens per second)
llama_print_timings:        eval time =   10842.08 ms /    49 runs   (  221.27
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   29343.08 ms /   204 tokens
No. of rows: 100%| | 1255/1258 [14:44:36<01:08, 22Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      12.96 ms /    34 runs   (    0.38
ms per token,  2624.47 tokens per second)
llama_print_timings: prompt eval time =   12235.90 ms /   109 tokens (  112.26
ms per token,     8.91 tokens per second)
llama_print_timings:        eval time =    8905.19 ms /    33 runs   (  269.85
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   21353.71 ms /   142 tokens
No. of rows: 100%| | 1256/1258 [14:44:57<00:44, 22Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =      10.39 ms /    25 runs   (    0.42
ms per token,  2406.85 tokens per second)
llama_print_timings: prompt eval time =   10855.22 ms /    92 tokens (  117.99
ms per token,     8.48 tokens per second)
llama_print_timings:        eval time =    5176.25 ms /    24 runs   (  215.68
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   16186.72 ms /   116 tokens
No. of rows: 100%| | 1257/1258 [14:45:14<00:20, 20Llama.generate: prefix-match
hit

llama_print_timings:        load time =   11457.89 ms
llama_print_timings:      sample time =       9.21 ms /    24 runs   (    0.38
ms per token,  2605.30 tokens per second)
llama_print_timings: prompt eval time =    9855.13 ms /    86 tokens (  114.59
ms per token,     8.73 tokens per second)
llama_print_timings:        eval time =    4952.48 ms /    23 runs   (  215.33
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   14954.81 ms /   109 tokens
No. of rows: 100%| | 1258/1258 [14:45:29<00:00, 42
```

[112]:
```
del phi2
del out
```

[107]:
```
sqlc = Llama(model_path="../sqlcoder-7b-q5_k_m.gguf")
```

```
llama_model_loader: loaded meta data with 22 key-value pairs and 291 tensors
from ../sqlcoder-7b-q5_k_m.gguf (version GGUF V3 (latest))
llama_model_loader: Dumping metadata keys/values. Note: KV overrides do not
apply in this output.
llama_model_loader: - kv   0:                       general.architecture str
= llama
llama_model_loader: - kv   1:                               general.name str
= .
llama_model_loader: - kv   2:                       llama.context_length u32
= 16384
llama_model_loader: - kv   3:                     llama.embedding_length u32
= 4096
llama_model_loader: - kv   4:                          llama.block_count u32
= 32
llama_model_loader: - kv   5:                  llama.feed_forward_length u32
= 11008
llama_model_loader: - kv   6:                 llama.rope.dimension_count u32
= 128
llama_model_loader: - kv   7:                 llama.attention.head_count u32
= 32
llama_model_loader: - kv   8:              llama.attention.head_count_kv u32
= 32
llama_model_loader: - kv   9:     llama.attention.layer_norm_rms_epsilon f32
= 0.000010
llama_model_loader: - kv  10:                       llama.rope.freq_base f32
= 1000000.000000
llama_model_loader: - kv  11:                          general.file_type u32
= 17
llama_model_loader: - kv  12:                       tokenizer.ggml.model str
= llama
llama_model_loader: - kv  13:                      tokenizer.ggml.tokens
arr[str,32016]   = ["<unk>", "<s>", "</s>", "<0x00>", "<…
llama_model_loader: - kv  14:                      tokenizer.ggml.scores
arr[f32,32016]   = [0.000000, 0.000000, 0.000000, 0.0000…
llama_model_loader: - kv  15:                  tokenizer.ggml.token_type
arr[i32,32016]   = [2, 3, 3, 6, 6, 6, 6, 6, 6, 6, 6, 6, …
llama_model_loader: - kv  16:                tokenizer.ggml.bos_token_id u32
= 1
llama_model_loader: - kv  17:                tokenizer.ggml.eos_token_id u32
= 2
llama_model_loader: - kv  18:            tokenizer.ggml.unknown_token_id u32
= 0
llama_model_loader: - kv  19:               tokenizer.ggml.add_bos_token bool
```

```
= true
llama_model_loader: - kv  20:               tokenizer.ggml.add_eos_token bool
= false
llama_model_loader: - kv  21:               general.quantization_version u32
= 2
llama_model_loader: - type  f32:   65 tensors
llama_model_loader: - type q5_K:  193 tensors
llama_model_loader: - type q6_K:   33 tensors
llm_load_vocab: mismatch in special tokens definition ( 264/32016 vs 259/32016
).
llm_load_print_meta: format           = GGUF V3 (latest)
llm_load_print_meta: arch             = llama
llm_load_print_meta: vocab type       = SPM
llm_load_print_meta: n_vocab          = 32016
llm_load_print_meta: n_merges         = 0
llm_load_print_meta: n_ctx_train      = 16384
llm_load_print_meta: n_embd           = 4096
llm_load_print_meta: n_head           = 32
llm_load_print_meta: n_head_kv        = 32
llm_load_print_meta: n_layer          = 32
llm_load_print_meta: n_rot            = 128
llm_load_print_meta: n_embd_head_k    = 128
llm_load_print_meta: n_embd_head_v    = 128
llm_load_print_meta: n_gqa            = 1
llm_load_print_meta: n_embd_k_gqa     = 4096
llm_load_print_meta: n_embd_v_gqa     = 4096
llm_load_print_meta: f_norm_eps       = 0.0e+00
llm_load_print_meta: f_norm_rms_eps   = 1.0e-05
llm_load_print_meta: f_clamp_kqv      = 0.0e+00
llm_load_print_meta: f_max_alibi_bias = 0.0e+00
llm_load_print_meta: n_ff             = 11008
llm_load_print_meta: n_expert         = 0
llm_load_print_meta: n_expert_used    = 0
llm_load_print_meta: rope scaling     = linear
llm_load_print_meta: freq_base_train  = 1000000.0
llm_load_print_meta: freq_scale_train = 1
llm_load_print_meta: n_yarn_orig_ctx  = 16384
llm_load_print_meta: rope_finetuned   = unknown
llm_load_print_meta: model type       = 7B
llm_load_print_meta: model ftype      = Q5_K - Medium
llm_load_print_meta: model params     = 6.74 B
llm_load_print_meta: model size       = 4.45 GiB (5.68 BPW)
llm_load_print_meta: general.name     = .
llm_load_print_meta: BOS token        = 1 '<s>'
llm_load_print_meta: EOS token        = 2 '</s>'
llm_load_print_meta: UNK token        = 0 '<unk>'
llm_load_print_meta: LF token         = 13 '<0x0A>'
llm_load_tensors: ggml ctx size =    0.11 MiB
```

```
llm_load_tensors:        CPU buffer size =  4560.96 MiB
…
…
llama_new_context_with_model: n_ctx      = 512
llama_new_context_with_model: freq_base  = 1000000.0
llama_new_context_with_model: freq_scale = 1
llama_kv_cache_init:        CPU KV buffer size =   256.00 MiB
llama_new_context_with_model: KV self size  =  256.00 MiB, K (f16):  128.00 MiB,
V (f16):  128.00 MiB
llama_new_context_with_model:        CPU input buffer size   =    10.01 MiB
llama_new_context_with_model:        CPU compute buffer size =    70.53 MiB
llama_new_context_with_model: graph splits (measure): 1
AVX = 1 | AVX_VNNI = 0 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI =
0 | FMA = 1 | NEON = 0 | ARM_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 |
BLAS = 0 | SSE3 = 1 | SSSE3 = 0 | VSX = 0 | MATMUL_INT8 = 0 |
Model metadata: {'general.name': '.', 'general.architecture': 'llama',
'llama.context_length': '16384', 'llama.rope.dimension_count': '128',
'llama.embedding_length': '4096', 'llama.block_count': '32',
'llama.feed_forward_length': '11008', 'llama.attention.head_count': '32',
'tokenizer.ggml.eos_token_id': '2', 'general.file_type': '17',
'llama.attention.head_count_kv': '32', 'llama.attention.layer_norm_rms_epsilon':
'0.000010', 'llama.rope.freq_base': '1000000.000000', 'tokenizer.ggml.model':
'llama', 'general.quantization_version': '2', 'tokenizer.ggml.bos_token_id':
'1', 'tokenizer.ggml.unknown_token_id': '0', 'tokenizer.ggml.add_bos_token':
'true', 'tokenizer.ggml.add_eos_token': 'false'}
```

[115]:
```
out = {"prompt": [], "pred": [], "actu": [], "inf_time": [], "temperature": [],␣
 ↪"difficulty": [], "token_in": [], "token_out": [], "tokens_per_sec": []}

out = predict(df_c, sqlc, out)
json.dump(out, open("sqlc_eval_df_c.json", "w"))
```

```
No. of rows:   0%|          | 0/1258 [00:00<?, ?it/s]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.40 ms /    32 runs   (    0.26
ms per token,  3811.34 tokens per second)
llama_print_timings: prompt eval time =   19408.26 ms /    94 tokens (  206.47
ms per token,     4.84 tokens per second)
llama_print_timings:        eval time =    8400.02 ms /    31 runs   (  270.97
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   27927.12 ms /   125 tokens
No. of rows:   0%| | 1/1258 [00:27<9:45:10, 27.93sLlama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.25 ms /    31 runs   (    0.27
```

```
ms per token,  3755.75 tokens per second)
llama_print_timings: prompt eval time =   22683.84 ms /   103 tokens (  220.23
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    8678.70 ms /    30 runs   (  289.29
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   31492.03 ms /   133 tokens
No. of rows:   0%| | 2/1258 [00:59<10:28:40, 30.03Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.30 ms /    47 runs   (    0.28
ms per token,  3532.51 tokens per second)
llama_print_timings: prompt eval time =   26905.17 ms /   117 tokens (  229.96
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =   13792.32 ms /    46 runs   (  299.83
ms per token,     3.34 tokens per second)
llama_print_timings:       total time =   40897.37 ms /   163 tokens
No. of rows:   0%| | 3/1258 [01:40<12:12:00, 35.00Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.90 ms /    25 runs   (    0.36
ms per token,  2809.62 tokens per second)
llama_print_timings: prompt eval time =   22374.85 ms /    95 tokens (  235.52
ms per token,     4.25 tokens per second)
llama_print_timings:        eval time =    9110.19 ms /    24 runs   (  379.59
ms per token,     2.63 tokens per second)
llama_print_timings:       total time =   31616.48 ms /   119 tokens
No. of rows:   0%| | 4/1258 [02:11<11:43:37, 33.67Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.96 ms /    50 runs   (    0.30
ms per token,  3342.25 tokens per second)
llama_print_timings: prompt eval time =   41644.19 ms /   121 tokens (  344.17
ms per token,     2.91 tokens per second)
llama_print_timings:        eval time =   16148.17 ms /    49 runs   (  329.55
ms per token,     3.03 tokens per second)
llama_print_timings:       total time =   58019.21 ms /   170 tokens
No. of rows:   0%| | 5/1258 [03:09<14:46:28, 42.45Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.23 ms /    24 runs   (    0.30
ms per token,  3319.50 tokens per second)
llama_print_timings: prompt eval time =   22898.34 ms /    93 tokens (  246.22
ms per token,     4.06 tokens per second)
llama_print_timings:        eval time =    7658.32 ms /    23 runs   (  332.97
```

```
ms per token,     3.00 tokens per second)
llama_print_timings:       total time =   30660.35 ms /   116 tokens
No. of rows:   0%| | 6/1258 [03:40<13:22:13, 38.44Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.30 ms /    42 runs   (    0.27
ms per token,  3717.47 tokens per second)
llama_print_timings: prompt eval time =   25277.37 ms /   113 tokens (  223.69
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   11411.12 ms /    41 runs   (  278.32
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   36851.90 ms /   154 tokens
No. of rows:   1%| | 7/1258 [04:17<13:10:46, 37.93Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.43 ms /    50 runs   (    0.27
ms per token,  3722.45 tokens per second)
llama_print_timings: prompt eval time =   26235.00 ms /   124 tokens (  211.57
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   13850.97 ms /    49 runs   (  282.67
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   40286.30 ms /   173 tokens
No. of rows:   1%| | 8/1258 [04:57<13:25:45, 38.68Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.89 ms /    34 runs   (    0.26
ms per token,  3824.95 tokens per second)
llama_print_timings: prompt eval time =   23512.41 ms /   110 tokens (  213.75
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    9006.54 ms /    33 runs   (  272.93
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   32649.18 ms /   143 tokens
No. of rows:   1%| | 9/1258 [05:30<12:46:00, 36.80Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.37 ms /    50 runs   (    0.27
ms per token,  3740.00 tokens per second)
llama_print_timings: prompt eval time =   51326.92 ms /   242 tokens (  212.09
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   13435.49 ms /    49 runs   (  274.19
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   64961.85 ms /   291 tokens
No. of rows:   1%| | 10/1258 [06:35<15:46:17, 45.5Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.05 ms /    31 runs   (    0.29
ms per token,  3425.04 tokens per second)
llama_print_timings: prompt eval time =   21727.71 ms /   101 tokens (  215.13
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    8162.02 ms /    30 runs   (  272.07
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   30012.31 ms /   131 tokens
No. of rows:   1%| | 11/1258 [07:05<14:07:06, 40.7Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.46 ms /    50 runs   (    0.29
ms per token,  3458.53 tokens per second)
llama_print_timings: prompt eval time =   27609.64 ms /   132 tokens (  209.16
ms per token,     4.78 tokens per second)
llama_print_timings:        eval time =   13769.07 ms /    49 runs   (  281.00
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   41590.01 ms /   181 tokens
No. of rows:   1%| | 12/1258 [07:47<14:11:42, 41.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.28 ms /    31 runs   (    0.27
ms per token,  3742.61 tokens per second)
llama_print_timings: prompt eval time =   17608.79 ms /    81 tokens (  217.39
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    8035.86 ms /    30 runs   (  267.86
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   25768.96 ms /   111 tokens
No. of rows:   1%| | 13/1258 [08:12<12:35:14, 36.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.57 ms /    29 runs   (    0.26
ms per token,  3833.44 tokens per second)
llama_print_timings: prompt eval time =   17280.05 ms /    79 tokens (  218.73
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    9549.82 ms /    28 runs   (  341.06
ms per token,     2.93 tokens per second)
llama_print_timings:       total time =   26945.05 ms /   107 tokens
No. of rows:   1%| | 14/1258 [08:39<11:35:29, 33.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.22 ms /    43 runs   (    0.26
ms per token,  3832.44 tokens per second)
```

```
llama_print_timings: prompt eval time =   24991.18 ms /   112 tokens (  223.14
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =   13515.55 ms /    42 runs   (  321.80
ms per token,     3.11 tokens per second)
llama_print_timings:       total time =   38675.37 ms /   154 tokens
No. of rows:   1%| | 15/1258 [09:18<12:07:00, 35.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.14 ms /    33 runs   (    0.28
ms per token,  3610.50 tokens per second)
llama_print_timings: prompt eval time =   19919.51 ms /    93 tokens (  214.19
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    9000.20 ms /    32 runs   (  281.26
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   29052.10 ms /   125 tokens
No. of rows:   1%| | 16/1258 [09:47<11:28:50, 33.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.14 ms /    37 runs   (    0.27
ms per token,  3647.84 tokens per second)
llama_print_timings: prompt eval time =   21413.00 ms /    91 tokens (  235.31
ms per token,     4.25 tokens per second)
llama_print_timings:        eval time =   11608.80 ms /    36 runs   (  322.47
ms per token,     3.10 tokens per second)
llama_print_timings:       total time =   33170.51 ms /   127 tokens
No. of rows:   1%| | 17/1258 [10:20<11:27:40, 33.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.69 ms /    50 runs   (    0.27
ms per token,  3651.23 tokens per second)
llama_print_timings: prompt eval time =   22359.85 ms /   100 tokens (  223.60
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   14151.88 ms /    49 runs   (  288.81
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   36708.26 ms /   149 tokens
No. of rows:   1%| | 18/1258 [10:57<11:48:39, 34.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.09 ms /    47 runs   (    0.26
ms per token,  3886.87 tokens per second)
llama_print_timings: prompt eval time =   19991.35 ms /    94 tokens (  212.67
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   14258.59 ms /    46 runs   (  309.97
ms per token,     3.23 tokens per second)
```

```
llama_print_timings:       total time =   34429.61 ms /   140 tokens
No. of rows:   2%| | 19/1258 [11:31<11:49:00, 34.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.52 ms /    21 runs   (    0.26
ms per token,  3801.59 tokens per second)
llama_print_timings: prompt eval time =   17134.86 ms /    79 tokens (  216.90
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    5535.69 ms /    20 runs   (  276.78
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   22750.22 ms /    99 tokens
No. of rows:   2%| | 20/1258 [11:54<10:36:42, 30.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.73 ms /    50 runs   (    0.29
ms per token,  3394.43 tokens per second)
llama_print_timings: prompt eval time =   40690.01 ms /   184 tokens (  221.14
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =   13929.67 ms /    49 runs   (  284.28
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   54819.75 ms /   233 tokens
No. of rows:   2%| | 21/1258 [12:49<13:04:31, 38.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.47 ms /    24 runs   (    0.27
ms per token,  3711.15 tokens per second)
llama_print_timings: prompt eval time =   19853.26 ms /    93 tokens (  213.48
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    6308.58 ms /    23 runs   (  274.29
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   26253.45 ms /   116 tokens
No. of rows:   2%| | 22/1258 [13:15<11:50:58, 34.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.95 ms /    37 runs   (    0.27
ms per token,  3720.46 tokens per second)
llama_print_timings: prompt eval time =   22369.21 ms /   104 tokens (  215.09
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   11372.12 ms /    36 runs   (  315.89
ms per token,     3.17 tokens per second)
llama_print_timings:       total time =   33882.50 ms /   140 tokens
No. of rows:   2%| | 23/1258 [13:49<11:46:32, 34.3Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.51 ms /    38 runs   (    0.28
ms per token,  3615.95 tokens per second)
llama_print_timings: prompt eval time =   20415.15 ms /    95 tokens (  214.90
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   10265.13 ms /    37 runs   (  277.44
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   30832.72 ms /   132 tokens
No. of rows:   2%| | 24/1258 [14:20<11:24:27, 33.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.68 ms /    44 runs   (    0.27
ms per token,  3766.80 tokens per second)
llama_print_timings: prompt eval time =   24146.64 ms /   106 tokens (  227.80
ms per token,     4.39 tokens per second)
llama_print_timings:        eval time =   12139.59 ms /    43 runs   (  282.32
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   36461.34 ms /   149 tokens
No. of rows:   2%| | 25/1258 [14:56<11:43:35, 34.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.06 ms /    42 runs   (    0.29
ms per token,  3483.45 tokens per second)
llama_print_timings: prompt eval time =   20559.62 ms /    94 tokens (  218.72
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =   11795.96 ms /    41 runs   (  287.71
ms per token,     3.48 tokens per second)
llama_print_timings:       total time =   32525.29 ms /   135 tokens
No. of rows:   2%| | 26/1258 [15:29<11:32:30, 33.7Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.10 ms /    30 runs   (    0.27
ms per token,  3704.16 tokens per second)
llama_print_timings: prompt eval time =   18421.73 ms /    77 tokens (  239.24
ms per token,     4.18 tokens per second)
llama_print_timings:        eval time =    8233.23 ms /    29 runs   (  283.90
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   26774.25 ms /   106 tokens
No. of rows:   2%| | 27/1258 [15:56<10:49:11, 31.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.52 ms /    38 runs   (    0.28
ms per token,  3611.14 tokens per second)
llama_print_timings: prompt eval time =   24942.17 ms /   109 tokens (  228.83
```

```
ms per token,     4.37 tokens per second)
llama_print_timings:        eval time =   11776.73 ms /    37 runs   (  318.29
ms per token,     3.14 tokens per second)
llama_print_timings:       total time =   36865.02 ms /   146 tokens
No. of rows:   2%| | 28/1258 [16:33<11:20:50, 33.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.30 ms /    50 runs   (    0.29
ms per token,  3496.26 tokens per second)
llama_print_timings: prompt eval time =   31588.84 ms /   147 tokens (  214.89
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13741.43 ms /    49 runs   (  280.44
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   45533.98 ms /   196 tokens
No. of rows:   2%| | 29/1258 [17:18<12:36:03, 36.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.64 ms /    24 runs   (    0.28
ms per token,  3613.37 tokens per second)
llama_print_timings: prompt eval time =   19145.93 ms /    88 tokens (  217.57
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    6239.82 ms /    23 runs   (  271.30
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   25479.48 ms /   111 tokens
No. of rows:   2%| | 30/1258 [17:44<11:25:17, 33.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.02 ms /    31 runs   (    0.26
ms per token,  3863.89 tokens per second)
llama_print_timings: prompt eval time =   21269.66 ms /    91 tokens (  233.73
ms per token,     4.28 tokens per second)
llama_print_timings:        eval time =    8180.15 ms /    30 runs   (  272.67
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   29566.50 ms /   121 tokens
No. of rows:   2%| | 31/1258 [18:13<11:00:42, 32.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.19 ms /    50 runs   (    0.28
ms per token,  3523.61 tokens per second)
llama_print_timings: prompt eval time =   25850.09 ms /   111 tokens (  232.88
ms per token,     4.29 tokens per second)
llama_print_timings:        eval time =   13863.41 ms /    49 runs   (  282.93
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   39917.28 ms /   160 tokens
```

```
No. of rows:   3%| | 32/1258 [18:53<11:46:49, 34.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.71 ms /    34 runs   (    0.26
ms per token,  3904.46 tokens per second)
llama_print_timings: prompt eval time =   21265.48 ms /    99 tokens (  214.80
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    9322.81 ms /    33 runs   (  282.51
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   30722.23 ms /   132 tokens
No. of rows:   3%| | 33/1258 [19:24<11:22:40, 33.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.21 ms /    50 runs   (    0.26
ms per token,  3784.44 tokens per second)
llama_print_timings: prompt eval time =   18891.70 ms /    88 tokens (  214.68
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13794.83 ms /    49 runs   (  281.53
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   32883.04 ms /   137 tokens
No. of rows:   3%| | 34/1258 [19:57<11:18:46, 33.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.74 ms /    29 runs   (    0.27
ms per token,  3745.32 tokens per second)
llama_print_timings: prompt eval time =   21277.98 ms /    90 tokens (  236.42
ms per token,     4.23 tokens per second)
llama_print_timings:        eval time =    7654.08 ms /    28 runs   (  273.36
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   29044.30 ms /   118 tokens
No. of rows:   3%| | 35/1258 [20:26<10:52:24, 32.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.40 ms /    32 runs   (    0.26
ms per token,  3808.16 tokens per second)
llama_print_timings: prompt eval time =   23410.42 ms /   101 tokens (  231.79
ms per token,     4.31 tokens per second)
llama_print_timings:        eval time =    8515.57 ms /    31 runs   (  274.70
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   32054.38 ms /   132 tokens
No. of rows:   3%| | 36/1258 [20:58<10:52:11, 32.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =       7.03 ms /    27 runs   (    0.26
ms per token,  3840.68 tokens per second)
llama_print_timings: prompt eval time =   19912.33 ms /    93 tokens (  214.11
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    7116.74 ms /    26 runs   (  273.72
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   27131.20 ms /   119 tokens
No. of rows:   3%| | 37/1258 [21:25<10:21:50, 30.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.84 ms /    34 runs   (    0.26
ms per token,  3844.85 tokens per second)
llama_print_timings: prompt eval time =   21594.78 ms /   101 tokens (  213.81
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    9156.57 ms /    33 runs   (  277.47
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   30878.56 ms /   134 tokens
No. of rows:   3%| | 38/1258 [21:56<10:23:19, 30.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.37 ms /    45 runs   (    0.25
ms per token,  3959.52 tokens per second)
llama_print_timings: prompt eval time =   30409.17 ms /   135 tokens (  225.25
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =   12059.01 ms /    44 runs   (  274.07
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   42641.21 ms /   179 tokens
No. of rows:   3%| | 39/1258 [22:39<11:35:55, 34.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.23 ms /    50 runs   (    0.26
ms per token,  3779.58 tokens per second)
llama_print_timings: prompt eval time =   21896.23 ms /   102 tokens (  214.67
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13529.05 ms /    49 runs   (  276.10
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   35615.89 ms /   151 tokens
No. of rows:   3%| | 40/1258 [23:14<11:43:41, 34.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.13 ms /    44 runs   (    0.28
ms per token,  3628.57 tokens per second)
llama_print_timings: prompt eval time =   25496.69 ms /   119 tokens (  214.26
ms per token,     4.67 tokens per second)
```

```
llama_print_timings:        eval time =   13819.63 ms /    43 runs   (  321.39
ms per token,     3.11 tokens per second)
llama_print_timings:       total time =   39492.09 ms /   162 tokens
No. of rows:   3%| | 41/1258 [23:54<12:12:32, 36.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.71 ms /    31 runs   (    0.28
ms per token,  3561.17 tokens per second)
llama_print_timings: prompt eval time =   18109.58 ms /    83 tokens (  218.19
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    8152.72 ms /    30 runs   (  271.76
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   26382.45 ms /   113 tokens
No. of rows:   3%| | 42/1258 [24:20<11:12:48, 33.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.24 ms /    48 runs   (    0.28
ms per token,  3624.83 tokens per second)
llama_print_timings: prompt eval time =   25304.45 ms /   112 tokens (  225.93
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   13235.41 ms /    47 runs   (  281.60
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   38740.97 ms /   159 tokens
No. of rows:   3%| | 43/1258 [24:59<11:45:57, 34.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.59 ms /    46 runs   (    0.27
ms per token,  3654.27 tokens per second)
llama_print_timings: prompt eval time =   22875.21 ms /   106 tokens (  215.80
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   12612.32 ms /    45 runs   (  280.27
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   35672.06 ms /   151 tokens
No. of rows:   3%| | 44/1258 [25:34<11:50:20, 35.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.58 ms /    31 runs   (    0.28
ms per token,  3613.47 tokens per second)
llama_print_timings: prompt eval time =   21221.06 ms /    97 tokens (  218.77
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    8257.43 ms /    30 runs   (  275.25
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   29598.49 ms /   127 tokens
No. of rows:   4%| | 45/1258 [26:04<11:16:23, 33.4Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.88 ms /    35 runs   (    0.25
ms per token,  3942.77 tokens per second)
llama_print_timings: prompt eval time =   20766.35 ms /    96 tokens (  216.32
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    9359.88 ms /    34 runs   (  275.29
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   30260.78 ms /   130 tokens
No. of rows:   4%| | 46/1258 [26:34<10:56:32, 32.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.42 ms /    50 runs   (    0.27
ms per token,  3726.62 tokens per second)
llama_print_timings: prompt eval time =   28772.39 ms /   126 tokens (  228.35
ms per token,     4.38 tokens per second)
llama_print_timings:        eval time =   13350.12 ms /    49 runs   (  272.45
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   42319.45 ms /   175 tokens
No. of rows:   4%| | 47/1258 [27:17<11:55:28, 35.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.21 ms /    40 runs   (    0.28
ms per token,  3567.92 tokens per second)
llama_print_timings: prompt eval time =   19816.16 ms /    93 tokens (  213.08
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   11143.45 ms /    39 runs   (  285.73
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   31117.84 ms /   132 tokens
No. of rows:   4%| | 48/1258 [27:48<11:28:43, 34.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.24 ms /    38 runs   (    0.27
ms per token,  3709.85 tokens per second)
llama_print_timings: prompt eval time =   24674.92 ms /   106 tokens (  232.78
ms per token,     4.30 tokens per second)
llama_print_timings:        eval time =   10141.25 ms /    37 runs   (  274.09
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   34966.41 ms /   143 tokens
No. of rows:   4%| | 49/1258 [28:23<11:33:07, 34.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.82 ms /    42 runs   (    0.26
```

```
ms per token,  3882.78 tokens per second)
llama_print_timings: prompt eval time =   15586.83 ms /    74 tokens (  210.63
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =   12749.97 ms /    41 runs   (  310.97
ms per token,     3.22 tokens per second)
llama_print_timings:       total time =   28499.25 ms /   115 tokens
No. of rows:   4%| | 50/1258 [28:51<10:56:57, 32.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.08 ms /    50 runs   (    0.26
ms per token,  3823.21 tokens per second)
llama_print_timings: prompt eval time =   25784.60 ms /   120 tokens (  214.87
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13487.08 ms /    49 runs   (  275.25
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   39467.29 ms /   169 tokens
No. of rows:   4%| | 51/1258 [29:31<11:37:43, 34.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.96 ms /    37 runs   (    0.27
ms per token,  3715.23 tokens per second)
llama_print_timings: prompt eval time =   22677.85 ms /   106 tokens (  213.94
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    9750.27 ms /    36 runs   (  270.84
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   32575.53 ms /   142 tokens
No. of rows:   4%| | 52/1258 [30:03<11:24:28, 34.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      16.11 ms /    50 runs   (    0.32
ms per token,  3103.08 tokens per second)
llama_print_timings: prompt eval time =   24702.14 ms /   108 tokens (  228.72
ms per token,     4.37 tokens per second)
llama_print_timings:        eval time =   13492.01 ms /    49 runs   (  275.35
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   38389.74 ms /   157 tokens
No. of rows:   4%| | 53/1258 [30:42<11:50:04, 35.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.90 ms /    31 runs   (    0.25
ms per token,  3922.06 tokens per second)
llama_print_timings: prompt eval time =   20045.89 ms /    94 tokens (  213.25
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    8144.71 ms /    30 runs   (  271.49
```

```
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   28308.47 ms /   124 tokens
No. of rows:   4%| | 54/1258 [31:10<11:07:06, 33.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.39 ms /    50 runs   (    0.27
ms per token,  3734.69 tokens per second)
llama_print_timings: prompt eval time =   21462.40 ms /   101 tokens (  212.50
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   13473.08 ms /    49 runs   (  274.96
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   35126.19 ms /   150 tokens
No. of rows:   4%| | 55/1258 [31:45<11:17:54, 33.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.71 ms /    48 runs   (    0.26
ms per token,  3776.85 tokens per second)
llama_print_timings: prompt eval time =   24064.42 ms /   112 tokens (  214.86
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   14406.38 ms /    47 runs   (  306.52
ms per token,     3.26 tokens per second)
llama_print_timings:       total time =   38659.75 ms /   159 tokens
No. of rows:   4%| | 56/1258 [32:24<11:46:32, 35.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.90 ms /    38 runs   (    0.26
ms per token,  3838.00 tokens per second)
llama_print_timings: prompt eval time =   25336.27 ms /   119 tokens (  212.91
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   10141.43 ms /    37 runs   (  274.09
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   35624.72 ms /   156 tokens
No. of rows:   5%| | 57/1258 [32:59<11:48:06, 35.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      16.01 ms /    50 runs   (    0.32
ms per token,  3122.85 tokens per second)
llama_print_timings: prompt eval time =   26170.44 ms /   117 tokens (  223.68
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   13382.25 ms /    49 runs   (  273.11
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   39747.87 ms /   166 tokens
No. of rows:   5%| | 58/1258 [33:39<12:13:49, 36.6Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.46 ms /    45 runs   (    0.28
ms per token,  3610.40 tokens per second)
llama_print_timings: prompt eval time =   24299.28 ms /   113 tokens (  215.04
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13756.22 ms /    44 runs   (  312.64
ms per token,     3.20 tokens per second)
llama_print_timings:       total time =   38235.72 ms /   157 tokens
No. of rows:   5%| | 59/1258 [34:17<12:22:31, 37.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.37 ms /    45 runs   (    0.25
ms per token,  3957.78 tokens per second)
llama_print_timings: prompt eval time =   22576.79 ms /   106 tokens (  212.99
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   11890.95 ms /    44 runs   (  270.25
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   34638.27 ms /   150 tokens
No. of rows:   5%| | 60/1258 [34:52<12:06:51, 36.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.49 ms /    31 runs   (    0.27
ms per token,  3652.21 tokens per second)
llama_print_timings: prompt eval time =   22102.48 ms /    94 tokens (  235.13
ms per token,     4.25 tokens per second)
llama_print_timings:        eval time =    8133.24 ms /    30 runs   (  271.11
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   30354.17 ms /   124 tokens
No. of rows:   5%| | 61/1258 [35:22<11:30:05, 34.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.08 ms /    50 runs   (    0.26
ms per token,  3821.46 tokens per second)
llama_print_timings: prompt eval time =   21673.32 ms /   102 tokens (  212.48
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   13510.41 ms /    49 runs   (  275.72
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   35380.07 ms /   151 tokens
No. of rows:   5%| | 62/1258 [35:58<11:34:16, 34.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.35 ms /    38 runs   (    0.27
ms per token,  3671.50 tokens per second)
```

```
llama_print_timings: prompt eval time =   18231.26 ms /    85 tokens (  214.49
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    9964.80 ms /    37 runs   (  269.32
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   28343.33 ms /   122 tokens
No. of rows:   5%| | 63/1258 [36:26<10:54:58, 32.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.71 ms /    44 runs   (    0.27
ms per token,  3757.79 tokens per second)
llama_print_timings: prompt eval time =   25162.94 ms /   118 tokens (  213.25
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   11810.01 ms /    43 runs   (  274.65
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   37143.24 ms /   161 tokens
No. of rows:   5%| | 64/1258 [37:03<11:19:52, 34.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.36 ms /    42 runs   (    0.29
ms per token,  3399.43 tokens per second)
llama_print_timings: prompt eval time =   23016.91 ms /   107 tokens (  215.11
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   11200.90 ms /    41 runs   (  273.19
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   34384.82 ms /   148 tokens
No. of rows:   5%| | 65/1258 [37:38<11:20:39, 34.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.37 ms /    50 runs   (    0.27
ms per token,  3739.44 tokens per second)
llama_print_timings: prompt eval time =   24122.09 ms /   105 tokens (  229.73
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =   13389.29 ms /    49 runs   (  273.25
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   37706.67 ms /   154 tokens
No. of rows:   5%| | 66/1258 [38:15<11:40:50, 35.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.79 ms /    50 runs   (    0.26
ms per token,  3910.22 tokens per second)
llama_print_timings: prompt eval time =   23545.19 ms /   102 tokens (  230.84
ms per token,     4.33 tokens per second)
llama_print_timings:        eval time =   14862.48 ms /    49 runs   (  303.32
ms per token,     3.30 tokens per second)
```

```
llama_print_timings:       total time =   38596.29 ms /   151 tokens
No. of rows:   5%| | 67/1258 [38:54<12:00:03, 36.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.67 ms /    29 runs   (    0.26
ms per token,  3778.50 tokens per second)
llama_print_timings: prompt eval time =   18408.91 ms /    85 tokens (  216.58
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7718.59 ms /    28 runs   (  275.66
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   26241.50 ms /   113 tokens
No. of rows:   5%| | 68/1258 [39:20<10:59:47, 33.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.57 ms /    44 runs   (    0.29
ms per token,  3499.28 tokens per second)
llama_print_timings: prompt eval time =   26655.58 ms /   126 tokens (  211.55
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   13778.20 ms /    43 runs   (  320.42
ms per token,     3.12 tokens per second)
llama_print_timings:       total time =   40619.25 ms /   169 tokens
No. of rows:   5%| | 69/1258 [40:01<11:43:00, 35.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.67 ms /    30 runs   (    0.26
ms per token,  3911.34 tokens per second)
llama_print_timings: prompt eval time =   20477.99 ms /    94 tokens (  217.85
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    9577.64 ms /    29 runs   (  330.26
ms per token,     3.03 tokens per second)
llama_print_timings:       total time =   30170.69 ms /   123 tokens
No. of rows:   6%| | 70/1258 [40:31<11:10:56, 33.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.37 ms /    46 runs   (    0.27
ms per token,  3718.07 tokens per second)
llama_print_timings: prompt eval time =   26603.60 ms /   124 tokens (  214.55
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   12515.59 ms /    45 runs   (  278.12
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   39307.58 ms /   169 tokens
No. of rows:   6%| | 71/1258 [41:10<11:42:35, 35.5Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      15.74 ms /    50 runs   (    0.31
ms per token,  3176.62 tokens per second)
llama_print_timings: prompt eval time =   21282.15 ms /   100 tokens (  212.82
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   13403.15 ms /    49 runs   (  273.53
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   34879.96 ms /   149 tokens
No. of rows:   6%| | 72/1258 [41:45<11:38:16, 35.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.83 ms /    33 runs   (    0.27
ms per token,  3739.38 tokens per second)
llama_print_timings: prompt eval time =   18822.58 ms /    85 tokens (  221.44
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    8783.34 ms /    32 runs   (  274.48
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   27732.42 ms /   117 tokens
No. of rows:   6%| | 73/1258 [42:13<10:52:43, 33.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.17 ms /    34 runs   (    0.27
ms per token,  3708.96 tokens per second)
llama_print_timings: prompt eval time =   18273.09 ms /    77 tokens (  237.31
ms per token,     4.21 tokens per second)
llama_print_timings:        eval time =    8958.26 ms /    33 runs   (  271.46
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   27361.50 ms /   110 tokens
No. of rows:   6%| | 74/1258 [42:40<10:18:33, 31.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.65 ms /    42 runs   (    0.28
ms per token,  3605.77 tokens per second)
llama_print_timings: prompt eval time =   25602.46 ms /   113 tokens (  226.57
ms per token,     4.41 tokens per second)
llama_print_timings:        eval time =   11642.23 ms /    41 runs   (  283.96
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   37412.49 ms /   154 tokens
No. of rows:   6%| | 75/1258 [43:18<10:53:57, 33.1Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.20 ms /    24 runs   (    0.38
ms per token,  2609.83 tokens per second)
llama_print_timings: prompt eval time =   17108.35 ms /    80 tokens (  213.85
```

```
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    6114.97 ms /    23 runs   (  265.87
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   23317.96 ms /   103 tokens
No. of rows:   6%| | 76/1258 [43:41<9:55:13, 30.21Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.78 ms /    32 runs   (    0.27
ms per token,  3646.31 tokens per second)
llama_print_timings: prompt eval time =   19937.67 ms /    93 tokens (  214.38
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    8304.74 ms /    31 runs   (  267.89
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   28364.41 ms /   124 tokens
No. of rows:   6%| | 77/1258 [44:10<9:43:50, 29.66Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.64 ms /    37 runs   (    0.26
ms per token,  3837.78 tokens per second)
llama_print_timings: prompt eval time =   22993.01 ms /   100 tokens (  229.93
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =    9812.02 ms /    36 runs   (  272.56
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   32943.79 ms /   136 tokens
No. of rows:   6%| | 78/1258 [44:43<10:02:45, 30.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.90 ms /    50 runs   (    0.28
ms per token,  3597.64 tokens per second)
llama_print_timings: prompt eval time =   18707.60 ms /    87 tokens (  215.03
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   15440.35 ms /    49 runs   (  315.11
ms per token,     3.17 tokens per second)
llama_print_timings:       total time =   34346.04 ms /   136 tokens
No. of rows:   6%| | 79/1258 [45:17<10:24:04, 31.7Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.05 ms /    20 runs   (    0.25
ms per token,  3960.40 tokens per second)
llama_print_timings: prompt eval time =   18201.50 ms /    83 tokens (  219.30
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    5139.95 ms /    19 runs   (  270.52
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   23415.98 ms /   102 tokens
```

```
No. of rows:   6%| | 80/1258 [45:40<9:34:27, 29.26Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.71 ms /    30 runs   (    0.26
ms per token,  3890.04 tokens per second)
llama_print_timings: prompt eval time =   19186.43 ms /    90 tokens (  213.18
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    9463.85 ms /    29 runs   (  326.34
ms per token,     3.06 tokens per second)
llama_print_timings:       total time =   28765.46 ms /   119 tokens
No. of rows:   6%| | 81/1258 [46:09<9:31:06, 29.11Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.35 ms /    28 runs   (    0.26
ms per token,  3807.97 tokens per second)
llama_print_timings: prompt eval time =   20572.62 ms /    96 tokens (  214.30
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    7261.62 ms /    27 runs   (  268.95
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   27945.55 ms /   123 tokens
No. of rows:   7%| | 82/1258 [46:37<9:23:47, 28.77Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.11 ms /    23 runs   (    0.27
ms per token,  3761.86 tokens per second)
llama_print_timings: prompt eval time =   17934.59 ms /    76 tokens (  235.98
ms per token,     4.24 tokens per second)
llama_print_timings:        eval time =    5895.68 ms /    22 runs   (  267.99
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   23916.79 ms /    98 tokens
No. of rows:   7%| | 83/1258 [47:01<8:54:52, 27.31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.18 ms /    43 runs   (    0.26
ms per token,  3846.50 tokens per second)
llama_print_timings: prompt eval time =   29497.89 ms /   130 tokens (  226.91
ms per token,     4.41 tokens per second)
llama_print_timings:        eval time =   11505.99 ms /    42 runs   (  273.95
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   41169.40 ms /   172 tokens
No. of rows:   7%| | 84/1258 [47:42<10:15:49, 31.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      12.98 ms /    50 runs   (    0.26
ms per token,  3852.08 tokens per second)
llama_print_timings: prompt eval time =   30268.18 ms /   127 tokens (  238.33
ms per token,     4.20 tokens per second)
llama_print_timings:        eval time =   15655.03 ms /    49 runs   (  319.49
ms per token,     3.13 tokens per second)
llama_print_timings:       total time =   46114.86 ms /   176 tokens
No. of rows:   7%| | 85/1258 [48:28<11:41:08, 35.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.59 ms /    46 runs   (    0.25
ms per token,  3968.25 tokens per second)
llama_print_timings: prompt eval time =   25091.78 ms /   112 tokens (  224.03
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =   14440.00 ms /    45 runs   (  320.89
ms per token,     3.12 tokens per second)
llama_print_timings:       total time =   39711.31 ms /   157 tokens
No. of rows:   7%| | 86/1258 [49:08<12:03:08, 37.0Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.31 ms /    43 runs   (    0.26
ms per token,  3800.60 tokens per second)
llama_print_timings: prompt eval time =   26122.12 ms /   117 tokens (  223.27
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =   11804.04 ms /    42 runs   (  281.05
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   38092.64 ms /   159 tokens
No. of rows:   7%| | 87/1258 [49:46<12:08:50, 37.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.48 ms /    28 runs   (    0.27
ms per token,  3743.32 tokens per second)
llama_print_timings: prompt eval time =   22732.72 ms /   100 tokens (  227.33
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =    9298.91 ms /    27 runs   (  344.40
ms per token,     2.90 tokens per second)
llama_print_timings:       total time =   32141.44 ms /   127 tokens
No. of rows:   7%| | 88/1258 [50:18<11:37:49, 35.7Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.24 ms /    50 runs   (    0.26
ms per token,  3776.72 tokens per second)
llama_print_timings: prompt eval time =   28179.45 ms /   126 tokens (  223.65
ms per token,     4.47 tokens per second)
```

```
llama_print_timings:        eval time =   15526.92 ms /    49 runs   (  316.88
ms per token,     3.16 tokens per second)
llama_print_timings:       total time =   43901.35 ms /   175 tokens
No. of rows:   7%| | 89/1258 [51:02<12:24:40, 38.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.99 ms /    31 runs   (    0.26
ms per token,  3879.36 tokens per second)
llama_print_timings: prompt eval time =   21016.12 ms /    95 tokens (  221.22
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    8416.39 ms /    30 runs   (  280.55
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   29554.56 ms /   125 tokens
No. of rows:   7%| | 90/1258 [51:32<11:33:30, 35.6Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.99 ms /    29 runs   (    0.28
ms per token,  3630.45 tokens per second)
llama_print_timings: prompt eval time =   17979.37 ms /    79 tokens (  227.59
ms per token,     4.39 tokens per second)
llama_print_timings:        eval time =    9655.49 ms /    28 runs   (  344.84
ms per token,     2.90 tokens per second)
llama_print_timings:       total time =   27748.71 ms /   107 tokens
No. of rows:   7%| | 91/1258 [51:59<10:47:01, 33.2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.68 ms /    31 runs   (    0.25
ms per token,  4035.41 tokens per second)
llama_print_timings: prompt eval time =   20111.75 ms /    87 tokens (  231.17
ms per token,     4.33 tokens per second)
llama_print_timings:        eval time =    8491.63 ms /    30 runs   (  283.05
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   28722.38 ms /   117 tokens
No. of rows:   7%| | 92/1258 [52:28<10:20:02, 31.9Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.40 ms /    38 runs   (    0.25
ms per token,  4040.40 tokens per second)
llama_print_timings: prompt eval time =   24173.88 ms /   110 tokens (  219.76
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   12129.19 ms /    37 runs   (  327.82
ms per token,     3.05 tokens per second)
llama_print_timings:       total time =   36448.28 ms /   147 tokens
No. of rows:   7%| | 93/1258 [53:05<10:45:56, 33.2Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.91 ms /    45 runs   (    0.26
ms per token,  3778.34 tokens per second)
llama_print_timings: prompt eval time =   21341.46 ms /    96 tokens (  222.31
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =   12369.60 ms /    44 runs   (  281.13
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   33884.77 ms /   140 tokens
No. of rows:   7%| | 94/1258 [53:38<10:49:03, 33.4Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.09 ms /    27 runs   (    0.26
ms per token,  3808.18 tokens per second)
llama_print_timings: prompt eval time =   19034.97 ms /    85 tokens (  223.94
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =    7252.70 ms /    26 runs   (  278.95
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   26391.88 ms /   111 tokens
No. of rows:   8%| | 95/1258 [54:05<10:07:28, 31.3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.50 ms /    33 runs   (    0.26
ms per token,  3884.64 tokens per second)
llama_print_timings: prompt eval time =   26269.20 ms /   112 tokens (  234.55
ms per token,     4.26 tokens per second)
llama_print_timings:        eval time =    8801.57 ms /    32 runs   (  275.05
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   35197.11 ms /   144 tokens
No. of rows:   8%| | 96/1258 [54:40<10:29:22, 32.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.26 ms /    50 runs   (    0.27
ms per token,  3771.88 tokens per second)
llama_print_timings: prompt eval time =   31769.12 ms /   138 tokens (  230.21
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =   14156.82 ms /    49 runs   (  288.91
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   46119.12 ms /   187 tokens
No. of rows:   8%| | 97/1258 [55:26<11:47:58, 36.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.84 ms /    36 runs   (    0.27
```

```
ms per token,  3658.16 tokens per second)
llama_print_timings: prompt eval time =   20471.73 ms /    91 tokens (  224.96
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   10183.44 ms /    35 runs   (  290.96
ms per token,     3.44 tokens per second)
llama_print_timings:       total time =   30803.74 ms /   126 tokens
No. of rows:   8%| | 98/1258 [55:57<11:13:52, 34.8Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.70 ms /    30 runs   (    0.26
ms per token,  3895.09 tokens per second)
llama_print_timings: prompt eval time =   22467.85 ms /   100 tokens (  224.68
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =    8061.66 ms /    29 runs   (  277.99
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   30646.20 ms /   129 tokens
No. of rows:   8%| | 99/1258 [56:28<10:48:55, 33.5Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.12 ms /    50 runs   (    0.26
ms per token,  3810.69 tokens per second)
llama_print_timings: prompt eval time =   23783.58 ms /   108 tokens (  220.22
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =   13781.72 ms /    49 runs   (  281.26
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   37758.24 ms /   157 tokens
No. of rows:   8%| | 100/1258 [57:05<11:12:27, 34.Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.68 ms /    36 runs   (    0.24
ms per token,  4149.38 tokens per second)
llama_print_timings: prompt eval time =   23240.61 ms /   104 tokens (  223.47
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =    9810.39 ms /    35 runs   (  280.30
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   33188.50 ms /   139 tokens
No. of rows:   8%| | 101/1258 [57:39<11:02:24, 34.Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.38 ms /    28 runs   (    0.26
ms per token,  3795.07 tokens per second)
llama_print_timings: prompt eval time =   21885.28 ms /    97 tokens (  225.62
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =    7546.65 ms /    27 runs   (  279.51
```

```
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   29538.22 ms /   124 tokens
No. of rows:   8%| | 102/1258 [58:08<10:34:01, 32.Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.55 ms /    34 runs   (    0.28
ms per token,  3558.72 tokens per second)
llama_print_timings: prompt eval time =   20205.88 ms /    90 tokens (  224.51
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =    9061.69 ms /    33 runs   (  274.60
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   29400.20 ms /   123 tokens
No. of rows:   8%| | 103/1258 [58:38<10:13:15, 31.Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.29 ms /    32 runs   (    0.26
ms per token,  3860.07 tokens per second)
llama_print_timings: prompt eval time =   22014.85 ms /    99 tokens (  222.37
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =    8532.55 ms /    31 runs   (  275.24
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   30668.11 ms /   130 tokens
No. of rows:   8%| | 104/1258 [59:08<10:05:53, 31.Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.30 ms /    25 runs   (    0.25
ms per token,  3966.99 tokens per second)
llama_print_timings: prompt eval time =   19121.62 ms /    76 tokens (  251.60
ms per token,     3.97 tokens per second)
llama_print_timings:        eval time =    6464.11 ms /    24 runs   (  269.34
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   25679.60 ms /   100 tokens
No. of rows:   8%| | 105/1258 [59:34<9:31:54, 29.7Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.72 ms /    30 runs   (    0.26
ms per token,  3884.50 tokens per second)
llama_print_timings: prompt eval time =   22121.05 ms /    99 tokens (  223.44
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =    8028.51 ms /    29 runs   (  276.85
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   30262.22 ms /   128 tokens
No. of rows:   8%| | 106/1258 [1:00:04<9:34:16, 29Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.21 ms /    29 runs   (    0.25
ms per token,  4021.08 tokens per second)
llama_print_timings: prompt eval time =   18947.59 ms /    75 tokens (  252.63
ms per token,     3.96 tokens per second)
llama_print_timings:        eval time =    7857.41 ms /    28 runs   (  280.62
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   26921.26 ms /   103 tokens
No. of rows:   9%| | 107/1258 [1:00:31<9:16:40, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.02 ms /    31 runs   (    0.26
ms per token,  3864.37 tokens per second)
llama_print_timings: prompt eval time =   19219.05 ms /    87 tokens (  220.91
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    8245.28 ms /    30 runs   (  274.84
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   27580.94 ms /   117 tokens
No. of rows:   9%| | 108/1258 [1:00:59<9:07:58, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.88 ms /    23 runs   (    0.26
ms per token,  3914.23 tokens per second)
llama_print_timings: prompt eval time =   23731.35 ms /   104 tokens (  228.19
ms per token,     4.38 tokens per second)
llama_print_timings:        eval time =    6244.36 ms /    22 runs   (  283.83
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   30064.05 ms /   126 tokens
No. of rows:   9%| | 109/1258 [1:01:29<9:15:57, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.36 ms /    24 runs   (    0.27
ms per token,  3771.21 tokens per second)
llama_print_timings: prompt eval time =   21329.67 ms /    95 tokens (  224.52
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =    6599.23 ms /    23 runs   (  286.92
ms per token,     3.49 tokens per second)
llama_print_timings:       total time =   28021.23 ms /   118 tokens
No. of rows:   9%| | 110/1258 [1:01:57<9:09:45, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.27 ms /    28 runs   (    0.26
ms per token,  3851.97 tokens per second)
```

```
llama_print_timings: prompt eval time =   20614.44 ms /    92 tokens (  224.07
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =    7533.05 ms /    27 runs   (  279.00
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   28252.34 ms /   119 tokens
No. of rows:   9%| | 111/1258 [1:02:25<9:06:29, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.70 ms /    50 runs   (    0.25
ms per token,  3935.77 tokens per second)
llama_print_timings: prompt eval time =   20081.28 ms /    90 tokens (  223.13
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =   15361.29 ms /    49 runs   (  313.50
ms per token,     3.19 tokens per second)
llama_print_timings:       total time =   35630.68 ms /   139 tokens
No. of rows:   9%| | 112/1258 [1:03:01<9:46:27, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.03 ms /    50 runs   (    0.26
ms per token,  3836.12 tokens per second)
llama_print_timings: prompt eval time =   28076.09 ms /   123 tokens (  228.26
ms per token,     4.38 tokens per second)
llama_print_timings:        eval time =   13786.58 ms /    49 runs   (  281.36
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   42056.34 ms /   172 tokens
No. of rows:   9%| | 113/1258 [1:03:43<10:51:00, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.09 ms /    31 runs   (    0.26
ms per token,  3830.94 tokens per second)
llama_print_timings: prompt eval time =   20445.84 ms /    92 tokens (  222.24
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =    8525.71 ms /    30 runs   (  284.19
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   29087.23 ms /   122 tokens
No. of rows:   9%| | 114/1258 [1:04:12<10:21:42, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.63 ms /    50 runs   (    0.27
ms per token,  3667.84 tokens per second)
llama_print_timings: prompt eval time =   25973.53 ms /   109 tokens (  238.29
ms per token,     4.20 tokens per second)
llama_print_timings:        eval time =   13905.68 ms /    49 runs   (  283.79
ms per token,     3.52 tokens per second)
```

```
llama_print_timings:       total time =   40076.75 ms /   158 tokens
No. of rows:   9%| | 115/1258 [1:04:52<11:03:54, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.75 ms /    33 runs   (    0.27
ms per token,  3771.43 tokens per second)
llama_print_timings: prompt eval time =   20268.39 ms /    92 tokens (  220.31
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =   10690.65 ms /    32 runs   (  334.08
ms per token,     2.99 tokens per second)
llama_print_timings:       total time =   31087.75 ms /   124 tokens
No. of rows:   9%| | 116/1258 [1:05:23<10:41:53, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.88 ms /    46 runs   (    0.26
ms per token,  3870.42 tokens per second)
llama_print_timings: prompt eval time =   20904.66 ms /    93 tokens (  224.78
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   13011.54 ms /    45 runs   (  289.15
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   34091.85 ms /   138 tokens
No. of rows:   9%| | 117/1258 [1:05:57<10:43:23, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.00 ms /    35 runs   (    0.29
ms per token,  3499.65 tokens per second)
llama_print_timings: prompt eval time =   26108.07 ms /   116 tokens (  225.07
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =    9877.81 ms /    34 runs   (  290.52
ms per token,     3.44 tokens per second)
llama_print_timings:       total time =   36123.89 ms /   150 tokens
No. of rows:   9%| | 118/1258 [1:06:33<10:55:56, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.02 ms /    33 runs   (    0.27
ms per token,  3659.75 tokens per second)
llama_print_timings: prompt eval time =   20708.74 ms /    93 tokens (  222.67
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =    9142.61 ms /    32 runs   (  285.71
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   29981.25 ms /   125 tokens
No. of rows:   9%| | 119/1258 [1:07:03<10:29:31, 3Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.46 ms /    50 runs   (    0.25
ms per token,  4014.45 tokens per second)
llama_print_timings: prompt eval time =   28090.63 ms /   119 tokens (  236.06
ms per token,     4.24 tokens per second)
llama_print_timings:        eval time =   13858.55 ms /    49 runs   (  282.83
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   42143.34 ms /   168 tokens
No. of rows:  10%| | 120/1258 [1:07:45<11:20:11, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.91 ms /    30 runs   (    0.26
ms per token,  3795.07 tokens per second)
llama_print_timings: prompt eval time =   21688.51 ms /    94 tokens (  230.73
ms per token,     4.33 tokens per second)
llama_print_timings:        eval time =    9680.22 ms /    29 runs   (  333.80
ms per token,     3.00 tokens per second)
llama_print_timings:       total time =   31484.09 ms /   123 tokens
No. of rows:  10%| | 121/1258 [1:08:17<10:54:41, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.41 ms /    33 runs   (    0.25
ms per token,  3924.83 tokens per second)
llama_print_timings: prompt eval time =   21061.26 ms /    95 tokens (  221.70
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    8812.00 ms /    32 runs   (  275.37
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   29996.87 ms /   127 tokens
No. of rows:  10%| | 122/1258 [1:08:47<10:28:17, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.06 ms /    50 runs   (    0.26
ms per token,  3827.31 tokens per second)
llama_print_timings: prompt eval time =   27001.69 ms /   114 tokens (  236.86
ms per token,     4.22 tokens per second)
llama_print_timings:        eval time =   15222.87 ms /    49 runs   (  310.67
ms per token,     3.22 tokens per second)
llama_print_timings:       total time =   42427.73 ms /   163 tokens
No. of rows:  10%| | 123/1258 [1:09:29<11:20:13, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.58 ms /    29 runs   (    0.26
ms per token,  3827.37 tokens per second)
llama_print_timings: prompt eval time =   22082.50 ms /    98 tokens (  225.33
```

```
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =    7886.72 ms /    28 runs   (  281.67
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   30078.14 ms /   126 tokens
No. of rows:  10%| | 124/1258 [1:09:59<10:46:22, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.94 ms /    50 runs   (    0.26
ms per token,  3865.18 tokens per second)
llama_print_timings: prompt eval time =   27331.47 ms /   120 tokens (  227.76
ms per token,     4.39 tokens per second)
llama_print_timings:        eval time =   13642.58 ms /    49 runs   (  278.42
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   41164.63 ms /   169 tokens
No. of rows:  10%| | 125/1258 [1:10:41<11:25:19, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.88 ms /    37 runs   (    0.27
ms per token,  3746.46 tokens per second)
llama_print_timings: prompt eval time =   23692.25 ms /   106 tokens (  223.51
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   10069.82 ms /    36 runs   (  279.72
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   33905.71 ms /   142 tokens
No. of rows:  10%| | 126/1258 [1:11:15<11:11:15, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.75 ms /    46 runs   (    0.30
ms per token,  3344.72 tokens per second)
llama_print_timings: prompt eval time =   25416.41 ms /   105 tokens (  242.06
ms per token,     4.13 tokens per second)
llama_print_timings:        eval time =   13672.48 ms /    45 runs   (  303.83
ms per token,     3.29 tokens per second)
llama_print_timings:       total time =   39294.58 ms /   150 tokens
No. of rows:  10%| | 127/1258 [1:11:54<11:31:41, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.59 ms /    44 runs   (    0.26
ms per token,  3796.38 tokens per second)
llama_print_timings: prompt eval time =   31074.33 ms /   125 tokens (  248.59
ms per token,     4.02 tokens per second)
llama_print_timings:        eval time =   12140.68 ms /    43 runs   (  282.34
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   43388.61 ms /   168 tokens
```

```
No. of rows:  10%| | 128/1258 [1:12:37<12:08:54, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.84 ms /    25 runs   (    0.27
ms per token,  3657.11 tokens per second)
llama_print_timings: prompt eval time =   19353.29 ms /    89 tokens (  217.45
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    6726.92 ms /    24 runs   (  280.29
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   26179.30 ms /   113 tokens
No. of rows:  10%| | 129/1258 [1:13:03<10:57:40, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.18 ms /    37 runs   (    0.28
ms per token,  3635.65 tokens per second)
llama_print_timings: prompt eval time =   20731.61 ms /    96 tokens (  215.95
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   11467.34 ms /    36 runs   (  318.54
ms per token,     3.14 tokens per second)
llama_print_timings:       total time =   32345.33 ms /   132 tokens
No. of rows:  10%| | 130/1258 [1:13:36<10:42:26, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.56 ms /    28 runs   (    0.31
ms per token,  3269.50 tokens per second)
llama_print_timings: prompt eval time =   21252.95 ms /    95 tokens (  223.72
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =    7620.77 ms /    27 runs   (  282.25
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   28987.36 ms /   122 tokens
No. of rows:  10%| | 131/1258 [1:14:05<10:12:39, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.94 ms /    33 runs   (    0.27
ms per token,  3691.69 tokens per second)
llama_print_timings: prompt eval time =   17545.14 ms /    82 tokens (  213.97
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    8803.28 ms /    32 runs   (  275.10
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   26481.04 ms /   114 tokens
No. of rows:  10%| | 132/1258 [1:14:31<9:37:35, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =       6.66 ms /    25 runs   (    0.27
ms per token,  3755.45 tokens per second)
llama_print_timings: prompt eval time =   18312.19 ms /    85 tokens (  215.44
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    6637.48 ms /    24 runs   (  276.56
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   25049.72 ms /   109 tokens
No. of rows:  11%| | 133/1258 [1:14:56<9:04:53, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.97 ms /    35 runs   (    0.26
ms per token,  3901.46 tokens per second)
llama_print_timings: prompt eval time =   20088.09 ms /    86 tokens (  233.58
ms per token,     4.28 tokens per second)
llama_print_timings:        eval time =    9495.63 ms /    34 runs   (  279.28
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   29718.97 ms /   120 tokens
No. of rows:  11%| | 134/1258 [1:15:26<9:08:14, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.08 ms /    50 runs   (    0.28
ms per token,  3550.13 tokens per second)
llama_print_timings: prompt eval time =   32139.83 ms /   132 tokens (  243.48
ms per token,     4.11 tokens per second)
llama_print_timings:        eval time =   13806.67 ms /    49 runs   (  281.77
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   46146.32 ms /   181 tokens
No. of rows:  11%| | 135/1258 [1:16:12<10:42:33, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.77 ms /    33 runs   (    0.39
ms per token,  2584.59 tokens per second)
llama_print_timings: prompt eval time =   21225.90 ms /    96 tokens (  221.10
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    9805.23 ms /    32 runs   (  306.41
ms per token,     3.26 tokens per second)
llama_print_timings:       total time =   31175.84 ms /   128 tokens
No. of rows:  11%| | 136/1258 [1:16:43<10:24:19, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.66 ms /    25 runs   (    0.27
ms per token,  3753.75 tokens per second)
llama_print_timings: prompt eval time =   18992.60 ms /    89 tokens (  213.40
ms per token,     4.69 tokens per second)
```

```
llama_print_timings:        eval time =    6848.22 ms /    24 runs   (  285.34
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   25941.94 ms /   113 tokens
No. of rows:  11%| | 137/1258 [1:17:09<9:42:05, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.96 ms /    29 runs   (    0.27
ms per token,  3643.67 tokens per second)
llama_print_timings: prompt eval time =   18864.40 ms /    82 tokens (  230.05
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =    8397.30 ms /    28 runs   (  299.90
ms per token,     3.33 tokens per second)
llama_print_timings:       total time =   27379.14 ms /   110 tokens
No. of rows:  11%| | 138/1258 [1:17:37<9:20:25, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.08 ms /    26 runs   (    0.27
ms per token,  3672.32 tokens per second)
llama_print_timings: prompt eval time =   20640.37 ms /    92 tokens (  224.35
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =    7092.41 ms /    25 runs   (  283.70
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   27838.16 ms /   117 tokens
No. of rows:  11%| | 139/1258 [1:18:05<9:07:47, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.59 ms /    28 runs   (    0.27
ms per token,  3689.06 tokens per second)
llama_print_timings: prompt eval time =   19046.60 ms /    88 tokens (  216.44
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7356.09 ms /    27 runs   (  272.45
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   26515.79 ms /   115 tokens
No. of rows:  11%| | 140/1258 [1:18:31<8:51:22, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.62 ms /    32 runs   (    0.27
ms per token,  3710.58 tokens per second)
llama_print_timings: prompt eval time =   20263.99 ms /    87 tokens (  232.92
ms per token,     4.29 tokens per second)
llama_print_timings:        eval time =    8584.05 ms /    31 runs   (  276.90
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   28972.11 ms /   118 tokens
No. of rows:  11%| | 141/1258 [1:19:00<8:53:28, 28Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.63 ms /    36 runs   (    0.27
ms per token,  3738.71 tokens per second)
llama_print_timings: prompt eval time =   20503.05 ms /    89 tokens (  230.37
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =    9501.22 ms /    35 runs   (  271.46
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   30142.89 ms /   124 tokens
No. of rows:  11%| | 142/1258 [1:19:30<9:01:20, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.27 ms /    33 runs   (    0.28
ms per token,  3560.25 tokens per second)
llama_print_timings: prompt eval time =   20484.49 ms /    95 tokens (  215.63
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    8903.32 ms /    32 runs   (  278.23
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   29519.53 ms /   127 tokens
No. of rows:  11%| | 143/1258 [1:20:00<9:03:13, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.72 ms /    26 runs   (    0.26
ms per token,  3870.20 tokens per second)
llama_print_timings: prompt eval time =   18856.06 ms /    87 tokens (  216.74
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    6905.40 ms /    25 runs   (  276.22
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   25861.71 ms /   112 tokens
No. of rows:  11%| | 144/1258 [1:20:26<8:44:00, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.35 ms /    36 runs   (    0.26
ms per token,  3849.86 tokens per second)
llama_print_timings: prompt eval time =   23873.22 ms /   106 tokens (  225.22
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =    9506.45 ms /    35 runs   (  271.61
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   33515.23 ms /   141 tokens
No. of rows:  12%| | 145/1258 [1:20:59<9:13:00, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.22 ms /    42 runs   (    0.27
```

```
ms per token,  3743.32 tokens per second)
llama_print_timings: prompt eval time =   17898.03 ms /    84 tokens (  213.07
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   11221.56 ms /    41 runs   (  273.70
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   29285.31 ms /   125 tokens
No. of rows:  12%| | 146/1258 [1:21:28<9:09:38, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.77 ms /    36 runs   (    0.27
ms per token,  3683.24 tokens per second)
llama_print_timings: prompt eval time =   24653.30 ms /   107 tokens (  230.40
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =    9740.12 ms /    35 runs   (  278.29
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   34536.30 ms /   142 tokens
No. of rows:  12%| | 147/1258 [1:22:03<9:36:17, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.85 ms /    32 runs   (    0.28
ms per token,  3615.00 tokens per second)
llama_print_timings: prompt eval time =   23460.57 ms /   109 tokens (  215.23
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    8398.21 ms /    31 runs   (  270.91
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   31980.20 ms /   140 tokens
No. of rows:  12%| | 148/1258 [1:22:35<9:40:34, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.86 ms /    34 runs   (    0.26
ms per token,  3836.61 tokens per second)
llama_print_timings: prompt eval time =   23314.06 ms /   109 tokens (  213.89
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    8995.87 ms /    33 runs   (  272.60
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   32443.26 ms /   142 tokens
No. of rows:  12%| | 149/1258 [1:23:07<9:45:58, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.86 ms /    45 runs   (    0.26
ms per token,  3795.87 tokens per second)
llama_print_timings: prompt eval time =   24247.52 ms /   108 tokens (  224.51
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   12158.22 ms /    44 runs   (  276.32
```

```
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   36583.39 ms /   152 tokens
No. of rows:  12%| | 150/1258 [1:23:44<10:12:30, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.17 ms /    45 runs   (    0.27
ms per token,  3697.31 tokens per second)
llama_print_timings: prompt eval time =   22445.08 ms /   104 tokens (  215.82
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   12294.08 ms /    44 runs   (  279.41
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   34917.24 ms /   148 tokens
No. of rows:  12%| | 151/1258 [1:24:19<10:21:40, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.48 ms /    47 runs   (    0.27
ms per token,  3765.42 tokens per second)
llama_print_timings: prompt eval time =   21497.58 ms /   101 tokens (  212.85
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   12396.13 ms /    46 runs   (  269.48
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   34084.78 ms /   147 tokens
No. of rows:  12%| | 152/1258 [1:24:53<10:23:18, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.82 ms /    29 runs   (    0.27
ms per token,  3710.81 tokens per second)
llama_print_timings: prompt eval time =   20543.87 ms /    89 tokens (  230.83
ms per token,     4.33 tokens per second)
llama_print_timings:        eval time =    7645.11 ms /    28 runs   (  273.04
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   28301.79 ms /   117 tokens
No. of rows:  12%| | 153/1258 [1:25:21<9:52:19, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.31 ms /    50 runs   (    0.27
ms per token,  3757.70 tokens per second)
llama_print_timings: prompt eval time =   35114.25 ms /   160 tokens (  219.46
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   13665.76 ms /    49 runs   (  278.89
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   48983.32 ms /   209 tokens
No. of rows:  12%| | 154/1258 [1:26:10<11:24:41, 3Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.82 ms /    50 runs   (    0.28
ms per token,  3616.90 tokens per second)
llama_print_timings: prompt eval time =   22010.60 ms /   103 tokens (  213.70
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13364.79 ms /    49 runs   (  272.75
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   35572.53 ms /   152 tokens
No. of rows:  12%| | 155/1258 [1:26:46<11:15:04, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.71 ms /    49 runs   (    0.26
ms per token,  3855.23 tokens per second)
llama_print_timings: prompt eval time =   19798.18 ms /    92 tokens (  215.20
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   12972.78 ms /    48 runs   (  270.27
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   32959.82 ms /   140 tokens
No. of rows:  12%| | 156/1258 [1:27:19<10:53:44, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.03 ms /    36 runs   (    0.28
ms per token,  3589.23 tokens per second)
llama_print_timings: prompt eval time =   21071.92 ms /    92 tokens (  229.04
ms per token,     4.37 tokens per second)
llama_print_timings:        eval time =   11503.36 ms /    35 runs   (  328.67
ms per token,     3.04 tokens per second)
llama_print_timings:       total time =   32715.89 ms /   127 tokens
No. of rows:  12%| | 157/1258 [1:27:52<10:37:21, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.81 ms /    32 runs   (    0.28
ms per token,  3632.65 tokens per second)
llama_print_timings: prompt eval time =   23220.29 ms /   108 tokens (  215.00
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    8532.10 ms /    31 runs   (  275.23
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   31877.95 ms /   139 tokens
No. of rows:  13%| | 158/1258 [1:28:24<10:21:07, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.42 ms /    45 runs   (    0.30
ms per token,  3352.21 tokens per second)
```

```
llama_print_timings: prompt eval time =   24382.09 ms /   108 tokens (  225.76
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   13446.46 ms /    44 runs   (  305.60
ms per token,     3.27 tokens per second)
llama_print_timings:       total time =   38016.91 ms /   152 tokens
No. of rows:  13%| | 159/1258 [1:29:02<10:43:20, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.58 ms /    50 runs   (    0.27
ms per token,  3681.34 tokens per second)
llama_print_timings: prompt eval time =   29180.96 ms /   138 tokens (  211.46
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   13426.15 ms /    49 runs   (  274.00
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   42806.41 ms /   187 tokens
No. of rows:  13%| | 160/1258 [1:29:44<11:24:59, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.88 ms /    50 runs   (    0.28
ms per token,  3602.31 tokens per second)
llama_print_timings: prompt eval time =   20385.62 ms /    85 tokens (  239.83
ms per token,     4.17 tokens per second)
llama_print_timings:        eval time =   13469.21 ms /    49 runs   (  274.88
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   34052.85 ms /   134 tokens
No. of rows:  13%| | 161/1258 [1:30:18<11:05:53, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.41 ms /    28 runs   (    0.26
ms per token,  3779.19 tokens per second)
llama_print_timings: prompt eval time =   18993.86 ms /    89 tokens (  213.41
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    7307.89 ms /    27 runs   (  270.66
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   26406.91 ms /   116 tokens
No. of rows:  13%| | 162/1258 [1:30:45<10:10:26, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.33 ms /    43 runs   (    0.29
ms per token,  3488.28 tokens per second)
llama_print_timings: prompt eval time =   28344.42 ms /   127 tokens (  223.18
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =   11758.24 ms /    42 runs   (  279.96
ms per token,     3.57 tokens per second)
```

```
llama_print_timings:       total time =   40275.29 ms /   169 tokens
No. of rows:  13%| | 163/1258 [1:31:25<10:47:28, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.96 ms /    30 runs   (    0.27
ms per token,  3767.42 tokens per second)
llama_print_timings: prompt eval time =   18577.98 ms /    88 tokens (  211.11
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =    8231.08 ms /    29 runs   (  283.83
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   26930.61 ms /   117 tokens
No. of rows:  13%| | 164/1258 [1:31:52<10:00:09, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.82 ms /    50 runs   (    0.28
ms per token,  3616.90 tokens per second)
llama_print_timings: prompt eval time =   27687.79 ms /   131 tokens (  211.36
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   13484.61 ms /    49 runs   (  275.20
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   41371.76 ms /   180 tokens
No. of rows:  13%| | 165/1258 [1:32:33<10:45:51, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.94 ms /    42 runs   (    0.26
ms per token,  3840.53 tokens per second)
llama_print_timings: prompt eval time =   16406.41 ms /    77 tokens (  213.07
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   12809.84 ms /    41 runs   (  312.44
ms per token,     3.20 tokens per second)
llama_print_timings:       total time =   29377.44 ms /   118 tokens
No. of rows:  13%| | 166/1258 [1:33:03<10:12:08, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.70 ms /    33 runs   (    0.26
ms per token,  3793.10 tokens per second)
llama_print_timings: prompt eval time =   19381.08 ms /    90 tokens (  215.35
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    8609.96 ms /    32 runs   (  269.06
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   28117.19 ms /   122 tokens
No. of rows:  13%| | 167/1258 [1:33:31<9:41:31, 31Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.50 ms /    27 runs   (    0.28
ms per token,  3599.04 tokens per second)
llama_print_timings: prompt eval time =   16956.91 ms /    80 tokens (  211.96
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =    7489.47 ms /    26 runs   (  288.06
ms per token,     3.47 tokens per second)
llama_print_timings:       total time =   24554.49 ms /   106 tokens
No. of rows:  13%| | 168/1258 [1:33:55<9:00:33, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.28 ms /    50 runs   (    0.27
ms per token,  3765.63 tokens per second)
llama_print_timings: prompt eval time =   22647.58 ms /   105 tokens (  215.69
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   13366.78 ms /    49 runs   (  272.79
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   36208.08 ms /   154 tokens
No. of rows:  13%| | 169/1258 [1:34:32<9:35:13, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.97 ms /    30 runs   (    0.27
ms per token,  3764.12 tokens per second)
llama_print_timings: prompt eval time =   18172.95 ms /    85 tokens (  213.80
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    7933.13 ms /    29 runs   (  273.56
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   26223.84 ms /   114 tokens
No. of rows:  14%| | 170/1258 [1:34:58<9:05:00, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.77 ms /    21 runs   (    0.27
ms per token,  3642.04 tokens per second)
llama_print_timings: prompt eval time =   20380.09 ms /    88 tokens (  231.59
ms per token,     4.32 tokens per second)
llama_print_timings:        eval time =    5502.60 ms /    20 runs   (  275.13
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   25964.86 ms /   108 tokens
No. of rows:  14%| | 171/1258 [1:35:24<8:42:18, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.98 ms /    26 runs   (    0.27
ms per token,  3722.79 tokens per second)
llama_print_timings: prompt eval time =   19005.21 ms /    89 tokens (  213.54
```

```
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    6855.10 ms /    25 runs   (  274.20
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   25959.42 ms /   114 tokens
No. of rows:  14%| | 172/1258 [1:35:50<8:26:16, 27Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.43 ms /    43 runs   (    0.27
ms per token,  3762.36 tokens per second)
llama_print_timings: prompt eval time =   28713.53 ms /   126 tokens (  227.89
ms per token,     4.39 tokens per second)
llama_print_timings:        eval time =   11416.72 ms /    42 runs   (  271.83
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   40296.49 ms /   168 tokens
No. of rows:  14%| | 173/1258 [1:36:30<9:32:43, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.35 ms /    35 runs   (    0.27
ms per token,  3742.11 tokens per second)
llama_print_timings: prompt eval time =   23512.32 ms /   110 tokens (  213.75
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    9138.06 ms /    34 runs   (  268.77
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   32786.97 ms /   144 tokens
No. of rows:  14%| | 174/1258 [1:37:03<9:38:16, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.92 ms /    38 runs   (    0.26
ms per token,  3831.03 tokens per second)
llama_print_timings: prompt eval time =   23517.06 ms /   101 tokens (  232.84
ms per token,     4.29 tokens per second)
llama_print_timings:        eval time =   11722.32 ms /    37 runs   (  316.82
ms per token,     3.16 tokens per second)
llama_print_timings:       total time =   35385.25 ms /   138 tokens
No. of rows:  14%| | 175/1258 [1:37:38<9:56:04, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.12 ms /    48 runs   (    0.27
ms per token,  3659.09 tokens per second)
llama_print_timings: prompt eval time =   24255.10 ms /   112 tokens (  216.56
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   13408.08 ms /    47 runs   (  285.28
ms per token,     3.51 tokens per second)
llama_print_timings:       total time =   37860.63 ms /   159 tokens
```

```
No. of rows:  14%| | 176/1258 [1:38:16<10:21:44, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.99 ms /    38 runs   (    0.26
ms per token,  3804.57 tokens per second)
llama_print_timings: prompt eval time =   23729.38 ms /   112 tokens (  211.87
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   11788.73 ms /    37 runs   (  318.61
ms per token,     3.14 tokens per second)
llama_print_timings:       total time =   35662.45 ms /   149 tokens
No. of rows:  14%| | 177/1258 [1:38:52<10:27:33, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.32 ms /    50 runs   (    0.27
ms per token,  3754.60 tokens per second)
llama_print_timings: prompt eval time =   25616.70 ms /   123 tokens (  208.27
ms per token,     4.80 tokens per second)
llama_print_timings:        eval time =   13407.33 ms /    49 runs   (  273.62
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   39220.14 ms /   172 tokens
No. of rows:  14%| | 178/1258 [1:39:31<10:50:46, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.19 ms /    49 runs   (    0.27
ms per token,  3716.34 tokens per second)
llama_print_timings: prompt eval time =   28077.44 ms /   134 tokens (  209.53
ms per token,     4.77 tokens per second)
llama_print_timings:        eval time =   13629.44 ms /    48 runs   (  283.95
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   41903.32 ms /   182 tokens
No. of rows:  14%| | 179/1258 [1:40:13<11:21:14, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.96 ms /    30 runs   (    0.27
ms per token,  3771.21 tokens per second)
llama_print_timings: prompt eval time =   18148.59 ms /    83 tokens (  218.66
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    9641.29 ms /    29 runs   (  332.46
ms per token,     3.01 tokens per second)
llama_print_timings:       total time =   27908.05 ms /   112 tokens
No. of rows:  14%| | 180/1258 [1:40:41<10:26:53, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      11.79 ms /    45 runs   (    0.26
ms per token,  3816.15 tokens per second)
llama_print_timings: prompt eval time =   21988.96 ms /   104 tokens (  211.43
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   11951.78 ms /    44 runs   (  271.63
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   34119.92 ms /   148 tokens
No. of rows:  14%| | 181/1258 [1:41:15<10:22:11, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.33 ms /    35 runs   (    0.27
ms per token,  3749.73 tokens per second)
llama_print_timings: prompt eval time =   21658.71 ms /   102 tokens (  212.34
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   10954.46 ms /    34 runs   (  322.19
ms per token,     3.10 tokens per second)
llama_print_timings:       total time =   32749.07 ms /   136 tokens
No. of rows:  14%| | 182/1258 [1:41:48<10:11:21, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.25 ms /    50 runs   (    0.27
ms per token,  3772.73 tokens per second)
llama_print_timings: prompt eval time =   23433.46 ms /   108 tokens (  216.98
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   13406.81 ms /    49 runs   (  273.61
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   37032.01 ms /   157 tokens
No. of rows:  15%| | 183/1258 [1:42:25<10:26:38, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.00 ms /    26 runs   (    0.27
ms per token,  3713.22 tokens per second)
llama_print_timings: prompt eval time =   21997.25 ms /   103 tokens (  213.57
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    6937.22 ms /    25 runs   (  277.49
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   29041.57 ms /   128 tokens
No. of rows:  15%| | 184/1258 [1:42:54<9:54:14, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.59 ms /    32 runs   (    0.27
ms per token,  3726.56 tokens per second)
llama_print_timings: prompt eval time =   22455.77 ms /    99 tokens (  226.83
ms per token,     4.41 tokens per second)
```

```
llama_print_timings:        eval time =    8596.84 ms /    31 runs   (  277.32
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   31180.97 ms /   130 tokens
No. of rows:  15%| | 185/1258 [1:43:25<9:42:53, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.42 ms /    48 runs   (    0.28
ms per token,  3577.28 tokens per second)
llama_print_timings: prompt eval time =   24590.48 ms /   108 tokens (  227.69
ms per token,     4.39 tokens per second)
llama_print_timings:        eval time =   15056.49 ms /    47 runs   (  320.35
ms per token,     3.12 tokens per second)
llama_print_timings:       total time =   39843.70 ms /   155 tokens
No. of rows:  15%| | 186/1258 [1:44:05<10:21:17, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.07 ms /    38 runs   (    0.27
ms per token,  3772.09 tokens per second)
llama_print_timings: prompt eval time =   21926.99 ms /   102 tokens (  214.97
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   10143.10 ms /    37 runs   (  274.14
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   32216.24 ms /   139 tokens
No. of rows:  15%| | 187/1258 [1:44:37<10:07:03, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.97 ms /    30 runs   (    0.27
ms per token,  3763.17 tokens per second)
llama_print_timings: prompt eval time =   20427.79 ms /    97 tokens (  210.60
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =    7840.22 ms /    29 runs   (  270.35
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   28383.39 ms /   126 tokens
No. of rows:  15%| | 188/1258 [1:45:06<9:36:25, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.46 ms /    28 runs   (    0.27
ms per token,  3754.36 tokens per second)
llama_print_timings: prompt eval time =   21469.13 ms /    99 tokens (  216.86
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    7323.15 ms /    27 runs   (  271.23
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   28902.00 ms /   126 tokens
No. of rows:  15%| | 189/1258 [1:45:35<9:17:37, 31Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.11 ms /    38 runs   (    0.27
ms per token,  3760.14 tokens per second)
llama_print_timings: prompt eval time =   23159.93 ms /   109 tokens (  212.48
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   10437.70 ms /    37 runs   (  282.10
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   33746.79 ms /   146 tokens
No. of rows:  15%| | 190/1258 [1:46:08<9:30:12, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.27 ms /    50 runs   (    0.27
ms per token,  3767.61 tokens per second)
llama_print_timings: prompt eval time =   27086.04 ms /   126 tokens (  214.97
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13194.92 ms /    49 runs   (  269.28
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   40476.89 ms /   175 tokens
No. of rows:  15%| | 191/1258 [1:46:49<10:14:45, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.79 ms /    31 runs   (    0.25
ms per token,  3978.95 tokens per second)
llama_print_timings: prompt eval time =   19384.96 ms /    92 tokens (  210.71
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =    8106.04 ms /    30 runs   (  270.20
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   27608.93 ms /   122 tokens
No. of rows:  15%| | 192/1258 [1:47:16<9:37:07, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.30 ms /    48 runs   (    0.26
ms per token,  3901.80 tokens per second)
llama_print_timings: prompt eval time =   24597.20 ms /   107 tokens (  229.88
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =   12808.07 ms /    47 runs   (  272.51
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   37589.28 ms /   154 tokens
No. of rows:  15%| | 193/1258 [1:47:54<10:03:48, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.29 ms /    28 runs   (    0.26
```

```
ms per token,  3839.30 tokens per second)
llama_print_timings: prompt eval time =   22031.23 ms /    94 tokens (  234.37
ms per token,     4.27 tokens per second)
llama_print_timings:        eval time =    7323.82 ms /    27 runs   (  271.25
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   29461.95 ms /   121 tokens
No. of rows:  15%| | 194/1258 [1:48:23<9:39:02, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.76 ms /    50 runs   (    0.28
ms per token,  3634.51 tokens per second)
llama_print_timings: prompt eval time =   29653.23 ms /   142 tokens (  208.83
ms per token,     4.79 tokens per second)
llama_print_timings:        eval time =   13384.03 ms /    49 runs   (  273.14
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   43237.08 ms /   191 tokens
No. of rows:  16%| | 195/1258 [1:49:07<10:34:48, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.98 ms /    50 runs   (    0.26
ms per token,  3852.08 tokens per second)
llama_print_timings: prompt eval time =   25976.27 ms /   116 tokens (  223.93
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   13256.13 ms /    49 runs   (  270.53
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   39431.60 ms /   165 tokens
No. of rows:  16%| | 196/1258 [1:49:46<10:53:21, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.43 ms /    50 runs   (    0.27
ms per token,  3724.39 tokens per second)
llama_print_timings: prompt eval time =   26195.40 ms /   115 tokens (  227.79
ms per token,     4.39 tokens per second)
llama_print_timings:        eval time =   13593.13 ms /    49 runs   (  277.41
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   39988.27 ms /   164 tokens
No. of rows:  16%| | 197/1258 [1:50:26<11:09:02, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.65 ms /    29 runs   (    0.26
ms per token,  3793.33 tokens per second)
llama_print_timings: prompt eval time =   19098.91 ms /    88 tokens (  217.03
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    7544.98 ms /    28 runs   (  269.46
```

```
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   26754.02 ms /   116 tokens
No. of rows:  16%| | 198/1258 [1:50:53<10:09:46, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.08 ms /    27 runs   (    0.26
ms per token,  3814.64 tokens per second)
llama_print_timings: prompt eval time =   18603.04 ms /    88 tokens (  211.40
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =    7196.59 ms /    26 runs   (  276.79
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   25905.02 ms /   114 tokens
No. of rows:  16%| | 199/1258 [1:51:19<9:23:38, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.54 ms /    38 runs   (    0.28
ms per token,  3603.60 tokens per second)
llama_print_timings: prompt eval time =   22243.81 ms /    97 tokens (  229.32
ms per token,     4.36 tokens per second)
llama_print_timings:        eval time =   11862.22 ms /    37 runs   (  320.60
ms per token,     3.12 tokens per second)
llama_print_timings:       total time =   34254.56 ms /   134 tokens
No. of rows:  16%| | 200/1258 [1:51:53<9:35:25, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      16.65 ms /    50 runs   (    0.33
ms per token,  3003.54 tokens per second)
llama_print_timings: prompt eval time =   25887.03 ms /   118 tokens (  219.38
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   13680.10 ms /    49 runs   (  279.19
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   39771.01 ms /   167 tokens
No. of rows:  16%| | 201/1258 [1:52:33<10:12:38, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.04 ms /    35 runs   (    0.26
ms per token,  3871.25 tokens per second)
llama_print_timings: prompt eval time =   20551.17 ms /    97 tokens (  211.87
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   10889.40 ms /    34 runs   (  320.28
ms per token,     3.12 tokens per second)
llama_print_timings:       total time =   31574.22 ms /   131 tokens
No. of rows:  16%| | 202/1258 [1:53:04<9:55:11, 33Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.96 ms /    31 runs   (    0.26
ms per token,  3894.96 tokens per second)
llama_print_timings: prompt eval time =   20085.89 ms /    94 tokens (  213.68
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    9764.27 ms /    30 runs   (  325.48
ms per token,     3.07 tokens per second)
llama_print_timings:       total time =   29966.65 ms /   124 tokens
No. of rows:  16%| | 203/1258 [1:53:34<9:34:21, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.81 ms /    34 runs   (    0.26
ms per token,  3857.94 tokens per second)
llama_print_timings: prompt eval time =   22945.00 ms /   109 tokens (  210.50
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =    9534.98 ms /    33 runs   (  288.94
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   32619.73 ms /   142 tokens
No. of rows:  16%| | 204/1258 [1:54:07<9:33:36, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.16 ms /    34 runs   (    0.27
ms per token,  3713.01 tokens per second)
llama_print_timings: prompt eval time =   23202.95 ms /   108 tokens (  214.84
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    8946.63 ms /    33 runs   (  271.11
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   32282.10 ms /   141 tokens
No. of rows:  16%| | 205/1258 [1:54:39<9:31:11, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.02 ms /    44 runs   (    0.27
ms per token,  3661.17 tokens per second)
llama_print_timings: prompt eval time =   25056.97 ms /   118 tokens (  212.35
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   11789.91 ms /    43 runs   (  274.18
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   37020.32 ms /   161 tokens
No. of rows:  16%| | 206/1258 [1:55:16<9:54:09, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.16 ms /    50 runs   (    0.26
ms per token,  3798.81 tokens per second)
```

```
llama_print_timings: prompt eval time =   20994.78 ms /    98 tokens (  214.23
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13248.72 ms /    49 runs   (  270.38
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   34441.85 ms /   147 tokens
No. of rows:  16%| | 207/1258 [1:55:51<9:56:36, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.17 ms /    27 runs   (    0.27
ms per token,  3763.59 tokens per second)
llama_print_timings: prompt eval time =   20567.32 ms /    94 tokens (  218.80
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    7148.76 ms /    26 runs   (  274.95
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   27821.23 ms /   120 tokens
No. of rows:  17%| | 208/1258 [1:56:19<9:23:19, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.16 ms /    41 runs   (    0.27
ms per token,  3672.52 tokens per second)
llama_print_timings: prompt eval time =   21586.64 ms /   100 tokens (  215.87
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   12665.75 ms /    40 runs   (  316.64
ms per token,     3.16 tokens per second)
llama_print_timings:       total time =   34413.87 ms /   140 tokens
No. of rows:  17%| | 209/1258 [1:56:53<9:34:29, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.24 ms /    50 runs   (    0.26
ms per token,  3776.72 tokens per second)
llama_print_timings: prompt eval time =   30219.06 ms /   144 tokens (  209.85
ms per token,     4.77 tokens per second)
llama_print_timings:        eval time =   13299.27 ms /    49 runs   (  271.41
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   43712.87 ms /   193 tokens
No. of rows:  17%| | 210/1258 [1:57:37<10:30:51, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.42 ms /    32 runs   (    0.26
ms per token,  3800.48 tokens per second)
llama_print_timings: prompt eval time =   19664.22 ms /    93 tokens (  211.44
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =    8385.93 ms /    31 runs   (  270.51
ms per token,     3.70 tokens per second)
```

```
llama_print_timings:       total time =   28174.74 ms /   124 tokens
No. of rows:  17%| | 211/1258 [1:58:05<9:48:38, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.33 ms /    30 runs   (    0.28
ms per token,  3599.71 tokens per second)
llama_print_timings: prompt eval time =   18088.24 ms /    83 tokens (  217.93
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    7913.72 ms /    29 runs   (  272.89
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   26124.08 ms /   112 tokens
No. of rows:  17%| | 212/1258 [1:58:31<9:08:20, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.74 ms /    50 runs   (    0.27
ms per token,  3640.07 tokens per second)
llama_print_timings: prompt eval time =   22789.65 ms /   106 tokens (  215.00
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13569.29 ms /    49 runs   (  276.92
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   36554.70 ms /   155 tokens
No. of rows:  17%| | 213/1258 [1:59:08<9:34:33, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.54 ms /    50 runs   (    0.27
ms per token,  3692.49 tokens per second)
llama_print_timings: prompt eval time =   26978.51 ms /   120 tokens (  224.82
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   15099.44 ms /    49 runs   (  308.15
ms per token,     3.25 tokens per second)
llama_print_timings:       total time =   42273.90 ms /   169 tokens
No. of rows:  17%| | 214/1258 [1:59:50<10:22:31, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.50 ms /    50 runs   (    0.27
ms per token,  3704.25 tokens per second)
llama_print_timings: prompt eval time =   35172.52 ms /   161 tokens (  218.46
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =   13427.56 ms /    49 runs   (  274.03
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   48803.20 ms /   210 tokens
No. of rows:  17%| | 215/1258 [2:00:39<11:29:49, 3Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.33 ms /    50 runs   (    0.27
ms per token,  3750.38 tokens per second)
llama_print_timings: prompt eval time =   26003.39 ms /   124 tokens (  209.70
ms per token,     4.77 tokens per second)
llama_print_timings:        eval time =   13317.40 ms /    49 runs   (  271.78
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   39515.58 ms /   173 tokens
No. of rows:  17%| | 216/1258 [2:01:18<11:28:23, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.00 ms /    43 runs   (    0.28
ms per token,  3582.44 tokens per second)
llama_print_timings: prompt eval time =   21665.84 ms /   102 tokens (  212.41
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   13232.30 ms /    42 runs   (  315.05
ms per token,     3.17 tokens per second)
llama_print_timings:       total time =   35071.67 ms /   144 tokens
No. of rows:  17%| | 217/1258 [2:01:53<11:03:59, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.99 ms /    45 runs   (    0.27
ms per token,  3753.75 tokens per second)
llama_print_timings: prompt eval time =   24638.48 ms /   114 tokens (  216.13
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   12041.79 ms /    44 runs   (  273.68
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   36861.77 ms /   158 tokens
No. of rows:  17%| | 218/1258 [2:02:30<10:56:04, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.37 ms /    32 runs   (    0.26
ms per token,  3821.35 tokens per second)
llama_print_timings: prompt eval time =   19029.30 ms /    89 tokens (  213.81
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    8407.31 ms /    31 runs   (  271.20
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   27561.41 ms /   120 tokens
No. of rows:  17%| | 219/1258 [2:02:58<10:02:01, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.51 ms /    50 runs   (    0.27
ms per token,  3702.06 tokens per second)
llama_print_timings: prompt eval time =   24555.79 ms /   116 tokens (  211.69
```

```
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   15221.18 ms /    49 runs   (  310.64
ms per token,     3.22 tokens per second)
llama_print_timings:       total time =   39978.99 ms /   165 tokens
No. of rows:  17%| | 220/1258 [2:03:38<10:28:32, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.30 ms /    45 runs   (    0.27
ms per token,  3659.73 tokens per second)
llama_print_timings: prompt eval time =   27681.92 ms /   130 tokens (  212.94
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   12811.62 ms /    44 runs   (  291.17
ms per token,     3.43 tokens per second)
llama_print_timings:       total time =   40682.27 ms /   174 tokens
No. of rows:  18%| | 221/1258 [2:04:18<10:50:31, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.80 ms /    43 runs   (    0.27
ms per token,  3642.83 tokens per second)
llama_print_timings: prompt eval time =   24358.84 ms /   115 tokens (  211.82
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   11520.68 ms /    42 runs   (  274.30
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   36051.24 ms /   157 tokens
No. of rows:  18%| | 222/1258 [2:04:54<10:41:43, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.12 ms /    23 runs   (    0.27
ms per token,  3760.63 tokens per second)
llama_print_timings: prompt eval time =   16934.05 ms /    78 tokens (  217.10
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    6056.86 ms /    22 runs   (  275.31
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   23082.67 ms /   100 tokens
No. of rows:  18%| | 223/1258 [2:05:18<9:28:15, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.56 ms /    50 runs   (    0.27
ms per token,  3687.04 tokens per second)
llama_print_timings: prompt eval time =   27509.20 ms /   123 tokens (  223.65
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   13220.74 ms /    49 runs   (  269.81
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   40928.28 ms /   172 tokens
```

```
No. of rows:  18%| | 224/1258 [2:05:59<10:09:02, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.36 ms /    26 runs   (    0.28
ms per token,  3533.57 tokens per second)
llama_print_timings: prompt eval time =   20278.70 ms /    86 tokens (  235.80
ms per token,     4.24 tokens per second)
llama_print_timings:        eval time =    6906.41 ms /    25 runs   (  276.26
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   27288.18 ms /   111 tokens
No. of rows:  18%| | 225/1258 [2:06:26<9:26:53, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.24 ms /    23 runs   (    0.27
ms per token,  3688.26 tokens per second)
llama_print_timings: prompt eval time =   17486.74 ms /    81 tokens (  215.89
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    5830.49 ms /    22 runs   (  265.02
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   23407.06 ms /   103 tokens
No. of rows:  18%| | 226/1258 [2:06:49<8:37:15, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.24 ms /    24 runs   (    0.26
ms per token,  3843.69 tokens per second)
llama_print_timings: prompt eval time =   18567.73 ms /    88 tokens (  211.00
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =    6200.04 ms /    23 runs   (  269.57
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   24861.08 ms /   111 tokens
No. of rows:  18%| | 227/1258 [2:07:14<8:09:55, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.92 ms /    50 runs   (    0.26
ms per token,  3870.57 tokens per second)
llama_print_timings: prompt eval time =   31050.15 ms /   141 tokens (  220.21
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =   13264.94 ms /    49 runs   (  270.71
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   44511.62 ms /   190 tokens
No. of rows:  18%| | 228/1258 [2:07:59<9:31:53, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =       6.77 ms /    23 runs   (    0.29
ms per token,  3396.34 tokens per second)
llama_print_timings: prompt eval time =   17986.99 ms /    75 tokens (  239.83
ms per token,     4.17 tokens per second)
llama_print_timings:        eval time =    6407.40 ms /    22 runs   (  291.25
ms per token,     3.43 tokens per second)
llama_print_timings:       total time =   24495.16 ms /    97 tokens
No. of rows:  18%| | 229/1258 [2:08:23<8:46:00, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.95 ms /    29 runs   (    0.27
ms per token,  3646.42 tokens per second)
llama_print_timings: prompt eval time =   16300.77 ms /    77 tokens (  211.70
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =    7543.17 ms /    28 runs   (  269.40
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   23957.80 ms /   105 tokens
No. of rows:  18%| | 230/1258 [2:08:47<8:11:01, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.10 ms /    27 runs   (    0.34
ms per token,  2967.36 tokens per second)
llama_print_timings: prompt eval time =   19640.69 ms /    92 tokens (  213.49
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    7718.87 ms /    26 runs   (  296.88
ms per token,     3.37 tokens per second)
llama_print_timings:       total time =   27478.92 ms /   118 tokens
No. of rows:  18%| | 231/1258 [2:09:15<8:04:29, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.66 ms /    39 runs   (    0.27
ms per token,  3657.85 tokens per second)
llama_print_timings: prompt eval time =   23669.37 ms /   104 tokens (  227.59
ms per token,     4.39 tokens per second)
llama_print_timings:        eval time =   10450.97 ms /    38 runs   (  275.03
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   34273.17 ms /   142 tokens
No. of rows:  18%| | 232/1258 [2:09:49<8:34:42, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.94 ms /    36 runs   (    0.28
ms per token,  3620.64 tokens per second)
llama_print_timings: prompt eval time =   21831.97 ms /    94 tokens (  232.25
ms per token,     4.31 tokens per second)
```

```
llama_print_timings:        eval time =   10900.69 ms /    35 runs   (  311.45
ms per token,     3.21 tokens per second)
llama_print_timings:       total time =   32885.63 ms /   129 tokens
No. of rows:  19%| | 233/1258 [2:10:22<8:48:33, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.10 ms /    22 runs   (    0.28
ms per token,  3604.78 tokens per second)
llama_print_timings: prompt eval time =   18842.19 ms /    86 tokens (  219.10
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    5822.42 ms /    21 runs   (  277.26
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   24749.04 ms /   107 tokens
No. of rows:  19%| | 234/1258 [2:10:46<8:16:22, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.51 ms /    34 runs   (    0.28
ms per token,  3575.94 tokens per second)
llama_print_timings: prompt eval time =   20263.17 ms /    95 tokens (  213.30
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    9053.91 ms /    33 runs   (  274.36
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   29450.68 ms /   128 tokens
No. of rows:  19%| | 235/1258 [2:11:16<8:17:48, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.39 ms /    42 runs   (    0.27
ms per token,  3687.12 tokens per second)
llama_print_timings: prompt eval time =   27384.71 ms /   131 tokens (  209.04
ms per token,     4.78 tokens per second)
llama_print_timings:        eval time =   11303.21 ms /    41 runs   (  275.69
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   38855.53 ms /   172 tokens
No. of rows:  19%| | 236/1258 [2:11:55<9:06:42, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.75 ms /    28 runs   (    0.28
ms per token,  3612.90 tokens per second)
llama_print_timings: prompt eval time =   21518.53 ms /   100 tokens (  215.19
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    7596.11 ms /    27 runs   (  281.34
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   29228.05 ms /   127 tokens
No. of rows:  19%| | 237/1258 [2:12:24<8:51:32, 31Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.18 ms /    27 runs   (    0.27
ms per token,  3762.54 tokens per second)
llama_print_timings: prompt eval time =   18199.80 ms /    85 tokens (  214.12
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    7000.08 ms /    26 runs   (  269.23
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   25302.42 ms /   111 tokens
No. of rows:  19%| | 238/1258 [2:12:49<8:20:48, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.02 ms /    50 runs   (    0.26
ms per token,  3840.84 tokens per second)
llama_print_timings: prompt eval time =   26560.78 ms /   126 tokens (  210.80
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =   14922.70 ms /    49 runs   (  304.54
ms per token,     3.28 tokens per second)
llama_print_timings:       total time =   41678.89 ms /   175 tokens
No. of rows:  19%| | 239/1258 [2:13:31<9:22:37, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.57 ms /    32 runs   (    0.27
ms per token,  3733.08 tokens per second)
llama_print_timings: prompt eval time =   18866.56 ms /    86 tokens (  219.38
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    8351.07 ms /    31 runs   (  269.39
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   27344.70 ms /   117 tokens
No. of rows:  19%| | 240/1258 [2:13:58<8:52:40, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.66 ms /    50 runs   (    0.25
ms per token,  3951.01 tokens per second)
llama_print_timings: prompt eval time =   28535.89 ms /   132 tokens (  216.18
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   13350.96 ms /    49 runs   (  272.47
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   42081.00 ms /   181 tokens
No. of rows:  19%| | 241/1258 [2:14:40<9:46:31, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.16 ms /    24 runs   (    0.26
```

```
ms per token,  3898.00 tokens per second)
llama_print_timings: prompt eval time =   16267.32 ms /    74 tokens (  219.83
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =    6271.17 ms /    23 runs   (  272.66
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   22630.57 ms /    97 tokens
No. of rows:  19%| | 242/1258 [2:15:03<8:45:09, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.39 ms /    46 runs   (    0.27
ms per token,  3711.17 tokens per second)
llama_print_timings: prompt eval time =   25448.97 ms /   111 tokens (  229.27
ms per token,     4.36 tokens per second)
llama_print_timings:        eval time =   12152.79 ms /    45 runs   (  270.06
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   37781.98 ms /   156 tokens
No. of rows:  19%| | 243/1258 [2:15:41<9:19:01, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.80 ms /    50 runs   (    0.28
ms per token,  3622.14 tokens per second)
llama_print_timings: prompt eval time =   22600.01 ms /    98 tokens (  230.61
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =   13792.14 ms /    49 runs   (  281.47
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   36592.13 ms /   147 tokens
No. of rows:  19%| | 244/1258 [2:16:18<9:36:29, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.42 ms /    50 runs   (    0.27
ms per token,  3724.95 tokens per second)
llama_print_timings: prompt eval time =   27959.97 ms /   121 tokens (  231.07
ms per token,     4.33 tokens per second)
llama_print_timings:        eval time =   13486.78 ms /    49 runs   (  275.24
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   41646.89 ms /   170 tokens
No. of rows:  19%| | 245/1258 [2:16:59<10:14:07, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.30 ms /    23 runs   (    0.27
ms per token,  3653.11 tokens per second)
llama_print_timings: prompt eval time =   25049.99 ms /   119 tokens (  210.50
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =    6015.05 ms /    22 runs   (  273.41
```

```
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   31155.33 ms /   141 tokens
No. of rows:  20%| | 246/1258 [2:17:30<9:47:08, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.59 ms /    50 runs   (    0.27
ms per token,  3679.99 tokens per second)
llama_print_timings: prompt eval time =   23545.59 ms /   110 tokens (  214.05
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13488.22 ms /    49 runs   (  275.27
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   37236.57 ms /   159 tokens
No. of rows:  20%| | 247/1258 [2:18:08<9:58:51, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.47 ms /    30 runs   (    0.28
ms per token,  3541.49 tokens per second)
llama_print_timings: prompt eval time =   19203.31 ms /    80 tokens (  240.04
ms per token,     4.17 tokens per second)
llama_print_timings:        eval time =    8381.42 ms /    29 runs   (  289.01
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   27710.88 ms /   109 tokens
No. of rows:  20%| | 248/1258 [2:18:35<9:18:46, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.65 ms /    50 runs   (    0.27
ms per token,  3661.93 tokens per second)
llama_print_timings: prompt eval time =   26482.25 ms /   125 tokens (  211.86
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   13400.70 ms /    49 runs   (  273.48
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   40078.54 ms /   174 tokens
No. of rows:  20%| | 249/1258 [2:19:15<9:52:58, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.74 ms /    29 runs   (    0.27
ms per token,  3746.29 tokens per second)
llama_print_timings: prompt eval time =   20405.48 ms /    95 tokens (  214.79
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    7571.72 ms /    28 runs   (  270.42
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   28090.86 ms /   123 tokens
No. of rows:  20%| | 250/1258 [2:19:43<9:16:16, 33Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.27 ms /    19 runs   (    0.28
ms per token,  3606.68 tokens per second)
llama_print_timings: prompt eval time =   18743.23 ms /    79 tokens (  237.26
ms per token,     4.21 tokens per second)
llama_print_timings:        eval time =    4801.20 ms /    18 runs   (  266.73
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   23619.02 ms /    97 tokens
No. of rows:  20%| | 251/1258 [2:20:07<8:27:58, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.85 ms /    40 runs   (    0.27
ms per token,  3685.96 tokens per second)
llama_print_timings: prompt eval time =   34139.05 ms /   151 tokens (  226.09
ms per token,     4.42 tokens per second)
llama_print_timings:        eval time =   10727.95 ms /    39 runs   (  275.08
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   45026.84 ms /   190 tokens
No. of rows:  20%| | 252/1258 [2:20:52<9:41:44, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.22 ms /    39 runs   (    0.26
ms per token,  3816.79 tokens per second)
llama_print_timings: prompt eval time =   21201.92 ms /    98 tokens (  216.35
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   10444.02 ms /    38 runs   (  274.84
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   31796.23 ms /   136 tokens
No. of rows:  20%| | 253/1258 [2:21:24<9:26:34, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.34 ms /    27 runs   (    0.27
ms per token,  3676.47 tokens per second)
llama_print_timings: prompt eval time =   19800.24 ms /    94 tokens (  210.64
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =    7174.41 ms /    26 runs   (  275.94
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   27081.87 ms /   120 tokens
No. of rows:  20%| | 254/1258 [2:21:51<8:52:14, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.67 ms /    48 runs   (    0.26
ms per token,  3787.28 tokens per second)
```

```
llama_print_timings: prompt eval time =   24482.78 ms /   113 tokens (  216.66
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   12872.69 ms /    47 runs   (  273.89
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   37545.13 ms /   160 tokens
No. of rows:  20%| | 255/1258 [2:22:29<9:20:31, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.51 ms /    44 runs   (    0.33
ms per token,  3033.02 tokens per second)
llama_print_timings: prompt eval time =   24011.72 ms /   104 tokens (  230.88
ms per token,     4.33 tokens per second)
llama_print_timings:        eval time =   11682.47 ms /    43 runs   (  271.69
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   35869.08 ms /   147 tokens
No. of rows:  20%| | 256/1258 [2:23:04<9:31:42, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.37 ms /    49 runs   (    0.27
ms per token,  3664.37 tokens per second)
llama_print_timings: prompt eval time =   23425.62 ms /   110 tokens (  212.96
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   13082.24 ms /    48 runs   (  272.55
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   36700.91 ms /   158 tokens
No. of rows:  20%| | 257/1258 [2:23:41<9:43:31, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.74 ms /    40 runs   (    0.27
ms per token,  3723.01 tokens per second)
llama_print_timings: prompt eval time =   24945.00 ms /   116 tokens (  215.04
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   11223.92 ms /    39 runs   (  287.79
ms per token,     3.47 tokens per second)
llama_print_timings:       total time =   36328.39 ms /   155 tokens
No. of rows:  21%| | 258/1258 [2:24:17<9:49:42, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.92 ms /    50 runs   (    0.28
ms per token,  3591.18 tokens per second)
llama_print_timings: prompt eval time =   29592.72 ms /   131 tokens (  225.90
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   13741.87 ms /    49 runs   (  280.45
ms per token,     3.57 tokens per second)
```

```
llama_print_timings:       total time =   43540.77 ms /   180 tokens
No. of rows:  21%| | 259/1258 [2:25:01<10:29:56, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.49 ms /    43 runs   (    0.27
ms per token,  3743.04 tokens per second)
llama_print_timings: prompt eval time =   18831.19 ms /    85 tokens (  221.54
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =   11504.36 ms /    42 runs   (  273.91
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   30510.31 ms /   127 tokens
No. of rows:  21%| | 260/1258 [2:25:32<9:52:47, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.43 ms /    30 runs   (    0.28
ms per token,  3560.83 tokens per second)
llama_print_timings: prompt eval time =   18265.02 ms /    85 tokens (  214.88
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    7974.69 ms /    29 runs   (  274.99
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   26358.51 ms /   114 tokens
No. of rows:  21%| | 261/1258 [2:25:58<9:05:58, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.81 ms /    50 runs   (    0.28
ms per token,  3620.56 tokens per second)
llama_print_timings: prompt eval time =   22662.27 ms /   105 tokens (  215.83
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   13571.33 ms /    49 runs   (  276.97
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   36436.31 ms /   154 tokens
No. of rows:  21%| | 262/1258 [2:26:34<9:23:16, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.83 ms /    48 runs   (    0.27
ms per token,  3742.69 tokens per second)
llama_print_timings: prompt eval time =   27910.81 ms /   130 tokens (  214.70
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   12941.84 ms /    47 runs   (  275.36
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   41041.42 ms /   177 tokens
No. of rows:  21%| | 263/1258 [2:27:15<9:58:07, 36Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.23 ms /    26 runs   (    0.28
ms per token,  3596.13 tokens per second)
llama_print_timings: prompt eval time =   18966.86 ms /    88 tokens (  215.53
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    6801.93 ms /    25 runs   (  272.08
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   25872.73 ms /   113 tokens
No. of rows:  21%| | 264/1258 [2:27:41<9:06:53, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.78 ms /    36 runs   (    0.27
ms per token,  3679.48 tokens per second)
llama_print_timings: prompt eval time =   25680.51 ms /   111 tokens (  231.36
ms per token,     4.32 tokens per second)
llama_print_timings:        eval time =   11532.97 ms /    35 runs   (  329.51
ms per token,     3.03 tokens per second)
llama_print_timings:       total time =   37359.33 ms /   146 tokens
No. of rows:  21%| | 265/1258 [2:28:19<9:27:55, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.16 ms /    50 runs   (    0.28
ms per token,  3530.08 tokens per second)
llama_print_timings: prompt eval time =   24304.09 ms /   113 tokens (  215.08
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13510.12 ms /    49 runs   (  275.72
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   38017.30 ms /   162 tokens
No. of rows:  21%| | 266/1258 [2:28:57<9:45:46, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.80 ms /    43 runs   (    0.30
ms per token,  3359.38 tokens per second)
llama_print_timings: prompt eval time =   19142.44 ms /    87 tokens (  220.03
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =   11487.81 ms /    42 runs   (  273.52
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   30804.13 ms /   129 tokens
No. of rows:  21%| | 267/1258 [2:29:28<9:22:18, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.08 ms /    45 runs   (    0.27
ms per token,  3724.55 tokens per second)
llama_print_timings: prompt eval time =   20905.22 ms /    90 tokens (  232.28
```

```
ms per token,     4.31 tokens per second)
llama_print_timings:        eval time =   12045.64 ms /    44 runs   (  273.76
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   33126.30 ms /   134 tokens
No. of rows:  21%| | 268/1258 [2:30:01<9:17:13, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.53 ms /    24 runs   (    0.27
ms per token,  3678.16 tokens per second)
llama_print_timings: prompt eval time =   17419.37 ms /    79 tokens (  220.50
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    6405.24 ms /    23 runs   (  278.49
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   23919.36 ms /   102 tokens
No. of rows:  21%| | 269/1258 [2:30:25<8:27:56, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.60 ms /    27 runs   (    0.28
ms per token,  3550.76 tokens per second)
llama_print_timings: prompt eval time =   19577.98 ms /    90 tokens (  217.53
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    7135.59 ms /    26 runs   (  274.45
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   26823.39 ms /   116 tokens
No. of rows:  21%| | 270/1258 [2:30:51<8:07:46, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.57 ms /    41 runs   (    0.26
ms per token,  3880.00 tokens per second)
llama_print_timings: prompt eval time =   19530.85 ms /    91 tokens (  214.62
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   12546.53 ms /    40 runs   (  313.66
ms per token,     3.19 tokens per second)
llama_print_timings:       total time =   32235.56 ms /   131 tokens
No. of rows:  22%| | 271/1258 [2:31:24<8:20:13, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.38 ms /    35 runs   (    0.27
ms per token,  3730.95 tokens per second)
llama_print_timings: prompt eval time =   17882.51 ms /    82 tokens (  218.08
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    9191.61 ms /    34 runs   (  270.34
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   27210.22 ms /   116 tokens
```

```
No. of rows:  22%| | 272/1258 [2:31:51<8:03:59, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.79 ms /    50 runs   (    0.28
ms per token,  3624.50 tokens per second)
llama_print_timings: prompt eval time =   20186.66 ms /    94 tokens (  214.75
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   15457.78 ms /    49 runs   (  315.46
ms per token,     3.17 tokens per second)
llama_print_timings:       total time =   35842.57 ms /   143 tokens
No. of rows:  22%| | 273/1258 [2:32:27<8:34:57, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.12 ms /    27 runs   (    0.26
ms per token,  3793.20 tokens per second)
llama_print_timings: prompt eval time =   19353.23 ms /    90 tokens (  215.04
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    7126.14 ms /    26 runs   (  274.08
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   26583.57 ms /   116 tokens
No. of rows:  22%| | 274/1258 [2:32:53<8:10:58, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.58 ms /    50 runs   (    0.27
ms per token,  3680.80 tokens per second)
llama_print_timings: prompt eval time =   26597.94 ms /   124 tokens (  214.50
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13384.10 ms /    49 runs   (  273.14
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   40178.31 ms /   173 tokens
No. of rows:  22%| | 275/1258 [2:33:33<9:00:50, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.90 ms /    50 runs   (    0.28
ms per token,  3597.90 tokens per second)
llama_print_timings: prompt eval time =   29604.65 ms /   133 tokens (  222.59
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =   13758.42 ms /    49 runs   (  280.78
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   43568.29 ms /   182 tokens
No. of rows:  22%| | 276/1258 [2:34:17<9:52:07, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      13.78 ms /    48 runs   (    0.29
ms per token,  3482.55 tokens per second)
llama_print_timings: prompt eval time =   21234.26 ms /    98 tokens (  216.68
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   13227.28 ms /    47 runs   (  281.43
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   34657.12 ms /   145 tokens
No. of rows:  22%| | 277/1258 [2:34:52<9:44:07, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.56 ms /    35 runs   (    0.27
ms per token,  3659.17 tokens per second)
llama_print_timings: prompt eval time =   24517.02 ms /   111 tokens (  220.87
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    9550.81 ms /    34 runs   (  280.91
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   34212.55 ms /   145 tokens
No. of rows:  22%| | 278/1258 [2:35:26<9:36:06, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.70 ms /    32 runs   (    0.27
ms per token,  3680.28 tokens per second)
llama_print_timings: prompt eval time =   21746.60 ms /   101 tokens (  215.31
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    8615.85 ms /    31 runs   (  277.93
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   30485.71 ms /   132 tokens
No. of rows:  22%| | 279/1258 [2:35:56<9:12:07, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.22 ms /    27 runs   (    0.27
ms per token,  3740.65 tokens per second)
llama_print_timings: prompt eval time =   19069.78 ms /    87 tokens (  219.19
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    7269.18 ms /    26 runs   (  279.58
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   26443.86 ms /   113 tokens
No. of rows:  22%| | 280/1258 [2:36:23<8:35:27, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.21 ms /    39 runs   (    0.29
ms per token,  3479.66 tokens per second)
llama_print_timings: prompt eval time =   21893.71 ms /    98 tokens (  223.41
ms per token,     4.48 tokens per second)
```

```
llama_print_timings:        eval time =   10576.81 ms /    38 runs   (  278.34
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   32629.05 ms /   136 tokens
No. of rows:  22%| | 281/1258 [2:36:56<8:39:54, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.88 ms /    36 runs   (    0.27
ms per token,  3641.88 tokens per second)
llama_print_timings: prompt eval time =   23990.72 ms /   110 tokens (  218.10
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    9724.69 ms /    35 runs   (  277.85
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   33860.42 ms /   145 tokens
No. of rows:  22%| | 282/1258 [2:37:29<8:48:50, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.31 ms /    27 runs   (    0.27
ms per token,  3692.56 tokens per second)
llama_print_timings: prompt eval time =   21471.88 ms /    99 tokens (  216.89
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    7221.38 ms /    26 runs   (  277.75
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   28801.22 ms /   125 tokens
No. of rows:  22%| | 283/1258 [2:37:58<8:30:15, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.59 ms /    50 runs   (    0.27
ms per token,  3679.45 tokens per second)
llama_print_timings: prompt eval time =   33623.87 ms /   148 tokens (  227.19
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =   15448.04 ms /    49 runs   (  315.27
ms per token,     3.17 tokens per second)
llama_print_timings:       total time =   49274.93 ms /   197 tokens
No. of rows:  23%| | 284/1258 [2:38:47<9:56:48, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.50 ms /    50 runs   (    0.27
ms per token,  3704.80 tokens per second)
llama_print_timings: prompt eval time =   19419.92 ms /    90 tokens (  215.78
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   13278.34 ms /    49 runs   (  270.99
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   32891.62 ms /   139 tokens
No. of rows:  23%| | 285/1258 [2:39:20<9:37:24, 35Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.66 ms /    39 runs   (    0.25
ms per token,  4038.94 tokens per second)
llama_print_timings: prompt eval time =   19972.57 ms /    93 tokens (  214.76
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   10239.88 ms /    38 runs   (  269.47
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   30363.44 ms /   131 tokens
No. of rows:  23%| | 286/1258 [2:39:51<9:11:21, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.03 ms /    26 runs   (    0.27
ms per token,  3700.54 tokens per second)
llama_print_timings: prompt eval time =   18671.85 ms /    79 tokens (  236.35
ms per token,     4.23 tokens per second)
llama_print_timings:        eval time =    8896.61 ms /    25 runs   (  355.86
ms per token,     2.81 tokens per second)
llama_print_timings:       total time =   27672.33 ms /   104 tokens
No. of rows:  23%| | 287/1258 [2:40:18<8:39:56, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.72 ms /    36 runs   (    0.27
ms per token,  3705.23 tokens per second)
llama_print_timings: prompt eval time =   22019.18 ms /   103 tokens (  213.78
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   10138.28 ms /    35 runs   (  289.67
ms per token,     3.45 tokens per second)
llama_print_timings:       total time =   32300.28 ms /   138 tokens
No. of rows:  23%| | 288/1258 [2:40:51<8:40:17, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.45 ms /    34 runs   (    0.28
ms per token,  3599.79 tokens per second)
llama_print_timings: prompt eval time =   22318.32 ms /   104 tokens (  214.60
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    8835.45 ms /    33 runs   (  267.74
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   31285.88 ms /   137 tokens
No. of rows:  23%| | 289/1258 [2:41:22<8:35:26, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.88 ms /    22 runs   (    0.27
```

```
ms per token,  3744.04 tokens per second)
llama_print_timings: prompt eval time =   17088.42 ms /    80 tokens (  213.61
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    5675.43 ms /    21 runs   (  270.26
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   22849.02 ms /   101 tokens
No. of rows:  23%| | 290/1258 [2:41:45<7:51:03, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.96 ms /    29 runs   (    0.27
ms per token,  3640.93 tokens per second)
llama_print_timings: prompt eval time =   19721.63 ms /    85 tokens (  232.02
ms per token,     4.31 tokens per second)
llama_print_timings:        eval time =    7563.40 ms /    28 runs   (  270.12
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   27396.07 ms /   113 tokens
No. of rows:  23%| | 291/1258 [2:42:12<7:41:53, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.94 ms /    38 runs   (    0.26
ms per token,  3824.86 tokens per second)
llama_print_timings: prompt eval time =   22533.64 ms /   105 tokens (  214.61
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   10144.54 ms /    37 runs   (  274.18
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   32826.50 ms /   142 tokens
No. of rows:  23%| | 292/1258 [2:42:45<8:01:34, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.54 ms /    25 runs   (    0.26
ms per token,  3819.71 tokens per second)
llama_print_timings: prompt eval time =   20120.63 ms /    93 tokens (  216.35
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    6683.88 ms /    24 runs   (  278.49
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   26904.12 ms /   117 tokens
No. of rows:  23%| | 293/1258 [2:43:12<7:46:36, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.56 ms /    50 runs   (    0.27
ms per token,  3686.50 tokens per second)
llama_print_timings: prompt eval time =   26687.06 ms /   125 tokens (  213.50
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13356.81 ms /    49 runs   (  272.59
```

```
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   40237.73 ms /   174 tokens
No. of rows:  23%| | 294/1258 [2:43:52<8:40:15, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.65 ms /    50 runs   (    0.27
ms per token,  3664.08 tokens per second)
llama_print_timings: prompt eval time =   25921.64 ms /   118 tokens (  219.67
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   13656.22 ms /    49 runs   (  278.70
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   39779.71 ms /   167 tokens
No. of rows:  23%| | 295/1258 [2:44:32<9:15:22, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.76 ms /    28 runs   (    0.28
ms per token,  3606.85 tokens per second)
llama_print_timings: prompt eval time =   24233.66 ms /   104 tokens (  233.02
ms per token,     4.29 tokens per second)
llama_print_timings:        eval time =    7273.68 ms /    27 runs   (  269.40
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   31617.39 ms /   131 tokens
No. of rows:  24%| | 296/1258 [2:45:04<9:00:28, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.41 ms /    38 runs   (    0.25
ms per token,  4036.54 tokens per second)
llama_print_timings: prompt eval time =   23222.12 ms /   106 tokens (  219.08
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   10243.40 ms /    37 runs   (  276.85
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   33611.18 ms /   143 tokens
No. of rows:  24%| | 297/1258 [2:45:37<8:59:28, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.42 ms /    34 runs   (    0.28
ms per token,  3609.34 tokens per second)
llama_print_timings: prompt eval time =   20252.43 ms /    94 tokens (  215.45
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   10599.69 ms /    33 runs   (  321.20
ms per token,     3.11 tokens per second)
llama_print_timings:       total time =   30986.40 ms /   127 tokens
No. of rows:  24%| | 298/1258 [2:46:08<8:46:00, 32Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.73 ms /    50 runs   (    0.27
ms per token,  3641.66 tokens per second)
llama_print_timings: prompt eval time =   22055.51 ms /   100 tokens (  220.56
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =   13486.11 ms /    49 runs   (  275.23
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   35738.77 ms /   149 tokens
No. of rows:  24%| | 299/1258 [2:46:44<8:59:13, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.66 ms /    29 runs   (    0.26
ms per token,  3783.92 tokens per second)
llama_print_timings: prompt eval time =   20718.40 ms /    94 tokens (  220.41
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    7526.76 ms /    28 runs   (  268.81
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   28356.86 ms /   122 tokens
No. of rows:  24%| | 300/1258 [2:47:12<8:32:55, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.97 ms /    37 runs   (    0.27
ms per token,  3710.76 tokens per second)
llama_print_timings: prompt eval time =   18155.71 ms /    85 tokens (  213.60
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   11433.62 ms /    36 runs   (  317.60
ms per token,     3.15 tokens per second)
llama_print_timings:       total time =   29732.67 ms /   121 tokens
No. of rows:  24%| | 301/1258 [2:47:42<8:20:58, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.75 ms /    37 runs   (    0.26
ms per token,  3793.70 tokens per second)
llama_print_timings: prompt eval time =   21308.45 ms /    99 tokens (  215.24
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9827.73 ms /    36 runs   (  272.99
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   31277.94 ms /   135 tokens
No. of rows:  24%| | 302/1258 [2:48:13<8:19:51, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.76 ms /    36 runs   (    0.27
ms per token,  3690.42 tokens per second)
```

```
llama_print_timings: prompt eval time =   27154.45 ms /   125 tokens (  217.24
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    9584.27 ms /    35 runs   (  273.84
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   36879.44 ms /   160 tokens
No. of rows:  24%| | 303/1258 [2:48:50<8:45:38, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.55 ms /    50 runs   (    0.27
ms per token,  3688.68 tokens per second)
llama_print_timings: prompt eval time =   30737.10 ms /   137 tokens (  224.36
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =   13263.33 ms /    49 runs   (  270.68
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   44197.91 ms /   186 tokens
No. of rows:  24%| | 304/1258 [2:49:35<9:38:26, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.37 ms /    50 runs   (    0.29
ms per token,  3479.96 tokens per second)
llama_print_timings: prompt eval time =   22424.21 ms /   105 tokens (  213.56
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   15037.50 ms /    49 runs   (  306.89
ms per token,     3.26 tokens per second)
llama_print_timings:       total time =   37659.40 ms /   154 tokens
No. of rows:  24%| | 305/1258 [2:50:12<9:43:58, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.57 ms /    35 runs   (    0.27
ms per token,  3658.79 tokens per second)
llama_print_timings: prompt eval time =   21480.18 ms /    96 tokens (  223.75
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =    9250.93 ms /    34 runs   (  272.09
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   30866.80 ms /   130 tokens
No. of rows:  24%| | 306/1258 [2:50:43<9:15:19, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.72 ms /    40 runs   (    0.27
ms per token,  3730.30 tokens per second)
llama_print_timings: prompt eval time =   18227.27 ms /    81 tokens (  225.03
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =   10562.43 ms /    39 runs   (  270.83
ms per token,     3.69 tokens per second)
```

```
llama_print_timings:       total time =   28947.17 ms /   120 tokens
No. of rows:  24%| | 307/1258 [2:51:12<8:45:55, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.06 ms /    44 runs   (    0.27
ms per token,  3646.91 tokens per second)
llama_print_timings: prompt eval time =   21557.85 ms /   100 tokens (  215.58
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   11750.75 ms /    43 runs   (  273.27
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   33486.81 ms /   143 tokens
No. of rows:  24%| | 308/1258 [2:51:46<8:46:54, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.65 ms /    22 runs   (    0.26
ms per token,  3893.12 tokens per second)
llama_print_timings: prompt eval time =   17789.26 ms /    83 tokens (  214.33
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    5852.69 ms /    21 runs   (  278.70
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   23726.19 ms /   104 tokens
No. of rows:  25%| | 309/1258 [2:52:09<8:01:06, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.94 ms /    39 runs   (    0.25
ms per token,  3923.54 tokens per second)
llama_print_timings: prompt eval time =   22992.73 ms /   106 tokens (  216.91
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   10371.56 ms /    38 runs   (  272.94
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   33515.78 ms /   144 tokens
No. of rows:  25%| | 310/1258 [2:52:43<8:15:18, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.18 ms /    34 runs   (    0.27
ms per token,  3704.11 tokens per second)
llama_print_timings: prompt eval time =   21843.87 ms /   100 tokens (  218.44
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    8917.44 ms /    33 runs   (  270.23
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   30895.19 ms /   133 tokens
No. of rows:  25%| | 311/1258 [2:53:14<8:12:41, 31Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.53 ms /    28 runs   (    0.27
ms per token,  3720.93 tokens per second)
llama_print_timings: prompt eval time =   20354.37 ms /    95 tokens (  214.26
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    7382.31 ms /    27 runs   (  273.42
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   27844.05 ms /   122 tokens
No. of rows:  25%| | 312/1258 [2:53:42<7:56:15, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.70 ms /    39 runs   (    0.27
ms per token,  3644.86 tokens per second)
llama_print_timings: prompt eval time =   26051.29 ms /   116 tokens (  224.58
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   10582.61 ms /    38 runs   (  278.49
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   36786.95 ms /   154 tokens
No. of rows:  25%| | 313/1258 [2:54:18<8:26:49, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.76 ms /    47 runs   (    0.29
ms per token,  3414.71 tokens per second)
llama_print_timings: prompt eval time =   29831.75 ms /   134 tokens (  222.63
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =   12832.53 ms /    46 runs   (  278.97
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   42859.15 ms /   180 tokens
No. of rows:  25%| | 314/1258 [2:55:01<9:16:46, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.13 ms /    26 runs   (    0.27
ms per token,  3645.03 tokens per second)
llama_print_timings: prompt eval time =   18900.78 ms /    86 tokens (  219.78
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =    6896.05 ms /    25 runs   (  275.84
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   25898.69 ms /   111 tokens
No. of rows:  25%| | 315/1258 [2:55:27<8:31:29, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.59 ms /    36 runs   (    0.27
ms per token,  3755.87 tokens per second)
llama_print_timings: prompt eval time =   20460.25 ms /    94 tokens (  217.66
```

```
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    9543.93 ms /    35 runs   (  272.68
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   30143.59 ms /   129 tokens
No. of rows:  25%| | 316/1258 [2:55:57<8:19:36, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.43 ms /    29 runs   (    0.29
ms per token,  3440.50 tokens per second)
llama_print_timings: prompt eval time =   18495.05 ms /    79 tokens (  234.11
ms per token,     4.27 tokens per second)
llama_print_timings:        eval time =    7915.96 ms /    28 runs   (  282.71
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   26529.10 ms /   107 tokens
No. of rows:  25%| | 317/1258 [2:56:24<7:54:12, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.37 ms /    50 runs   (    0.27
ms per token,  3740.56 tokens per second)
llama_print_timings: prompt eval time =   24043.12 ms /   112 tokens (  214.67
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13672.37 ms /    49 runs   (  279.03
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   37914.06 ms /   161 tokens
No. of rows:  25%| | 318/1258 [2:57:02<8:29:48, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.00 ms /    30 runs   (    0.27
ms per token,  3751.41 tokens per second)
llama_print_timings: prompt eval time =   19927.22 ms /    92 tokens (  216.60
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7878.63 ms /    29 runs   (  271.68
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   27923.23 ms /   121 tokens
No. of rows:  25%| | 319/1258 [2:57:30<8:07:40, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.61 ms /    32 runs   (    0.27
ms per token,  3717.47 tokens per second)
llama_print_timings: prompt eval time =   19837.31 ms /    92 tokens (  215.62
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    8329.06 ms /    31 runs   (  268.68
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   28293.16 ms /   123 tokens
```

```
No. of rows:  25%| | 320/1258 [2:57:58<7:53:42, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.75 ms /    28 runs   (    0.28
ms per token,  3614.30 tokens per second)
llama_print_timings: prompt eval time =   19830.63 ms /    84 tokens (  236.08
ms per token,     4.24 tokens per second)
llama_print_timings:        eval time =    7732.88 ms /    27 runs   (  286.40
ms per token,     3.49 tokens per second)
llama_print_timings:       total time =   27674.95 ms /   111 tokens
No. of rows:  26%| | 321/1258 [2:58:26<7:40:57, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.20 ms /    32 runs   (    0.26
ms per token,  3902.91 tokens per second)
llama_print_timings: prompt eval time =   20474.58 ms /    94 tokens (  217.81
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    8581.12 ms /    31 runs   (  276.81
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   29179.27 ms /   125 tokens
No. of rows:  26%| | 322/1258 [2:58:55<7:38:55, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.27 ms /    20 runs   (    0.26
ms per token,  3794.35 tokens per second)
llama_print_timings: prompt eval time =   17194.54 ms /    78 tokens (  220.44
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    5155.57 ms /    19 runs   (  271.35
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   22427.23 ms /    97 tokens
No. of rows:  26%| | 323/1258 [2:59:17<7:05:46, 27Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.73 ms /    22 runs   (    0.26
ms per token,  3842.12 tokens per second)
llama_print_timings: prompt eval time =   17278.45 ms /    80 tokens (  215.98
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    5770.81 ms /    21 runs   (  274.80
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   23132.90 ms /   101 tokens
No. of rows:  26%| | 324/1258 [2:59:40<6:45:48, 26Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =       6.54 ms /    25 runs   (    0.26
ms per token,  3823.80 tokens per second)
llama_print_timings: prompt eval time =   17368.47 ms /    80 tokens (  217.11
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    6520.43 ms /    24 runs   (  271.68
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   23987.50 ms /   104 tokens
No. of rows:  26%| | 325/1258 [3:00:04<6:35:41, 25Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.58 ms /    44 runs   (    0.26
ms per token,  3799.33 tokens per second)
llama_print_timings: prompt eval time =   24095.42 ms /   103 tokens (  233.94
ms per token,     4.27 tokens per second)
llama_print_timings:        eval time =   11765.88 ms /    43 runs   (  273.62
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   36034.26 ms /   146 tokens
No. of rows:  26%| | 326/1258 [3:00:40<7:24:38, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.72 ms /    25 runs   (    0.27
ms per token,  3718.02 tokens per second)
llama_print_timings: prompt eval time =   18080.86 ms /    82 tokens (  220.50
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    6867.28 ms /    24 runs   (  286.14
ms per token,     3.49 tokens per second)
llama_print_timings:       total time =   25046.55 ms /   106 tokens
No. of rows:  26%| | 327/1258 [3:01:05<7:07:30, 27Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.68 ms /    50 runs   (    0.27
ms per token,  3655.24 tokens per second)
llama_print_timings: prompt eval time =   25593.54 ms /   120 tokens (  213.28
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   13337.02 ms /    49 runs   (  272.18
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   39128.34 ms /   169 tokens
No. of rows:  26%| | 328/1258 [3:01:45<8:00:56, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.66 ms /    47 runs   (    0.27
ms per token,  3712.77 tokens per second)
llama_print_timings: prompt eval time =   21413.46 ms /    97 tokens (  220.76
ms per token,     4.53 tokens per second)
```

```
llama_print_timings:        eval time =   12940.19 ms /    46 runs   (  281.31
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   34540.41 ms /   143 tokens
No. of rows:  26%| | 329/1258 [3:02:19<8:16:47, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.46 ms /    34 runs   (    0.28
ms per token,  3595.60 tokens per second)
llama_print_timings: prompt eval time =   22620.94 ms /   104 tokens (  217.51
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    9605.95 ms /    33 runs   (  291.09
ms per token,     3.44 tokens per second)
llama_print_timings:       total time =   32360.40 ms /   137 tokens
No. of rows:  26%| | 330/1258 [3:02:52<8:17:33, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.28 ms /    39 runs   (    0.26
ms per token,  3794.88 tokens per second)
llama_print_timings: prompt eval time =   18486.18 ms /    83 tokens (  222.73
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =   10354.03 ms /    38 runs   (  272.47
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   28993.69 ms /   121 tokens
No. of rows:  26%| | 331/1258 [3:03:21<8:02:20, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.42 ms /    25 runs   (    0.26
ms per token,  3896.51 tokens per second)
llama_print_timings: prompt eval time =   18673.74 ms /    86 tokens (  217.14
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    6418.12 ms /    24 runs   (  267.42
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   25186.70 ms /   110 tokens
No. of rows:  26%| | 332/1258 [3:03:46<7:33:55, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.15 ms /    24 runs   (    0.26
ms per token,  3900.54 tokens per second)
llama_print_timings: prompt eval time =   19869.16 ms /    91 tokens (  218.34
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    6210.43 ms /    23 runs   (  270.02
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   26170.77 ms /   114 tokens
No. of rows:  26%| | 333/1258 [3:04:12<7:18:28, 28Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.92 ms /    25 runs   (    0.28
ms per token,  3612.72 tokens per second)
llama_print_timings: prompt eval time =   19612.00 ms /    89 tokens (  220.36
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    6543.49 ms /    24 runs   (  272.65
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   26256.10 ms /   113 tokens
No. of rows:  27%| | 334/1258 [3:04:38<7:07:52, 27Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.49 ms /    47 runs   (    0.27
ms per token,  3763.01 tokens per second)
llama_print_timings: prompt eval time =   20172.30 ms /    93 tokens (  216.91
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   12870.40 ms /    46 runs   (  279.79
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   33228.46 ms /   139 tokens
No. of rows:  27%| | 335/1258 [3:05:11<7:32:37, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.35 ms /    36 runs   (    0.26
ms per token,  3851.92 tokens per second)
llama_print_timings: prompt eval time =   23904.80 ms /   109 tokens (  219.31
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    9596.32 ms /    35 runs   (  274.18
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   33641.28 ms /   144 tokens
No. of rows:  27%| | 336/1258 [3:05:45<7:51:36, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.93 ms /    50 runs   (    0.28
ms per token,  3589.89 tokens per second)
llama_print_timings: prompt eval time =   25243.26 ms /   117 tokens (  215.75
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   15354.21 ms /    49 runs   (  313.35
ms per token,     3.19 tokens per second)
llama_print_timings:       total time =   40800.37 ms /   166 tokens
No. of rows:  27%| | 337/1258 [3:06:26<8:37:41, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.13 ms /    43 runs   (    0.28
```

```
ms per token,  3544.93 tokens per second)
llama_print_timings: prompt eval time =   21426.81 ms /    97 tokens (  220.89
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =   13068.22 ms /    42 runs   (  311.15
ms per token,     3.21 tokens per second)
llama_print_timings:       total time =   34668.67 ms /   139 tokens
No. of rows:  27%| | 338/1258 [3:07:01<8:41:27, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.64 ms /    36 runs   (    0.27
ms per token,  3734.05 tokens per second)
llama_print_timings: prompt eval time =   20946.41 ms /    95 tokens (  220.49
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    9661.02 ms /    35 runs   (  276.03
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   30746.67 ms /   130 tokens
No. of rows:  27%| | 339/1258 [3:07:31<8:25:59, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.23 ms /    31 runs   (    0.27
ms per token,  3766.71 tokens per second)
llama_print_timings: prompt eval time =   21601.11 ms /   101 tokens (  213.87
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    8252.14 ms /    30 runs   (  275.07
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   29971.63 ms /   131 tokens
No. of rows:  27%| | 340/1258 [3:08:01<8:11:24, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.26 ms /    41 runs   (    0.27
ms per token,  3641.85 tokens per second)
llama_print_timings: prompt eval time =   24520.22 ms /   107 tokens (  229.16
ms per token,     4.36 tokens per second)
llama_print_timings:        eval time =   12831.67 ms /    40 runs   (  320.79
ms per token,     3.12 tokens per second)
llama_print_timings:       total time =   37511.52 ms /   147 tokens
No. of rows:  27%| | 341/1258 [3:08:39<8:35:38, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.01 ms /    33 runs   (    0.27
ms per token,  3663.00 tokens per second)
llama_print_timings: prompt eval time =   20012.15 ms /    92 tokens (  217.52
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =   10466.64 ms /    32 runs   (  327.08
```

```
ms per token,     3.06 tokens per second)
llama_print_timings:       total time =   30608.77 ms /   124 tokens
No. of rows:  27%| | 342/1258 [3:09:09<8:20:45, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.39 ms /    26 runs   (    0.28
ms per token,  3516.84 tokens per second)
llama_print_timings: prompt eval time =   17912.15 ms /    83 tokens (  215.81
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    6900.79 ms /    25 runs   (  276.03
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   24914.81 ms /   108 tokens
No. of rows:  27%| | 343/1258 [3:09:34<7:44:11, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.01 ms /    50 runs   (    0.26
ms per token,  3844.68 tokens per second)
llama_print_timings: prompt eval time =   29515.91 ms /   138 tokens (  213.88
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   15186.33 ms /    49 runs   (  309.93
ms per token,     3.23 tokens per second)
llama_print_timings:       total time =   44899.62 ms /   187 tokens
No. of rows:  27%| | 344/1258 [3:10:19<8:49:48, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.86 ms /    50 runs   (    0.28
ms per token,  3608.28 tokens per second)
llama_print_timings: prompt eval time =   24468.95 ms /   109 tokens (  224.49
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   13315.70 ms /    49 runs   (  271.75
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   37982.99 ms /   158 tokens
No. of rows:  27%| | 345/1258 [3:10:57<9:03:53, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.17 ms /    26 runs   (    0.28
ms per token,  3626.22 tokens per second)
llama_print_timings: prompt eval time =   18582.73 ms /    84 tokens (  221.22
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    6749.36 ms /    25 runs   (  269.97
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   25434.99 ms /   109 tokens
No. of rows:  28%| | 346/1258 [3:11:23<8:16:19, 32Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.59 ms /    35 runs   (    0.27
ms per token,  3649.25 tokens per second)
llama_print_timings: prompt eval time =   19809.28 ms /    93 tokens (  213.00
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    9309.99 ms /    34 runs   (  273.82
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   29255.76 ms /   127 tokens
No. of rows:  28%| | 347/1258 [3:11:52<8:00:20, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.85 ms /    50 runs   (    0.28
ms per token,  3609.33 tokens per second)
llama_print_timings: prompt eval time =   17962.73 ms /    82 tokens (  219.06
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =   13638.67 ms /    49 runs   (  278.34
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   31805.08 ms /   131 tokens
No. of rows:  28%| | 348/1258 [3:12:24<8:00:36, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.83 ms /    50 runs   (    0.28
ms per token,  3616.64 tokens per second)
llama_print_timings: prompt eval time =   28255.65 ms /   125 tokens (  226.05
ms per token,     4.42 tokens per second)
llama_print_timings:        eval time =   13721.67 ms /    49 runs   (  280.03
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   42180.21 ms /   174 tokens
No. of rows:  28%| | 349/1258 [3:13:06<8:47:44, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.61 ms /    32 runs   (    0.27
ms per token,  3715.31 tokens per second)
llama_print_timings: prompt eval time =   20373.77 ms /    93 tokens (  219.07
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    8480.05 ms /    31 runs   (  273.55
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   28980.95 ms /   124 tokens
No. of rows:  28%| | 350/1258 [3:13:35<8:20:40, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.86 ms /    37 runs   (    0.27
ms per token,  3751.77 tokens per second)
```

```
llama_print_timings: prompt eval time =   21779.06 ms /   100 tokens (  217.79
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    9782.84 ms /    36 runs   (  271.75
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   31705.22 ms /   136 tokens
No. of rows:  28%| | 351/1258 [3:14:07<8:13:54, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.86 ms /    33 runs   (    0.27
ms per token,  3723.34 tokens per second)
llama_print_timings: prompt eval time =   22936.51 ms /   106 tokens (  216.38
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    8776.98 ms /    32 runs   (  274.28
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   31841.96 ms /   138 tokens
No. of rows:  28%| | 352/1258 [3:14:38<8:09:37, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.15 ms /    30 runs   (    0.27
ms per token,  3680.08 tokens per second)
llama_print_timings: prompt eval time =   18529.01 ms /    85 tokens (  217.99
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    9519.32 ms /    29 runs   (  328.25
ms per token,     3.05 tokens per second)
llama_print_timings:       total time =   28164.76 ms /   114 tokens
No. of rows:  28%| | 353/1258 [3:15:07<7:49:50, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.45 ms /    25 runs   (    0.26
ms per token,  3874.17 tokens per second)
llama_print_timings: prompt eval time =   20948.00 ms /    96 tokens (  218.21
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    6672.33 ms /    24 runs   (  278.01
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   27716.77 ms /   120 tokens
No. of rows:  28%| | 354/1258 [3:15:34<7:33:52, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.47 ms /    35 runs   (    0.27
ms per token,  3697.44 tokens per second)
llama_print_timings: prompt eval time =   22812.36 ms /   106 tokens (  215.21
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9308.93 ms /    34 runs   (  273.79
ms per token,     3.65 tokens per second)
```

```
llama_print_timings:       total time =   32255.96 ms /   140 tokens
No. of rows:  28%| | 355/1258 [3:16:07<7:43:01, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.67 ms /    34 runs   (    0.28
ms per token,  3514.58 tokens per second)
llama_print_timings: prompt eval time =   19125.45 ms /    87 tokens (  219.83
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =    9042.69 ms /    33 runs   (  274.02
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   28306.93 ms /   120 tokens
No. of rows:  28%| | 356/1258 [3:16:35<7:31:27, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.04 ms /    41 runs   (    0.27
ms per token,  3712.76 tokens per second)
llama_print_timings: prompt eval time =   22720.07 ms /    99 tokens (  229.50
ms per token,     4.36 tokens per second)
llama_print_timings:        eval time =   12906.05 ms /    40 runs   (  322.65
ms per token,     3.10 tokens per second)
llama_print_timings:       total time =   35784.66 ms /   139 tokens
No. of rows:  28%| | 357/1258 [3:17:11<7:56:55, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.45 ms /    38 runs   (    0.28
ms per token,  3636.36 tokens per second)
llama_print_timings: prompt eval time =   24384.26 ms /   111 tokens (  219.68
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   10204.26 ms /    37 runs   (  275.79
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   34738.89 ms /   148 tokens
No. of rows:  28%| | 358/1258 [3:17:46<8:09:50, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.25 ms /    27 runs   (    0.27
ms per token,  3723.62 tokens per second)
llama_print_timings: prompt eval time =   21779.50 ms /   100 tokens (  217.80
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    7158.84 ms /    26 runs   (  275.34
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   29043.03 ms /   126 tokens
No. of rows:  29%| | 359/1258 [3:18:15<7:53:05, 31Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.70 ms /    47 runs   (    0.27
ms per token,  3700.79 tokens per second)
llama_print_timings: prompt eval time =   24388.53 ms /   106 tokens (  230.08
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =   12697.86 ms /    46 runs   (  276.04
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   37291.85 ms /   152 tokens
No. of rows:  29%| | 360/1258 [3:18:52<8:18:15, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.09 ms /    33 runs   (    0.28
ms per token,  3631.96 tokens per second)
llama_print_timings: prompt eval time =   20385.98 ms /    95 tokens (  214.59
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    9051.20 ms /    32 runs   (  282.85
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   29563.32 ms /   127 tokens
No. of rows:  29%| | 361/1258 [3:19:21<8:00:57, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.84 ms /    38 runs   (    0.26
ms per token,  3862.18 tokens per second)
llama_print_timings: prompt eval time =   22873.41 ms /   106 tokens (  215.79
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   10084.56 ms /    37 runs   (  272.56
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   33106.71 ms /   143 tokens
No. of rows:  29%| | 362/1258 [3:19:55<8:04:42, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.68 ms /    40 runs   (    0.27
ms per token,  3744.62 tokens per second)
llama_print_timings: prompt eval time =   21529.15 ms /    99 tokens (  217.47
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =   10893.15 ms /    39 runs   (  279.31
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   32583.05 ms /   138 tokens
No. of rows:  29%| | 363/1258 [3:20:27<8:04:45, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.76 ms /    50 runs   (    0.26
ms per token,  3917.27 tokens per second)
llama_print_timings: prompt eval time =   28486.14 ms /   133 tokens (  214.18
```

```
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13483.23 ms /    49 runs   (  275.17
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   42162.03 ms /   182 tokens
No. of rows:  29%| | 364/1258 [3:21:09<8:47:26, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.58 ms /    24 runs   (    0.27
ms per token,  3646.31 tokens per second)
llama_print_timings: prompt eval time =   17184.08 ms /    78 tokens (  220.31
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    6382.03 ms /    23 runs   (  277.48
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   23660.12 ms /   101 tokens
No. of rows:  29%| | 365/1258 [3:21:33<7:54:28, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.09 ms /    39 runs   (    0.26
ms per token,  3865.98 tokens per second)
llama_print_timings: prompt eval time =   20354.02 ms /    94 tokens (  216.53
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   10403.50 ms /    38 runs   (  273.78
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   30916.16 ms /   132 tokens
No. of rows:  29%| | 366/1258 [3:22:04<7:49:40, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.93 ms /    34 runs   (    0.26
ms per token,  3806.54 tokens per second)
llama_print_timings: prompt eval time =   23064.08 ms /   107 tokens (  215.55
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    9216.04 ms /    33 runs   (  279.27
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   32411.17 ms /   140 tokens
No. of rows:  29%| | 367/1258 [3:22:36<7:52:49, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.92 ms /    34 runs   (    0.26
ms per token,  3813.80 tokens per second)
llama_print_timings: prompt eval time =   21897.24 ms /   101 tokens (  216.80
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    9039.49 ms /    33 runs   (  273.92
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   31067.22 ms /   134 tokens
```

```
No. of rows:  29%| | 368/1258 [3:23:07<7:48:53, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.01 ms /    31 runs   (    0.26
ms per token,  3870.16 tokens per second)
llama_print_timings: prompt eval time =   21528.35 ms /    99 tokens (  217.46
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    8215.85 ms /    30 runs   (  273.86
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   29862.73 ms /   129 tokens
No. of rows:  29%| | 369/1258 [3:23:37<7:40:37, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.29 ms /    24 runs   (    0.26
ms per token,  3816.19 tokens per second)
llama_print_timings: prompt eval time =   14680.12 ms /    67 tokens (  219.11
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    6428.40 ms /    23 runs   (  279.50
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   21203.80 ms /    90 tokens
No. of rows:  29%| | 370/1258 [3:23:58<6:56:14, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.29 ms /    50 runs   (    0.27
ms per token,  3763.64 tokens per second)
llama_print_timings: prompt eval time =   29248.67 ms /   135 tokens (  216.66
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   13560.21 ms /    49 runs   (  276.74
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   43003.83 ms /   184 tokens
No. of rows:  29%| | 371/1258 [3:24:41<8:01:48, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.60 ms /    36 runs   (    0.27
ms per token,  3749.22 tokens per second)
llama_print_timings: prompt eval time =   20236.41 ms /    85 tokens (  238.08
ms per token,     4.20 tokens per second)
llama_print_timings:        eval time =    9756.51 ms /    35 runs   (  278.76
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   30139.46 ms /   120 tokens
No. of rows:  30%| | 372/1258 [3:25:12<7:50:26, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      12.81 ms /    48 runs   (    0.27
ms per token,  3747.66 tokens per second)
llama_print_timings: prompt eval time =   25219.48 ms /   109 tokens (  231.37
ms per token,     4.32 tokens per second)
llama_print_timings:        eval time =   12857.82 ms /    47 runs   (  273.57
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   38264.94 ms /   156 tokens
No. of rows:  30%| | 373/1258 [3:25:50<8:18:17, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.72 ms /    38 runs   (    0.26
ms per token,  3907.46 tokens per second)
llama_print_timings: prompt eval time =   21868.03 ms /   100 tokens (  218.68
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =   10083.69 ms /    37 runs   (  272.53
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   32097.36 ms /   137 tokens
No. of rows:  30%| | 374/1258 [3:26:22<8:10:18, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.34 ms /    47 runs   (    0.26
ms per token,  3809.37 tokens per second)
llama_print_timings: prompt eval time =   25231.04 ms /   115 tokens (  219.40
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   12690.00 ms /    46 runs   (  275.87
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   38111.87 ms /   161 tokens
No. of rows:  30%| | 375/1258 [3:27:00<8:31:04, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.26 ms /    35 runs   (    0.26
ms per token,  3780.51 tokens per second)
llama_print_timings: prompt eval time =   23518.93 ms /   102 tokens (  230.58
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =    9403.98 ms /    34 runs   (  276.59
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   33058.36 ms /   136 tokens
No. of rows:  30%| | 376/1258 [3:27:33<8:23:12, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.76 ms /    50 runs   (    0.28
ms per token,  3633.72 tokens per second)
llama_print_timings: prompt eval time =   23048.40 ms /   108 tokens (  213.41
ms per token,     4.69 tokens per second)
```

```
llama_print_timings:        eval time =   13825.13 ms /    49 runs   (  282.15
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   37073.88 ms /   157 tokens
No. of rows:  30%| | 377/1258 [3:28:10<8:35:11, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.38 ms /    50 runs   (    0.27
ms per token,  3737.48 tokens per second)
llama_print_timings: prompt eval time =   30754.13 ms /   136 tokens (  226.13
ms per token,     4.42 tokens per second)
llama_print_timings:        eval time =   14052.74 ms /    49 runs   (  286.79
ms per token,     3.49 tokens per second)
llama_print_timings:       total time =   45003.97 ms /   185 tokens
No. of rows:  30%| | 378/1258 [3:28:55<9:18:16, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.60 ms /    44 runs   (    0.26
ms per token,  3794.41 tokens per second)
llama_print_timings: prompt eval time =   31561.78 ms /   141 tokens (  223.84
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   11739.27 ms /    43 runs   (  273.01
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   43470.99 ms /   184 tokens
No. of rows:  30%| | 379/1258 [3:29:39<9:41:25, 39Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.35 ms /    50 runs   (    0.27
ms per token,  3746.72 tokens per second)
llama_print_timings: prompt eval time =   22590.42 ms /   105 tokens (  215.15
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13530.00 ms /    49 runs   (  276.12
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   36315.53 ms /   154 tokens
No. of rows:  30%| | 380/1258 [3:30:15<9:25:59, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.86 ms /    31 runs   (    0.25
ms per token,  3946.03 tokens per second)
llama_print_timings: prompt eval time =   21679.62 ms /    99 tokens (  218.99
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    8170.66 ms /    30 runs   (  272.36
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   29969.36 ms /   129 tokens
No. of rows:  30%| | 381/1258 [3:30:45<8:47:11, 36Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.40 ms /    38 runs   (    0.27
ms per token,  3653.85 tokens per second)
llama_print_timings: prompt eval time =   20100.38 ms /    85 tokens (  236.47
ms per token,     4.23 tokens per second)
llama_print_timings:        eval time =   10034.63 ms /    37 runs   (  271.21
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   30281.95 ms /   122 tokens
No. of rows:  30%| | 382/1258 [3:31:15<8:21:16, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.87 ms /    30 runs   (    0.26
ms per token,  3812.43 tokens per second)
llama_print_timings: prompt eval time =   22799.78 ms /    98 tokens (  232.65
ms per token,     4.30 tokens per second)
llama_print_timings:        eval time =    7893.69 ms /    29 runs   (  272.20
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   30806.32 ms /   127 tokens
No. of rows:  30%| | 383/1258 [3:31:46<8:05:18, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.75 ms /    50 runs   (    0.27
ms per token,  3636.63 tokens per second)
llama_print_timings: prompt eval time =   23285.45 ms /   110 tokens (  211.69
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   15523.18 ms /    49 runs   (  316.80
ms per token,     3.16 tokens per second)
llama_print_timings:       total time =   39012.15 ms /   159 tokens
No. of rows:  31%| | 384/1258 [3:32:25<8:29:50, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.94 ms /    33 runs   (    0.27
ms per token,  3692.93 tokens per second)
llama_print_timings: prompt eval time =   19739.07 ms /    90 tokens (  219.32
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    8664.70 ms /    32 runs   (  270.77
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   28536.14 ms /   122 tokens
No. of rows:  31%| | 385/1258 [3:32:54<8:01:04, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.72 ms /    45 runs   (    0.28
```

```
ms per token,  3537.18 tokens per second)
llama_print_timings: prompt eval time =   23941.81 ms /   112 tokens (  213.77
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   12387.69 ms /    44 runs   (  281.54
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   36508.28 ms /   156 tokens
No. of rows:  31%| | 386/1258 [3:33:30<8:15:34, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.53 ms /    50 runs   (    0.27
ms per token,  3696.31 tokens per second)
llama_print_timings: prompt eval time =   27092.69 ms /   128 tokens (  211.66
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   13350.79 ms /    49 runs   (  272.47
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   40641.53 ms /   177 tokens
No. of rows:  31%| | 387/1258 [3:34:11<8:43:31, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.95 ms /    30 runs   (    0.26
ms per token,  3775.01 tokens per second)
llama_print_timings: prompt eval time =   19192.79 ms /    85 tokens (  225.80
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =    7892.03 ms /    29 runs   (  272.14
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   27201.04 ms /   114 tokens
No. of rows:  31%| | 388/1258 [3:34:38<8:04:24, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.31 ms /    32 runs   (    0.26
ms per token,  3848.93 tokens per second)
llama_print_timings: prompt eval time =   18119.43 ms /    83 tokens (  218.31
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    8457.45 ms /    31 runs   (  272.82
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   26698.74 ms /   114 tokens
No. of rows:  31%| | 389/1258 [3:35:05<7:34:44, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.32 ms /    45 runs   (    0.27
ms per token,  3653.78 tokens per second)
llama_print_timings: prompt eval time =   22635.03 ms /   105 tokens (  215.57
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   12207.01 ms /    44 runs   (  277.43
```

```
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   35027.25 ms /   149 tokens
No. of rows:  31%| | 390/1258 [3:35:40<7:49:59, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.84 ms /    30 runs   (    0.26
ms per token,  3827.51 tokens per second)
llama_print_timings: prompt eval time =   18603.05 ms /    86 tokens (  216.31
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7930.55 ms /    29 runs   (  273.47
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   26650.04 ms /   115 tokens
No. of rows:  31%| | 391/1258 [3:36:06<7:24:10, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.26 ms /    44 runs   (    0.28
ms per token,  3589.20 tokens per second)
llama_print_timings: prompt eval time =   22606.61 ms /    97 tokens (  233.06
ms per token,     4.29 tokens per second)
llama_print_timings:        eval time =   11881.82 ms /    43 runs   (  276.32
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   34661.59 ms /   140 tokens
No. of rows:  31%| | 392/1258 [3:36:41<7:40:40, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.15 ms /    40 runs   (    0.28
ms per token,  3588.73 tokens per second)
llama_print_timings: prompt eval time =   26251.58 ms /   123 tokens (  213.43
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   10816.63 ms /    39 runs   (  277.35
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   37228.04 ms /   162 tokens
No. of rows:  31%| | 393/1258 [3:37:18<8:03:09, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.02 ms /    35 runs   (    0.26
ms per token,  3880.27 tokens per second)
llama_print_timings: prompt eval time =   24642.53 ms /   113 tokens (  218.08
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    9357.60 ms /    34 runs   (  275.22
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   34135.15 ms /   147 tokens
No. of rows:  31%| | 394/1258 [3:37:53<8:05:18, 33Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.68 ms /    40 runs   (    0.27
ms per token,  3747.07 tokens per second)
llama_print_timings: prompt eval time =   18877.92 ms /    86 tokens (  219.51
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   10578.49 ms /    39 runs   (  271.24
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   29618.26 ms /   125 tokens
No. of rows:  31%| | 395/1258 [3:38:22<7:47:10, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.60 ms /    32 runs   (    0.27
ms per token,  3721.36 tokens per second)
llama_print_timings: prompt eval time =   20550.53 ms /    94 tokens (  218.62
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    8309.72 ms /    31 runs   (  268.06
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   28987.27 ms /   125 tokens
No. of rows:  31%| | 396/1258 [3:38:51<7:31:36, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.54 ms /    25 runs   (    0.26
ms per token,  3823.21 tokens per second)
llama_print_timings: prompt eval time =   17460.02 ms /    81 tokens (  215.56
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    6510.05 ms /    24 runs   (  271.25
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   24065.99 ms /   105 tokens
No. of rows:  32%| | 397/1258 [3:39:15<6:59:23, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.03 ms /    48 runs   (    0.25
ms per token,  3988.70 tokens per second)
llama_print_timings: prompt eval time =   22620.13 ms /   106 tokens (  213.40
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   12745.95 ms /    47 runs   (  271.19
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   35555.18 ms /   153 tokens
No. of rows:  32%| | 398/1258 [3:39:51<7:26:09, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.32 ms /    39 runs   (    0.29
ms per token,  3444.93 tokens per second)
```

```
llama_print_timings: prompt eval time =   21726.89 ms /   101 tokens (  215.12
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   10960.25 ms /    38 runs   (  288.43
ms per token,     3.47 tokens per second)
llama_print_timings:       total time =   32849.40 ms /   139 tokens
No. of rows:  32%| | 399/1258 [3:40:24<7:33:04, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.26 ms /    42 runs   (    0.27
ms per token,  3729.36 tokens per second)
llama_print_timings: prompt eval time =   24609.32 ms /   106 tokens (  232.16
ms per token,     4.31 tokens per second)
llama_print_timings:        eval time =   11398.88 ms /    41 runs   (  278.02
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   36181.54 ms /   147 tokens
No. of rows:  32%| | 400/1258 [3:41:00<7:52:01, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.86 ms /    50 runs   (    0.28
ms per token,  3606.72 tokens per second)
llama_print_timings: prompt eval time =   26233.86 ms /   122 tokens (  215.03
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13733.43 ms /    49 runs   (  280.27
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   40168.19 ms /   171 tokens
No. of rows:  32%| | 401/1258 [3:41:40<8:22:11, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.30 ms /    33 runs   (    0.25
ms per token,  3973.51 tokens per second)
llama_print_timings: prompt eval time =   21246.04 ms /    99 tokens (  214.61
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    8800.11 ms /    32 runs   (  275.00
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   30173.42 ms /   131 tokens
No. of rows:  32%| | 402/1258 [3:42:10<8:00:17, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.27 ms /    35 runs   (    0.29
ms per token,  3408.98 tokens per second)
llama_print_timings: prompt eval time =   22733.83 ms /    98 tokens (  231.98
ms per token,     4.31 tokens per second)
llama_print_timings:        eval time =   10856.09 ms /    34 runs   (  319.30
ms per token,     3.13 tokens per second)
```

```
llama_print_timings:       total time =   33745.01 ms /   132 tokens
No. of rows:  32%| | 403/1258 [3:42:44<8:00:03, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.82 ms /    46 runs   (    0.26
ms per token,  3891.38 tokens per second)
llama_print_timings: prompt eval time =   29991.23 ms /   140 tokens (  214.22
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   11910.07 ms /    45 runs   (  264.67
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   42073.94 ms /   185 tokens
No. of rows:  32%| | 404/1258 [3:43:26<8:35:23, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.43 ms /    50 runs   (    0.25
ms per token,  4022.53 tokens per second)
llama_print_timings: prompt eval time =   22869.40 ms /   107 tokens (  213.73
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13005.68 ms /    49 runs   (  265.42
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   36064.36 ms /   156 tokens
No. of rows:  32%| | 405/1258 [3:44:02<8:34:11, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.43 ms /    39 runs   (    0.27
ms per token,  3738.86 tokens per second)
llama_print_timings: prompt eval time =   25361.92 ms /   119 tokens (  213.13
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   10301.06 ms /    38 runs   (  271.08
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   35812.34 ms /   157 tokens
No. of rows:  32%| | 406/1258 [3:44:38<8:32:06, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.08 ms /    40 runs   (    0.25
ms per token,  3967.07 tokens per second)
llama_print_timings: prompt eval time =   21875.60 ms /   105 tokens (  208.34
ms per token,     4.80 tokens per second)
llama_print_timings:        eval time =   10481.74 ms /    39 runs   (  268.76
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   32508.38 ms /   144 tokens
No. of rows:  32%| | 407/1258 [3:45:10<8:16:24, 35Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       4.94 ms /    19 runs   (    0.26
ms per token,  3849.27 tokens per second)
llama_print_timings: prompt eval time =   14308.88 ms /    65 tokens (  220.14
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    4955.21 ms /    18 runs   (  275.29
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   19337.19 ms /    83 tokens
No. of rows:  32%| | 408/1258 [3:45:30<7:09:17, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.05 ms /    50 runs   (    0.26
ms per token,  3832.00 tokens per second)
llama_print_timings: prompt eval time =   19667.01 ms /    89 tokens (  220.98
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =   13220.41 ms /    49 runs   (  269.80
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   33085.27 ms /   138 tokens
No. of rows:  33%| | 409/1258 [3:46:03<7:20:37, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.78 ms /    38 runs   (    0.26
ms per token,  3886.28 tokens per second)
llama_print_timings: prompt eval time =   25246.74 ms /   117 tokens (  215.78
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   10267.99 ms /    37 runs   (  277.51
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   35664.55 ms /   154 tokens
No. of rows:  33%| | 410/1258 [3:46:39<7:39:15, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.10 ms /    33 runs   (    0.28
ms per token,  3624.78 tokens per second)
llama_print_timings: prompt eval time =   22635.47 ms /   106 tokens (  213.54
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    8635.43 ms /    32 runs   (  269.86
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   31399.46 ms /   138 tokens
No. of rows:  33%| | 411/1258 [3:47:10<7:34:10, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.75 ms /    29 runs   (    0.27
ms per token,  3740.01 tokens per second)
llama_print_timings: prompt eval time =   18714.99 ms /    86 tokens (  217.62
```

```
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    7743.69 ms /    28 runs   (  276.56
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   26573.56 ms /   114 tokens
No. of rows:  33%| | 412/1258 [3:47:37<7:09:55, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.89 ms /    33 runs   (    0.27
ms per token,  3712.04 tokens per second)
llama_print_timings: prompt eval time =   21445.88 ms /    93 tokens (  230.60
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =    8766.01 ms /    32 runs   (  273.94
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   30340.11 ms /   125 tokens
No. of rows:  33%| | 413/1258 [3:48:07<7:08:52, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.06 ms /    50 runs   (    0.26
ms per token,  3828.19 tokens per second)
llama_print_timings: prompt eval time =   28341.04 ms /   132 tokens (  214.70
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13316.58 ms /    49 runs   (  271.77
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   41857.45 ms /   181 tokens
No. of rows:  33%| | 414/1258 [3:48:49<7:56:31, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.99 ms /    50 runs   (    0.28
ms per token,  3573.73 tokens per second)
llama_print_timings: prompt eval time =   25095.67 ms /   110 tokens (  228.14
ms per token,     4.38 tokens per second)
llama_print_timings:        eval time =   13417.00 ms /    49 runs   (  273.82
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   38709.75 ms /   159 tokens
No. of rows:  33%| | 415/1258 [3:49:27<8:16:21, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.95 ms /    26 runs   (    0.27
ms per token,  3742.08 tokens per second)
llama_print_timings: prompt eval time =   21170.57 ms /    90 tokens (  235.23
ms per token,     4.25 tokens per second)
llama_print_timings:        eval time =    6917.71 ms /    25 runs   (  276.71
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   28189.92 ms /   115 tokens
```

```
No. of rows:  33%| | 416/1258 [3:49:56<7:45:45, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.92 ms /    32 runs   (    0.28
ms per token,  3589.05 tokens per second)
llama_print_timings: prompt eval time =   22500.01 ms /   103 tokens (  218.45
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    8817.44 ms /    31 runs   (  284.43
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   31446.60 ms /   134 tokens
No. of rows:  33%| | 417/1258 [3:50:27<7:37:54, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.46 ms /    30 runs   (    0.28
ms per token,  3545.68 tokens per second)
llama_print_timings: prompt eval time =   19588.56 ms /    83 tokens (  236.01
ms per token,     4.24 tokens per second)
llama_print_timings:        eval time =    8034.77 ms /    29 runs   (  277.06
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   27741.71 ms /   112 tokens
No. of rows:  33%| | 418/1258 [3:50:55<7:16:41, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.44 ms /    39 runs   (    0.27
ms per token,  3737.42 tokens per second)
llama_print_timings: prompt eval time =   26010.65 ms /   116 tokens (  224.23
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =   10262.71 ms /    38 runs   (  270.07
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   36427.87 ms /   154 tokens
No. of rows:  33%| | 419/1258 [3:51:31<7:38:07, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.65 ms /    43 runs   (    0.27
ms per token,  3691.62 tokens per second)
llama_print_timings: prompt eval time =   21848.63 ms /   100 tokens (  218.49
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =   11508.84 ms /    42 runs   (  274.02
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   33525.56 ms /   142 tokens
No. of rows:  33%| | 420/1258 [3:52:05<7:40:49, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =       8.42 ms /    30 runs   (    0.28
ms per token,  3561.25 tokens per second)
llama_print_timings: prompt eval time =   21047.14 ms /    97 tokens (  216.98
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    8245.95 ms /    29 runs   (  284.34
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   29412.96 ms /   126 tokens
No. of rows:  33%| | 421/1258 [3:52:34<7:25:18, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.22 ms /    31 runs   (    0.27
ms per token,  3769.00 tokens per second)
llama_print_timings: prompt eval time =   20412.47 ms /    95 tokens (  214.87
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9800.24 ms /    30 runs   (  326.67
ms per token,     3.06 tokens per second)
llama_print_timings:       total time =   30334.16 ms /   125 tokens
No. of rows:  34%| | 422/1258 [3:53:05<7:18:09, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.92 ms /    50 runs   (    0.28
ms per token,  3592.21 tokens per second)
llama_print_timings: prompt eval time =   20152.78 ms /    94 tokens (  214.39
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   15122.97 ms /    49 runs   (  308.63
ms per token,     3.24 tokens per second)
llama_print_timings:       total time =   35474.57 ms /   143 tokens
No. of rows:  34%| | 423/1258 [3:53:40<7:34:31, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.30 ms /    46 runs   (    0.27
ms per token,  3739.53 tokens per second)
llama_print_timings: prompt eval time =   25105.92 ms /   114 tokens (  220.23
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =   12347.34 ms /    45 runs   (  274.39
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   37632.99 ms /   159 tokens
No. of rows:  34%| | 424/1258 [3:54:18<7:54:44, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.52 ms /    29 runs   (    0.26
ms per token,  3854.33 tokens per second)
llama_print_timings: prompt eval time =   19555.57 ms /    87 tokens (  224.78
ms per token,     4.45 tokens per second)
```

```
llama_print_timings:        eval time =    7662.40 ms /    28 runs   (  273.66
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   27333.36 ms /   115 tokens
No. of rows:  34%| | 425/1258 [3:54:45<7:25:45, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.34 ms /    37 runs   (    0.28
ms per token,  3578.68 tokens per second)
llama_print_timings: prompt eval time =   22785.06 ms /   106 tokens (  214.95
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9893.84 ms /    36 runs   (  274.83
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   32826.46 ms /   142 tokens
No. of rows:  34%| | 426/1258 [3:55:18<7:28:16, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.88 ms /    50 runs   (    0.28
ms per token,  3602.31 tokens per second)
llama_print_timings: prompt eval time =   26931.62 ms /   117 tokens (  230.18
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =   13925.97 ms /    49 runs   (  284.20
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   41055.78 ms /   166 tokens
No. of rows:  34%| | 427/1258 [3:55:59<8:04:02, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.91 ms /    22 runs   (    0.27
ms per token,  3719.99 tokens per second)
llama_print_timings: prompt eval time =   18423.99 ms /    84 tokens (  219.33
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    5663.08 ms /    21 runs   (  269.67
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   24171.74 ms /   105 tokens
No. of rows:  34%| | 428/1258 [3:56:23<7:18:46, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.84 ms /    50 runs   (    0.26
ms per token,  3893.17 tokens per second)
llama_print_timings: prompt eval time =   25105.68 ms /   111 tokens (  226.18
ms per token,     4.42 tokens per second)
llama_print_timings:        eval time =   15155.45 ms /    49 runs   (  309.29
ms per token,     3.23 tokens per second)
llama_print_timings:       total time =   40459.62 ms /   160 tokens
No. of rows:  34%| | 429/1258 [3:57:04<7:54:31, 34Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.86 ms /    39 runs   (    0.28
ms per token,  3592.48 tokens per second)
llama_print_timings: prompt eval time =   24221.74 ms /   111 tokens (  218.21
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =   10288.90 ms /    38 runs   (  270.76
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   34667.10 ms /   149 tokens
No. of rows:  34%| | 430/1258 [3:57:38<7:55:17, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.03 ms /    50 runs   (    0.28
ms per token,  3564.05 tokens per second)
llama_print_timings: prompt eval time =   25075.56 ms /   115 tokens (  218.05
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =   13546.04 ms /    49 runs   (  276.45
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   38818.99 ms /   164 tokens
No. of rows:  34%| | 431/1258 [3:58:17<8:12:51, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.00 ms /    26 runs   (    0.27
ms per token,  3712.69 tokens per second)
llama_print_timings: prompt eval time =   22033.33 ms /   102 tokens (  216.01
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    6936.61 ms /    25 runs   (  277.46
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   29069.75 ms /   127 tokens
No. of rows:  34%| | 432/1258 [3:58:46<7:44:41, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.31 ms /    27 runs   (    0.27
ms per token,  3695.09 tokens per second)
llama_print_timings: prompt eval time =   18060.31 ms /    83 tokens (  217.59
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    7076.20 ms /    26 runs   (  272.16
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   25241.55 ms /   109 tokens
No. of rows:  34%| | 433/1258 [3:59:11<7:09:00, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.17 ms /    31 runs   (    0.26
```

```
ms per token,  3795.30 tokens per second)
llama_print_timings: prompt eval time =   22609.65 ms /   103 tokens (  219.51
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    8082.38 ms /    30 runs   (  269.41
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   30812.03 ms /   133 tokens
No. of rows:  34%| | 434/1258 [3:59:42<7:06:57, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.85 ms /    33 runs   (    0.27
ms per token,  3730.08 tokens per second)
llama_print_timings: prompt eval time =   23547.40 ms /   102 tokens (  230.86
ms per token,     4.33 tokens per second)
llama_print_timings:        eval time =    8847.97 ms /    32 runs   (  276.50
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   32523.13 ms /   134 tokens
No. of rows:  35%| | 435/1258 [4:00:15<7:12:22, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.08 ms /    50 runs   (    0.26
ms per token,  3822.63 tokens per second)
llama_print_timings: prompt eval time =   31789.45 ms /   142 tokens (  223.87
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   13429.60 ms /    49 runs   (  274.07
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   45422.86 ms /   191 tokens
No. of rows:  35%| | 436/1258 [4:01:00<8:09:00, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.87 ms /    25 runs   (    0.27
ms per token,  3639.01 tokens per second)
llama_print_timings: prompt eval time =   21035.57 ms /    97 tokens (  216.86
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    6565.29 ms /    24 runs   (  273.55
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   27701.43 ms /   121 tokens
No. of rows:  35%| | 437/1258 [4:01:28<7:35:38, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.91 ms /    29 runs   (    0.27
ms per token,  3667.17 tokens per second)
llama_print_timings: prompt eval time =   19349.22 ms /    90 tokens (  214.99
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    7995.26 ms /    28 runs   (  285.55
```

```
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   27460.55 ms /   118 tokens
No. of rows:  35%| | 438/1258 [4:01:55<7:11:10, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.64 ms /    38 runs   (    0.28
ms per token,  3571.09 tokens per second)
llama_print_timings: prompt eval time =   23329.20 ms /   109 tokens (  214.03
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   10336.47 ms /    37 runs   (  279.36
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   33817.07 ms /   146 tokens
No. of rows:  35%| | 439/1258 [4:02:29<7:19:57, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.61 ms /    34 runs   (    0.28
ms per token,  3539.09 tokens per second)
llama_print_timings: prompt eval time =   20133.24 ms /    92 tokens (  218.84
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =   10831.45 ms /    33 runs   (  328.23
ms per token,     3.05 tokens per second)
llama_print_timings:       total time =   31097.14 ms /   125 tokens
No. of rows:  35%| | 440/1258 [4:03:00<7:14:48, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.79 ms /    36 runs   (    0.27
ms per token,  3677.97 tokens per second)
llama_print_timings: prompt eval time =   18354.53 ms /    85 tokens (  215.94
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    9661.98 ms /    35 runs   (  276.06
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   28155.45 ms /   120 tokens
No. of rows:  35%| | 441/1258 [4:03:28<6:58:58, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.35 ms /    35 runs   (    0.27
ms per token,  3744.52 tokens per second)
llama_print_timings: prompt eval time =   26500.89 ms /   116 tokens (  228.46
ms per token,     4.38 tokens per second)
llama_print_timings:        eval time =    9714.37 ms /    34 runs   (  285.72
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   36352.58 ms /   150 tokens
No. of rows:  35%| | 442/1258 [4:04:05<7:21:16, 32Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.99 ms /    50 runs   (    0.26
ms per token,  3850.30 tokens per second)
llama_print_timings: prompt eval time =   27748.48 ms /   129 tokens (  215.10
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13423.02 ms /    49 runs   (  273.94
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   41367.17 ms /   178 tokens
No. of rows:  35%| | 443/1258 [4:04:46<7:57:10, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.68 ms /    50 runs   (    0.27
ms per token,  3654.70 tokens per second)
llama_print_timings: prompt eval time =   30894.87 ms /   139 tokens (  222.27
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =   13647.55 ms /    49 runs   (  278.52
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   44739.33 ms /   188 tokens
No. of rows:  35%| | 444/1258 [4:05:31<8:35:43, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.97 ms /    50 runs   (    0.28
ms per token,  3579.87 tokens per second)
llama_print_timings: prompt eval time =   25616.31 ms /   113 tokens (  226.69
ms per token,     4.41 tokens per second)
llama_print_timings:        eval time =   13396.77 ms /    49 runs   (  273.40
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   39209.91 ms /   162 tokens
No. of rows:  35%| | 445/1258 [4:06:10<8:39:59, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.45 ms /    31 runs   (    0.27
ms per token,  3670.81 tokens per second)
llama_print_timings: prompt eval time =   19136.96 ms /    88 tokens (  217.47
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    8421.43 ms /    30 runs   (  280.71
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   27682.33 ms /   118 tokens
No. of rows:  35%| | 446/1258 [4:06:38<7:55:57, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.27 ms /    37 runs   (    0.28
ms per token,  3601.67 tokens per second)
```

```
llama_print_timings: prompt eval time =   18735.97 ms /    87 tokens (  215.36
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    9785.92 ms /    36 runs   (  271.83
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   28665.03 ms /   123 tokens
No. of rows:  36%| | 447/1258 [4:07:07<7:28:59, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.72 ms /    23 runs   (    0.25
ms per token,  4022.39 tokens per second)
llama_print_timings: prompt eval time =   16233.57 ms /    73 tokens (  222.38
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =    5985.63 ms /    22 runs   (  272.07
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   22306.56 ms /    95 tokens
No. of rows:  36%| | 448/1258 [4:07:29<6:44:19, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.09 ms /    24 runs   (    0.38
ms per token,  2641.14 tokens per second)
llama_print_timings: prompt eval time =   19343.56 ms /    89 tokens (  217.34
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    6331.19 ms /    23 runs   (  275.27
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   25772.16 ms /   112 tokens
No. of rows:  36%| | 449/1258 [4:07:55<6:26:57, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.03 ms /    34 runs   (    0.27
ms per token,  3763.56 tokens per second)
llama_print_timings: prompt eval time =   19318.33 ms /    90 tokens (  214.65
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    8956.04 ms /    33 runs   (  271.40
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   28407.46 ms /   123 tokens
No. of rows:  36%| | 450/1258 [4:08:23<6:25:20, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.34 ms /    38 runs   (    0.27
ms per token,  3675.05 tokens per second)
llama_print_timings: prompt eval time =   18682.70 ms /    86 tokens (  217.24
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =   10316.43 ms /    37 runs   (  278.82
ms per token,     3.59 tokens per second)
```

```
llama_print_timings:       total time =   29152.98 ms /   123 tokens
No. of rows:  36%| | 451/1258 [4:08:52<6:27:03, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.02 ms /    42 runs   (    0.26
ms per token,  3810.91 tokens per second)
llama_print_timings: prompt eval time =   26564.78 ms /   118 tokens (  225.13
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =   11221.75 ms /    41 runs   (  273.70
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   37953.10 ms /   159 tokens
No. of rows:  36%| | 452/1258 [4:09:30<7:03:35, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.43 ms /    35 runs   (    0.30
ms per token,  3354.42 tokens per second)
llama_print_timings: prompt eval time =   21031.13 ms /    97 tokens (  216.82
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    9639.32 ms /    34 runs   (  283.51
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   30817.38 ms /   131 tokens
No. of rows:  36%| | 453/1258 [4:10:01<7:00:12, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.45 ms /    50 runs   (    0.27
ms per token,  3716.64 tokens per second)
llama_print_timings: prompt eval time =   22639.88 ms /   103 tokens (  219.80
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   13867.07 ms /    49 runs   (  283.00
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   36705.54 ms /   152 tokens
No. of rows:  36%| | 454/1258 [4:10:38<7:21:22, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.93 ms /    27 runs   (    0.26
ms per token,  3896.10 tokens per second)
llama_print_timings: prompt eval time =   22036.10 ms /   102 tokens (  216.04
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    7077.32 ms /    26 runs   (  272.20
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   29221.50 ms /   128 tokens
No. of rows:  36%| | 455/1258 [4:11:07<7:05:56, 31Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.91 ms /    50 runs   (    0.26
ms per token,  3872.67 tokens per second)
llama_print_timings: prompt eval time =   26638.12 ms /   125 tokens (  213.10
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   13474.90 ms /    49 runs   (  275.00
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   40311.29 ms /   174 tokens
No. of rows:  36%| | 456/1258 [4:11:47<7:39:27, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.67 ms /    25 runs   (    0.27
ms per token,  3750.38 tokens per second)
llama_print_timings: prompt eval time =   17944.74 ms /    81 tokens (  221.54
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    6775.97 ms /    24 runs   (  282.33
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   24817.34 ms /   105 tokens
No. of rows:  36%| | 457/1258 [4:12:12<7:00:40, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.34 ms /    24 runs   (    0.26
ms per token,  3786.68 tokens per second)
llama_print_timings: prompt eval time =   20208.68 ms /    93 tokens (  217.30
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    6350.35 ms /    23 runs   (  276.10
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   26651.16 ms /   116 tokens
No. of rows:  36%| | 458/1258 [4:12:39<6:40:44, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.30 ms /    24 runs   (    0.26
ms per token,  3808.31 tokens per second)
llama_print_timings: prompt eval time =   17459.52 ms /    79 tokens (  221.01
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    6241.88 ms /    23 runs   (  271.39
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   23792.45 ms /   102 tokens
No. of rows:  36%| | 459/1258 [4:13:03<6:15:14, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.69 ms /    46 runs   (    0.28
ms per token,  3626.33 tokens per second)
llama_print_timings: prompt eval time =   25314.96 ms /   118 tokens (  214.53
```

```
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   12398.40 ms /    45 runs   (  275.52
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   37895.67 ms /   163 tokens
No. of rows:  37%| | 460/1258 [4:13:40<6:53:34, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.15 ms /    30 runs   (    0.27
ms per token,  3682.34 tokens per second)
llama_print_timings: prompt eval time =   19583.33 ms /    90 tokens (  217.59
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    8202.22 ms /    29 runs   (  282.84
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   27907.34 ms /   119 tokens
No. of rows:  37%| | 461/1258 [4:14:08<6:40:22, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.80 ms /    36 runs   (    0.27
ms per token,  3674.59 tokens per second)
llama_print_timings: prompt eval time =   24050.82 ms /   111 tokens (  216.67
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    9710.56 ms /    35 runs   (  277.44
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   33903.64 ms /   146 tokens
No. of rows:  37%| | 462/1258 [4:14:42<6:54:53, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.17 ms /    27 runs   (    0.27
ms per token,  3763.07 tokens per second)
llama_print_timings: prompt eval time =   19150.21 ms /    89 tokens (  215.17
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    7129.16 ms /    26 runs   (  274.20
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   26382.21 ms /   115 tokens
No. of rows:  37%| | 463/1258 [4:15:09<6:34:57, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.91 ms /    50 runs   (    0.28
ms per token,  3593.50 tokens per second)
llama_print_timings: prompt eval time =   23025.15 ms /   100 tokens (  230.25
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =   13212.70 ms /    49 runs   (  269.65
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   36436.17 ms /   149 tokens
```

```
No. of rows:  37%| | 464/1258 [4:15:45<7:00:47, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.66 ms /    24 runs   (    0.28
ms per token,  3601.98 tokens per second)
llama_print_timings: prompt eval time =   18255.93 ms /    83 tokens (  219.95
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =    6387.10 ms /    23 runs   (  277.70
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   24737.12 ms /   106 tokens
No. of rows:  37%| | 465/1258 [4:16:10<6:32:16, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.11 ms /    34 runs   (    0.27
ms per token,  3730.93 tokens per second)
llama_print_timings: prompt eval time =   20523.35 ms /    94 tokens (  218.33
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    9377.58 ms /    33 runs   (  284.17
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   30034.54 ms /   127 tokens
No. of rows:  37%| | 466/1258 [4:16:40<6:33:13, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.57 ms /    24 runs   (    0.27
ms per token,  3652.41 tokens per second)
llama_print_timings: prompt eval time =   18740.13 ms /    86 tokens (  217.91
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    6313.51 ms /    23 runs   (  274.50
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   25149.73 ms /   109 tokens
No. of rows:  37%| | 467/1258 [4:17:05<6:14:24, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.20 ms /    31 runs   (    0.26
ms per token,  3781.41 tokens per second)
llama_print_timings: prompt eval time =   20937.82 ms /    96 tokens (  218.10
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    8192.32 ms /    30 runs   (  273.08
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   29251.03 ms /   126 tokens
No. of rows:  37%| | 468/1258 [4:17:34<6:17:19, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      11.19 ms /    41 runs   (    0.27
ms per token,  3663.33 tokens per second)
llama_print_timings: prompt eval time =   18964.31 ms /    87 tokens (  217.98
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =   10947.26 ms /    40 runs   (  273.68
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   30071.60 ms /   127 tokens
No. of rows:  37%| | 469/1258 [4:18:04<6:22:26, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.27 ms /    35 runs   (    0.26
ms per token,  3773.99 tokens per second)
llama_print_timings: prompt eval time =   21297.04 ms /    99 tokens (  215.12
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9652.42 ms /    34 runs   (  283.89
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   31093.20 ms /   133 tokens
No. of rows:  37%| | 470/1258 [4:18:35<6:29:54, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.12 ms /    49 runs   (    0.27
ms per token,  3735.33 tokens per second)
llama_print_timings: prompt eval time =   31790.35 ms /   148 tokens (  214.80
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13089.01 ms /    48 runs   (  272.69
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   45070.45 ms /   196 tokens
No. of rows:  37%| | 471/1258 [4:19:21<7:29:55, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.60 ms /    36 runs   (    0.27
ms per token,  3750.39 tokens per second)
llama_print_timings: prompt eval time =   21861.26 ms /   102 tokens (  214.33
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   11272.85 ms /    35 runs   (  322.08
ms per token,     3.10 tokens per second)
llama_print_timings:       total time =   33274.73 ms /   137 tokens
No. of rows:  38%| | 472/1258 [4:19:54<7:25:23, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.14 ms /    50 runs   (    0.26
ms per token,  3803.73 tokens per second)
llama_print_timings: prompt eval time =   23499.97 ms /   107 tokens (  219.63
ms per token,     4.55 tokens per second)
```

```
llama_print_timings:        eval time =   13911.68 ms /    49 runs   (  283.91
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   37612.26 ms /   156 tokens
No. of rows:  38%| | 473/1258 [4:20:31<7:39:01, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.33 ms /    45 runs   (    0.27
ms per token,  3649.34 tokens per second)
llama_print_timings: prompt eval time =   25882.50 ms /   119 tokens (  217.50
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =   12001.85 ms /    44 runs   (  272.77
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   38060.92 ms /   163 tokens
No. of rows:  38%| | 474/1258 [4:21:10<7:50:05, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.64 ms /    50 runs   (    0.27
ms per token,  3664.88 tokens per second)
llama_print_timings: prompt eval time =   28848.46 ms /   128 tokens (  225.38
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =   13508.64 ms /    49 runs   (  275.69
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   42556.28 ms /   177 tokens
No. of rows:  38%| | 475/1258 [4:21:52<8:15:19, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.86 ms /    32 runs   (    0.28
ms per token,  3613.37 tokens per second)
llama_print_timings: prompt eval time =   22404.36 ms /    97 tokens (  230.97
ms per token,     4.33 tokens per second)
llama_print_timings:        eval time =    8296.41 ms /    31 runs   (  267.63
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   30826.22 ms /   128 tokens
No. of rows:  38%| | 476/1258 [4:22:23<7:46:47, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.78 ms /    38 runs   (    0.26
ms per token,  3885.08 tokens per second)
llama_print_timings: prompt eval time =   20337.53 ms /    90 tokens (  225.97
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   11818.68 ms /    37 runs   (  319.42
ms per token,     3.13 tokens per second)
llama_print_timings:       total time =   32302.67 ms /   127 tokens
No. of rows:  38%| | 477/1258 [4:22:55<7:32:33, 34Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.43 ms /    43 runs   (    0.34
ms per token,  2980.52 tokens per second)
llama_print_timings: prompt eval time =   25446.66 ms /   105 tokens (  242.35
ms per token,     4.13 tokens per second)
llama_print_timings:        eval time =   13427.50 ms /    42 runs   (  319.70
ms per token,     3.13 tokens per second)
llama_print_timings:       total time =   39075.40 ms /   147 tokens
No. of rows:  38%| | 478/1258 [4:23:34<7:48:46, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.14 ms /    33 runs   (    0.37
ms per token,  2718.73 tokens per second)
llama_print_timings: prompt eval time =   25013.72 ms /   103 tokens (  242.85
ms per token,     4.12 tokens per second)
llama_print_timings:        eval time =   12405.94 ms /    32 runs   (  387.69
ms per token,     2.58 tokens per second)
llama_print_timings:       total time =   37609.37 ms /   135 tokens
No. of rows:  38%| | 479/1258 [4:24:12<7:54:17, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.42 ms /    43 runs   (    0.34
ms per token,  2981.56 tokens per second)
llama_print_timings: prompt eval time =   27864.65 ms /    98 tokens (  284.33
ms per token,     3.52 tokens per second)
llama_print_timings:        eval time =   14250.19 ms /    42 runs   (  339.29
ms per token,     2.95 tokens per second)
llama_print_timings:       total time =   42321.38 ms /   140 tokens
No. of rows:  38%| | 480/1258 [4:24:54<8:16:14, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.47 ms /    50 runs   (    0.27
ms per token,  3711.40 tokens per second)
llama_print_timings: prompt eval time =   26498.52 ms /   119 tokens (  222.68
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =   13593.77 ms /    49 runs   (  277.42
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   40290.26 ms /   168 tokens
No. of rows:  38%| | 481/1258 [4:25:35<8:23:26, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.99 ms /    31 runs   (    0.26
```

```
ms per token,  3878.39 tokens per second)
llama_print_timings: prompt eval time =   19468.33 ms /    89 tokens (  218.75
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    8198.76 ms /    30 runs   (  273.29
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   27787.16 ms /   119 tokens
No. of rows:  38%| | 482/1258 [4:26:02<7:39:50, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.68 ms /    26 runs   (    0.26
ms per token,  3892.80 tokens per second)
llama_print_timings: prompt eval time =   20345.15 ms /    87 tokens (  233.85
ms per token,     4.28 tokens per second)
llama_print_timings:        eval time =    6956.28 ms /    25 runs   (  278.25
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   27398.70 ms /   112 tokens
No. of rows:  38%| | 483/1258 [4:26:30<7:07:40, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.07 ms /    38 runs   (    0.27
ms per token,  3772.84 tokens per second)
llama_print_timings: prompt eval time =   24619.39 ms /   112 tokens (  219.82
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   10225.51 ms /    37 runs   (  276.37
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   34990.54 ms /   149 tokens
No. of rows:  38%| | 484/1258 [4:27:05<7:14:25, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.75 ms /    37 runs   (    0.26
ms per token,  3794.09 tokens per second)
llama_print_timings: prompt eval time =   23239.43 ms /   108 tokens (  215.18
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9744.55 ms /    36 runs   (  270.68
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   33127.91 ms /   144 tokens
No. of rows:  39%| | 485/1258 [4:27:38<7:11:46, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.27 ms /    35 runs   (    0.26
ms per token,  3777.25 tokens per second)
llama_print_timings: prompt eval time =   22384.94 ms /    99 tokens (  226.11
ms per token,     4.42 tokens per second)
llama_print_timings:        eval time =    9203.96 ms /    34 runs   (  270.70
```

```
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   31723.73 ms /   133 tokens
No. of rows:  39%| | 486/1258 [4:28:10<7:04:20, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.78 ms /    37 runs   (    0.29
ms per token,  3433.24 tokens per second)
llama_print_timings: prompt eval time =   22390.33 ms /   105 tokens (  213.24
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   10062.16 ms /    36 runs   (  279.50
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   32601.94 ms /   141 tokens
No. of rows:  39%| | 487/1258 [4:28:42<7:02:21, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.26 ms /    38 runs   (    0.27
ms per token,  3704.43 tokens per second)
llama_print_timings: prompt eval time =   22255.95 ms /   103 tokens (  216.08
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   10061.20 ms /    37 runs   (  271.92
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   32467.81 ms /   140 tokens
No. of rows:  39%| | 488/1258 [4:29:15<7:00:17, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.31 ms /    32 runs   (    0.26
ms per token,  3851.25 tokens per second)
llama_print_timings: prompt eval time =   21097.33 ms /    98 tokens (  215.28
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    8347.74 ms /    31 runs   (  269.28
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   29567.43 ms /   129 tokens
No. of rows:  39%| | 489/1258 [4:29:44<6:47:31, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.35 ms /    44 runs   (    0.30
ms per token,  3296.87 tokens per second)
llama_print_timings: prompt eval time =   24525.20 ms /   109 tokens (  225.00
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =   12333.24 ms /    43 runs   (  286.82
ms per token,     3.49 tokens per second)
llama_print_timings:       total time =   37047.62 ms /   152 tokens
No. of rows:  39%| | 490/1258 [4:30:21<7:07:12, 33Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.76 ms /    25 runs   (    0.27
ms per token,  3698.77 tokens per second)
llama_print_timings: prompt eval time =   19872.32 ms /    90 tokens (  220.80
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    6553.09 ms /    24 runs   (  273.05
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   26524.72 ms /   114 tokens
No. of rows:  39%| | 491/1258 [4:30:48<6:40:24, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.06 ms /    41 runs   (    0.27
ms per token,  3706.05 tokens per second)
llama_print_timings: prompt eval time =   27212.31 ms /   127 tokens (  214.27
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   10925.82 ms /    40 runs   (  273.15
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   38303.91 ms /   167 tokens
No. of rows:  39%| | 492/1258 [4:31:26<7:06:36, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.05 ms /    39 runs   (    0.26
ms per token,  3879.05 tokens per second)
llama_print_timings: prompt eval time =   22431.93 ms /   104 tokens (  215.69
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   12007.86 ms /    38 runs   (  316.00
ms per token,     3.16 tokens per second)
llama_print_timings:       total time =   34590.20 ms /   142 tokens
No. of rows:  39%| | 493/1258 [4:32:01<7:10:35, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.87 ms /    40 runs   (    0.30
ms per token,  3369.84 tokens per second)
llama_print_timings: prompt eval time =   28343.24 ms /   130 tokens (  218.02
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =   11258.82 ms /    39 runs   (  288.69
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   39767.13 ms /   169 tokens
No. of rows:  39%| | 494/1258 [4:32:41<7:32:56, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.71 ms /    48 runs   (    0.26
ms per token,  3776.55 tokens per second)
```

```
llama_print_timings: prompt eval time =   23696.19 ms /   111 tokens (  213.48
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   12865.47 ms /    47 runs   (  273.73
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   36750.85 ms /   158 tokens
No. of rows:  39%| | 495/1258 [4:33:17<7:36:54, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.30 ms /    36 runs   (    0.29
ms per token,  3495.48 tokens per second)
llama_print_timings: prompt eval time =   20090.78 ms /    93 tokens (  216.03
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    9513.15 ms /    35 runs   (  271.80
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   29751.29 ms /   128 tokens
No. of rows:  39%| | 496/1258 [4:33:47<7:12:44, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.28 ms /    28 runs   (    0.26
ms per token,  3846.15 tokens per second)
llama_print_timings: prompt eval time =   19507.35 ms /    89 tokens (  219.18
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    9009.58 ms /    27 runs   (  333.69
ms per token,     3.00 tokens per second)
llama_print_timings:       total time =   28624.39 ms /   116 tokens
No. of rows:  40%| | 497/1258 [4:34:16<6:51:30, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.98 ms /    31 runs   (    0.29
ms per token,  3452.50 tokens per second)
llama_print_timings: prompt eval time =   18304.50 ms /    81 tokens (  225.98
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =    8695.52 ms /    30 runs   (  289.85
ms per token,     3.45 tokens per second)
llama_print_timings:       total time =   27126.29 ms /   111 tokens
No. of rows:  40%| | 498/1258 [4:34:43<6:30:48, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.48 ms /    30 runs   (    0.28
ms per token,  3536.48 tokens per second)
llama_print_timings: prompt eval time =   22040.37 ms /   102 tokens (  216.08
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    7951.29 ms /    29 runs   (  274.18
ms per token,     3.65 tokens per second)
```

```
llama_print_timings:       total time =   30109.25 ms /   131 tokens
No. of rows:  40%| | 499/1258 [4:35:13<6:27:29, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.69 ms /    29 runs   (    0.27
ms per token,  3770.15 tokens per second)
llama_print_timings: prompt eval time =   20452.33 ms /    95 tokens (  215.29
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    9270.04 ms /    28 runs   (  331.07
ms per token,     3.02 tokens per second)
llama_print_timings:       total time =   29833.58 ms /   123 tokens
No. of rows:  40%| | 500/1258 [4:35:43<6:23:59, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.36 ms /    28 runs   (    0.26
ms per token,  3802.80 tokens per second)
llama_print_timings: prompt eval time =   18359.04 ms /    85 tokens (  215.99
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    7291.11 ms /    27 runs   (  270.04
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   25759.71 ms /   112 tokens
No. of rows:  40%| | 501/1258 [4:36:09<6:05:58, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.00 ms /    31 runs   (    0.29
ms per token,  3445.59 tokens per second)
llama_print_timings: prompt eval time =   23137.93 ms /   109 tokens (  212.27
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =    8401.58 ms /    30 runs   (  280.05
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   31666.73 ms /   139 tokens
No. of rows:  40%| | 502/1258 [4:36:40<6:15:34, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.02 ms /    25 runs   (    0.28
ms per token,  3559.73 tokens per second)
llama_print_timings: prompt eval time =   17666.35 ms /    81 tokens (  218.10
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    6578.42 ms /    24 runs   (  274.10
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   24343.32 ms /   105 tokens
No. of rows:  40%| | 503/1258 [4:37:05<5:54:28, 28Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.31 ms /    50 runs   (    0.27
ms per token,  3755.73 tokens per second)
llama_print_timings: prompt eval time =   31756.81 ms /   143 tokens (  222.08
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =   14856.33 ms /    49 runs   (  303.19
ms per token,     3.30 tokens per second)
llama_print_timings:       total time =   46808.74 ms /   192 tokens
No. of rows:  40%| | 504/1258 [4:37:51<7:04:17, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.40 ms /    35 runs   (    0.27
ms per token,  3723.40 tokens per second)
llama_print_timings: prompt eval time =   30172.12 ms /   141 tokens (  213.99
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    9591.08 ms /    34 runs   (  282.09
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   39904.16 ms /   175 tokens
No. of rows:  40%| | 505/1258 [4:38:31<7:26:52, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.49 ms /    50 runs   (    0.27
ms per token,  3707.00 tokens per second)
llama_print_timings: prompt eval time =   26444.68 ms /   121 tokens (  218.55
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =   13422.20 ms /    49 runs   (  273.92
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   40062.73 ms /   170 tokens
No. of rows:  40%| | 506/1258 [4:39:11<7:43:02, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.50 ms /    26 runs   (    0.25
ms per token,  3998.77 tokens per second)
llama_print_timings: prompt eval time =   17394.69 ms /    80 tokens (  217.43
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    6867.64 ms /    25 runs   (  274.71
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   24362.36 ms /   105 tokens
No. of rows:  40%| | 507/1258 [4:39:36<6:55:13, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.78 ms /    36 runs   (    0.27
ms per token,  3681.36 tokens per second)
llama_print_timings: prompt eval time =   22039.93 ms /   101 tokens (  218.22
```

```
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    9626.98 ms /    35 runs   (  275.06
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   31808.94 ms /   136 tokens
No. of rows:  40%| | 508/1258 [4:40:08<6:49:34, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.30 ms /    35 runs   (    0.27
ms per token,  3763.85 tokens per second)
llama_print_timings: prompt eval time =   22587.51 ms /    97 tokens (  232.86
ms per token,     4.29 tokens per second)
llama_print_timings:        eval time =    9463.74 ms /    34 runs   (  278.35
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   32189.98 ms /   131 tokens
No. of rows:  40%| | 509/1258 [4:40:40<6:46:54, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.91 ms /    33 runs   (    0.27
ms per token,  3702.87 tokens per second)
llama_print_timings: prompt eval time =   19857.57 ms /    91 tokens (  218.22
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =   10535.73 ms /    32 runs   (  329.24
ms per token,     3.04 tokens per second)
llama_print_timings:       total time =   30519.46 ms /   123 tokens
No. of rows:  41%| | 510/1258 [4:41:10<6:38:35, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.93 ms /    44 runs   (    0.27
ms per token,  3689.42 tokens per second)
llama_print_timings: prompt eval time =   24818.45 ms /   115 tokens (  215.81
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   11689.57 ms /    43 runs   (  271.85
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   36681.56 ms /   158 tokens
No. of rows:  41%| | 511/1258 [4:41:47<6:55:41, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.96 ms /    31 runs   (    0.39
ms per token,  2593.06 tokens per second)
llama_print_timings: prompt eval time =   20265.04 ms /    87 tokens (  232.93
ms per token,     4.29 tokens per second)
llama_print_timings:        eval time =   10189.63 ms /    30 runs   (  339.65
ms per token,     2.94 tokens per second)
llama_print_timings:       total time =   30581.81 ms /   117 tokens
```

```
No. of rows:  41%| | 512/1258 [4:42:18<6:44:41, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.17 ms /    23 runs   (    0.27
ms per token,  3730.13 tokens per second)
llama_print_timings: prompt eval time =   20255.66 ms /    91 tokens (  222.59
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =    6266.46 ms /    22 runs   (  284.84
ms per token,     3.51 tokens per second)
llama_print_timings:       total time =   26610.41 ms /   113 tokens
No. of rows:  41%| | 513/1258 [4:42:44<6:22:03, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.18 ms /    23 runs   (    0.27
ms per token,  3722.29 tokens per second)
llama_print_timings: prompt eval time =   17566.42 ms /    81 tokens (  216.87
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    5908.60 ms /    22 runs   (  268.57
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   23562.82 ms /   103 tokens
No. of rows:  41%| | 514/1258 [4:43:08<5:54:45, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.32 ms /    35 runs   (    0.27
ms per token,  3754.96 tokens per second)
llama_print_timings: prompt eval time =   23458.23 ms /   110 tokens (  213.26
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    9359.44 ms /    34 runs   (  275.28
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   32953.10 ms /   144 tokens
No. of rows:  41%| | 515/1258 [4:43:41<6:10:26, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.95 ms /    50 runs   (    0.28
ms per token,  3583.46 tokens per second)
llama_print_timings: prompt eval time =   26042.68 ms /   122 tokens (  213.46
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13301.68 ms /    49 runs   (  271.46
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   39544.06 ms /   171 tokens
No. of rows:  41%| | 516/1258 [4:44:20<6:45:40, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      12.01 ms /    44 runs   (    0.27
ms per token,  3663.92 tokens per second)
llama_print_timings: prompt eval time =   21462.10 ms /    97 tokens (  221.26
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =   11847.32 ms /    43 runs   (  275.52
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   33484.60 ms /   140 tokens
No. of rows:  41%| | 517/1258 [4:44:54<6:47:38, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.73 ms /    50 runs   (    0.27
ms per token,  3641.93 tokens per second)
llama_print_timings: prompt eval time =   32013.40 ms /   148 tokens (  216.31
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   13391.01 ms /    49 runs   (  273.29
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   45605.28 ms /   197 tokens
No. of rows:  41%| | 518/1258 [4:45:39<7:33:46, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.04 ms /    30 runs   (    0.27
ms per token,  3729.95 tokens per second)
llama_print_timings: prompt eval time =   21705.94 ms /    93 tokens (  233.40
ms per token,     4.28 tokens per second)
llama_print_timings:        eval time =    7866.75 ms /    29 runs   (  271.27
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   29689.19 ms /   122 tokens
No. of rows:  41%| | 519/1258 [4:46:09<7:06:58, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.42 ms /    20 runs   (    0.27
ms per token,  3686.64 tokens per second)
llama_print_timings: prompt eval time =   20222.57 ms /    91 tokens (  222.23
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =    5158.47 ms /    19 runs   (  271.50
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   25457.39 ms /   110 tokens
No. of rows:  41%| | 520/1258 [4:46:35<6:32:24, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.31 ms /    39 runs   (    0.26
ms per token,  3783.84 tokens per second)
llama_print_timings: prompt eval time =   22763.52 ms /   104 tokens (  218.88
ms per token,     4.57 tokens per second)
```

```
llama_print_timings:        eval time =   10278.55 ms /    38 runs   (  270.49
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   33193.66 ms /   142 tokens
No. of rows:  41%| | 521/1258 [4:47:08<6:36:41, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.32 ms /    43 runs   (    0.29
ms per token,  3489.13 tokens per second)
llama_print_timings: prompt eval time =   24355.27 ms /   112 tokens (  217.46
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =   11466.72 ms /    42 runs   (  273.02
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   35991.49 ms /   154 tokens
No. of rows:  41%| | 522/1258 [4:47:44<6:49:47, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.04 ms /    50 runs   (    0.26
ms per token,  3834.36 tokens per second)
llama_print_timings: prompt eval time =   26416.88 ms /   117 tokens (  225.79
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   13412.09 ms /    49 runs   (  273.72
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   40021.99 ms /   166 tokens
No. of rows:  42%| | 523/1258 [4:48:24<7:13:34, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.01 ms /    30 runs   (    0.27
ms per token,  3744.85 tokens per second)
llama_print_timings: prompt eval time =   20825.64 ms /    87 tokens (  239.38
ms per token,     4.18 tokens per second)
llama_print_timings:        eval time =    7937.32 ms /    29 runs   (  273.70
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   28878.24 ms /   116 tokens
No. of rows:  42%| | 524/1258 [4:48:53<6:49:05, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.17 ms /    50 runs   (    0.26
ms per token,  3796.22 tokens per second)
llama_print_timings: prompt eval time =   30396.79 ms /   142 tokens (  214.06
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13235.57 ms /    49 runs   (  270.11
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   43827.48 ms /   191 tokens
No. of rows:  42%| | 525/1258 [4:49:36<7:26:35, 36Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.89 ms /    26 runs   (    0.27
ms per token,  3771.94 tokens per second)
llama_print_timings: prompt eval time =   19152.84 ms /    88 tokens (  217.65
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    6674.69 ms /    25 runs   (  266.99
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   25927.89 ms /   113 tokens
No. of rows:  42%| | 526/1258 [4:50:02<6:47:08, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.99 ms /    42 runs   (    0.29
ms per token,  3502.04 tokens per second)
llama_print_timings: prompt eval time =   21463.42 ms /    93 tokens (  230.79
ms per token,     4.33 tokens per second)
llama_print_timings:        eval time =   11348.28 ms /    41 runs   (  276.79
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   32978.65 ms /   134 tokens
No. of rows:  42%| | 527/1258 [4:50:35<6:45:08, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.46 ms /    50 runs   (    0.27
ms per token,  3713.88 tokens per second)
llama_print_timings: prompt eval time =   20973.01 ms /    95 tokens (  220.77
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =   13238.06 ms /    49 runs   (  270.16
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   34406.85 ms /   144 tokens
No. of rows:  42%| | 528/1258 [4:51:10<6:48:51, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.59 ms /    32 runs   (    0.27
ms per token,  3724.39 tokens per second)
llama_print_timings: prompt eval time =   20763.68 ms /    96 tokens (  216.29
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    8394.45 ms /    31 runs   (  270.79
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   29286.54 ms /   127 tokens
No. of rows:  42%| | 529/1258 [4:51:39<6:32:35, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.79 ms /    32 runs   (    0.27
```

```
ms per token,  3638.84 tokens per second)
llama_print_timings: prompt eval time =   18954.23 ms /    86 tokens (  220.40
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    8791.87 ms /    31 runs   (  283.61
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   27874.28 ms /   117 tokens
No. of rows:  42%| | 530/1258 [4:52:07<6:15:53, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.03 ms /    29 runs   (    0.28
ms per token,  3611.46 tokens per second)
llama_print_timings: prompt eval time =   20410.83 ms /    93 tokens (  219.47
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    8039.28 ms /    28 runs   (  287.12
ms per token,     3.48 tokens per second)
llama_print_timings:       total time =   28567.79 ms /   121 tokens
No. of rows:  42%| | 531/1258 [4:52:36<6:06:39, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.83 ms /    46 runs   (    0.26
ms per token,  3888.75 tokens per second)
llama_print_timings: prompt eval time =   23836.63 ms /   110 tokens (  216.70
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   12259.31 ms /    45 runs   (  272.43
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   36273.78 ms /   155 tokens
No. of rows:  42%| | 532/1258 [4:53:12<6:28:00, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.84 ms /    50 runs   (    0.28
ms per token,  3612.72 tokens per second)
llama_print_timings: prompt eval time =   27718.79 ms /   131 tokens (  211.59
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   13258.42 ms /    49 runs   (  270.58
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   41176.49 ms /   180 tokens
No. of rows:  42%| | 533/1258 [4:53:53<7:00:31, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.51 ms /    27 runs   (    0.28
ms per token,  3595.21 tokens per second)
llama_print_timings: prompt eval time =   19397.68 ms /    90 tokens (  215.53
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    7049.48 ms /    26 runs   (  271.13
```

```
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   26552.36 ms /   116 tokens
No. of rows:  42%| | 534/1258 [4:54:20<6:30:06, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.87 ms /    50 runs   (    0.26
ms per token,  3885.00 tokens per second)
llama_print_timings: prompt eval time =   27598.68 ms /   120 tokens (  229.99
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =   13670.85 ms /    49 runs   (  279.00
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   41465.11 ms /   169 tokens
No. of rows:  43%| | 535/1258 [4:55:01<7:02:37, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.59 ms /    33 runs   (    0.29
ms per token,  3439.65 tokens per second)
llama_print_timings: prompt eval time =   23021.89 ms /   106 tokens (  217.19
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    8776.23 ms /    32 runs   (  274.26
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   31926.43 ms /   138 tokens
No. of rows:  43%| | 536/1258 [4:55:33<6:50:42, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.56 ms /    40 runs   (    0.26
ms per token,  3787.52 tokens per second)
llama_print_timings: prompt eval time =   22513.98 ms /    99 tokens (  227.41
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =   10629.33 ms /    39 runs   (  272.55
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   33299.15 ms /   138 tokens
No. of rows:  43%| | 537/1258 [4:56:06<6:47:08, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.31 ms /    50 runs   (    0.27
ms per token,  3755.73 tokens per second)
llama_print_timings: prompt eval time =   21744.11 ms /   102 tokens (  213.18
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   13724.50 ms /    49 runs   (  280.09
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   35668.10 ms /   151 tokens
No. of rows:  43%| | 538/1258 [4:56:42<6:53:03, 34Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.10 ms /    27 runs   (    0.26
ms per token,  3803.35 tokens per second)
llama_print_timings: prompt eval time =   17195.18 ms /    78 tokens (  220.45
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    7100.69 ms /    26 runs   (  273.10
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   24399.09 ms /   104 tokens
No. of rows:  43%| | 539/1258 [4:57:06<6:16:28, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.34 ms /    28 runs   (    0.26
ms per token,  3815.23 tokens per second)
llama_print_timings: prompt eval time =   18316.43 ms /    84 tokens (  218.05
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    9041.04 ms /    27 runs   (  334.85
ms per token,     2.99 tokens per second)
llama_print_timings:       total time =   27463.68 ms /   111 tokens
No. of rows:  43%| | 540/1258 [4:57:34<6:01:47, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.82 ms /    50 runs   (    0.28
ms per token,  3617.95 tokens per second)
llama_print_timings: prompt eval time =   24470.32 ms /   114 tokens (  214.65
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13386.96 ms /    49 runs   (  273.20
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   38055.40 ms /   163 tokens
No. of rows:  43%| | 541/1258 [4:58:12<6:29:20, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.27 ms /    50 runs   (    0.27
ms per token,  3769.03 tokens per second)
llama_print_timings: prompt eval time =   30119.62 ms /   134 tokens (  224.77
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   13625.06 ms /    49 runs   (  278.06
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   43941.73 ms /   183 tokens
No. of rows:  43%| | 542/1258 [4:58:56<7:09:30, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       4.94 ms /    19 runs   (    0.26
ms per token,  3846.93 tokens per second)
```

```
llama_print_timings: prompt eval time =   23300.04 ms /   108 tokens (  215.74
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    4950.14 ms /    18 runs   (  275.01
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   28323.96 ms /   126 tokens
No. of rows:  43%| | 543/1258 [4:59:24<6:41:31, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.46 ms /    50 runs   (    0.27
ms per token,  3715.81 tokens per second)
llama_print_timings: prompt eval time =   24866.00 ms /   110 tokens (  226.05
ms per token,     4.42 tokens per second)
llama_print_timings:        eval time =   15110.80 ms /    49 runs   (  308.38
ms per token,     3.24 tokens per second)
llama_print_timings:       total time =   40172.66 ms /   159 tokens
No. of rows:  43%| | 544/1258 [5:00:04<7:04:07, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.15 ms /    26 runs   (    0.27
ms per token,  3637.38 tokens per second)
llama_print_timings: prompt eval time =   19173.89 ms /    87 tokens (  220.39
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    6784.05 ms /    25 runs   (  271.36
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   26061.54 ms /   112 tokens
No. of rows:  43%| | 545/1258 [5:00:30<6:29:24, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.28 ms /    35 runs   (    0.27
ms per token,  3770.33 tokens per second)
llama_print_timings: prompt eval time =   21581.22 ms /    93 tokens (  232.06
ms per token,     4.31 tokens per second)
llama_print_timings:        eval time =    9304.52 ms /    34 runs   (  273.66
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   31022.07 ms /   127 tokens
No. of rows:  43%| | 546/1258 [5:01:02<6:22:39, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.66 ms /    44 runs   (    0.26
ms per token,  3773.91 tokens per second)
llama_print_timings: prompt eval time =   20267.74 ms /    94 tokens (  215.61
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   11688.32 ms /    43 runs   (  271.82
ms per token,     3.68 tokens per second)
```

```
llama_print_timings:       total time =   32126.07 ms /   137 tokens
No. of rows:  43%| | 547/1258 [5:01:34<6:21:43, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.07 ms /    31 runs   (    0.26
ms per token,  3842.82 tokens per second)
llama_print_timings: prompt eval time =   23040.75 ms /   108 tokens (  213.34
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    8174.13 ms /    30 runs   (  272.47
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   31334.93 ms /   138 tokens
No. of rows:  44%| | 548/1258 [5:02:05<6:18:07, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.29 ms /    32 runs   (    0.29
ms per token,  3445.68 tokens per second)
llama_print_timings: prompt eval time =   22227.87 ms /    97 tokens (  229.15
ms per token,     4.36 tokens per second)
llama_print_timings:        eval time =   10930.62 ms /    31 runs   (  352.60
ms per token,     2.84 tokens per second)
llama_print_timings:       total time =   33296.29 ms /   128 tokens
No. of rows:  44%| | 549/1258 [5:02:38<6:22:21, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.04 ms /    27 runs   (    0.26
ms per token,  3836.32 tokens per second)
llama_print_timings: prompt eval time =   17712.64 ms /    79 tokens (  224.21
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =    7023.81 ms /    26 runs   (  270.15
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   24839.99 ms /   105 tokens
No. of rows:  44%| | 550/1258 [5:03:03<5:55:14, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.46 ms /    32 runs   (    0.26
ms per token,  3780.27 tokens per second)
llama_print_timings: prompt eval time =   20872.06 ms /    96 tokens (  217.42
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    8468.78 ms /    31 runs   (  273.19
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   29467.05 ms /   127 tokens
No. of rows:  44%| | 551/1258 [5:03:33<5:52:30, 29Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.02 ms /    26 runs   (    0.27
ms per token,  3702.65 tokens per second)
llama_print_timings: prompt eval time =   17182.35 ms /    78 tokens (  220.29
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    6687.34 ms /    25 runs   (  267.49
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   23970.34 ms /   103 tokens
No. of rows:  44%| | 552/1258 [5:03:57<5:31:01, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.00 ms /    26 runs   (    0.27
ms per token,  3716.94 tokens per second)
llama_print_timings: prompt eval time =   20263.01 ms /    94 tokens (  215.56
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    8496.62 ms /    25 runs   (  339.86
ms per token,     2.94 tokens per second)
llama_print_timings:       total time =   28860.32 ms /   119 tokens
No. of rows:  44%| | 553/1258 [5:04:25<5:33:10, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.30 ms /    50 runs   (    0.27
ms per token,  3758.27 tokens per second)
llama_print_timings: prompt eval time =   23457.30 ms /   106 tokens (  221.30
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =   13505.16 ms /    49 runs   (  275.62
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   37159.69 ms /   155 tokens
No. of rows:  44%| | 554/1258 [5:05:03<6:03:42, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.10 ms /    50 runs   (    0.26
ms per token,  3817.09 tokens per second)
llama_print_timings: prompt eval time =   23269.75 ms /   107 tokens (  217.47
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =   13429.37 ms /    49 runs   (  274.07
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   36897.80 ms /   156 tokens
No. of rows:  44%| | 555/1258 [5:05:40<6:23:55, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.06 ms /    34 runs   (    0.27
ms per token,  3752.35 tokens per second)
llama_print_timings: prompt eval time =   21127.95 ms /    98 tokens (  215.59
```

```
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    9002.03 ms /    33 runs   (  272.79
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   30261.04 ms /   131 tokens
No. of rows:  44%| | 556/1258 [5:06:10<6:14:36, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.02 ms /    37 runs   (    0.27
ms per token,  3693.35 tokens per second)
llama_print_timings: prompt eval time =   22242.81 ms /   103 tokens (  215.95
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   10000.18 ms /    36 runs   (  277.78
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   32393.93 ms /   139 tokens
No. of rows:  44%| | 557/1258 [5:06:42<6:15:25, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.37 ms /    41 runs   (    0.28
ms per token,  3605.66 tokens per second)
llama_print_timings: prompt eval time =   26810.03 ms /   124 tokens (  216.21
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   11062.46 ms /    40 runs   (  276.56
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   38034.45 ms /   164 tokens
No. of rows:  44%| | 558/1258 [5:07:20<6:35:35, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.69 ms /    50 runs   (    0.27
ms per token,  3653.64 tokens per second)
llama_print_timings: prompt eval time =   22537.42 ms /   105 tokens (  214.64
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13501.17 ms /    49 runs   (  275.53
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   36233.73 ms /   154 tokens
No. of rows:  44%| | 559/1258 [5:07:56<6:43:08, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.16 ms /    50 runs   (    0.26
ms per token,  3799.39 tokens per second)
llama_print_timings: prompt eval time =   25446.10 ms /   119 tokens (  213.83
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13450.84 ms /    49 runs   (  274.51
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   39092.99 ms /   168 tokens
```

```
No. of rows:  45%| | 560/1258 [5:08:36<6:58:17, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.32 ms /    34 runs   (    0.27
ms per token,  3648.46 tokens per second)
llama_print_timings: prompt eval time =   17698.94 ms /    80 tokens (  221.24
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    8967.59 ms /    33 runs   (  271.75
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   26799.14 ms /   113 tokens
No. of rows:  45%| | 561/1258 [5:09:02<6:25:47, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.91 ms /    34 runs   (    0.26
ms per token,  3817.22 tokens per second)
llama_print_timings: prompt eval time =   24653.21 ms /   114 tokens (  216.26
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    8959.44 ms /    33 runs   (  271.50
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   33744.77 ms /   147 tokens
No. of rows:  45%| | 562/1258 [5:09:36<6:27:07, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.46 ms /    29 runs   (    0.26
ms per token,  3885.32 tokens per second)
llama_print_timings: prompt eval time =   22034.50 ms /    96 tokens (  229.53
ms per token,     4.36 tokens per second)
llama_print_timings:        eval time =    7837.10 ms /    28 runs   (  279.90
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   29983.68 ms /   124 tokens
No. of rows:  45%| | 563/1258 [5:10:06<6:14:48, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.11 ms /    50 runs   (    0.28
ms per token,  3543.59 tokens per second)
llama_print_timings: prompt eval time =   29133.97 ms /   126 tokens (  231.22
ms per token,     4.32 tokens per second)
llama_print_timings:        eval time =   14031.59 ms /    49 runs   (  286.36
ms per token,     3.49 tokens per second)
llama_print_timings:       total time =   43370.49 ms /   175 tokens
No. of rows:  45%| | 564/1258 [5:10:49<6:52:31, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      13.67 ms /    50 runs   (    0.27
ms per token,  3657.91 tokens per second)
llama_print_timings: prompt eval time =   25823.17 ms /   118 tokens (  218.84
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =   13390.26 ms /    49 runs   (  273.27
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   39409.85 ms /   167 tokens
No. of rows:  45%| | 565/1258 [5:11:29<7:04:52, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.25 ms /    41 runs   (    0.27
ms per token,  3645.42 tokens per second)
llama_print_timings: prompt eval time =   23009.10 ms /   107 tokens (  215.04
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   10863.37 ms /    40 runs   (  271.58
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   34031.05 ms /   147 tokens
No. of rows:  45%| | 566/1258 [5:12:03<6:54:47, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.59 ms /    32 runs   (    0.27
ms per token,  3723.53 tokens per second)
llama_print_timings: prompt eval time =   18535.27 ms /    79 tokens (  234.62
ms per token,     4.26 tokens per second)
llama_print_timings:        eval time =    8444.90 ms /    31 runs   (  272.42
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   27111.37 ms /   110 tokens
No. of rows:  45%| | 567/1258 [5:12:30<6:23:38, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.14 ms /    27 runs   (    0.26
ms per token,  3783.10 tokens per second)
llama_print_timings: prompt eval time =   20560.89 ms /    87 tokens (  236.33
ms per token,     4.23 tokens per second)
llama_print_timings:        eval time =    7069.12 ms /    26 runs   (  271.89
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   27735.02 ms /   113 tokens
No. of rows:  45%| | 568/1258 [5:12:58<6:03:49, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.81 ms /    50 runs   (    0.28
ms per token,  3620.83 tokens per second)
llama_print_timings: prompt eval time =   22603.19 ms /   104 tokens (  217.34
ms per token,     4.60 tokens per second)
```

```
llama_print_timings:        eval time =   13277.09 ms /    49 runs   (  270.96
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   36075.15 ms /   153 tokens
No. of rows:  45%| | 569/1258 [5:13:34<6:18:39, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.23 ms /    43 runs   (    0.26
ms per token,  3829.71 tokens per second)
llama_print_timings: prompt eval time =   20386.22 ms /    95 tokens (  214.59
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   11359.74 ms /    42 runs   (  270.47
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   31914.01 ms /   137 tokens
No. of rows:  45%| | 570/1258 [5:14:06<6:14:29, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.30 ms /    43 runs   (    0.29
ms per token,  3496.79 tokens per second)
llama_print_timings: prompt eval time =   28215.44 ms /   125 tokens (  225.72
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   12397.37 ms /    42 runs   (  295.18
ms per token,     3.39 tokens per second)
llama_print_timings:       total time =   40800.32 ms /   167 tokens
No. of rows:  45%| | 571/1258 [5:14:47<6:41:57, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.18 ms /    35 runs   (    0.26
ms per token,  3812.64 tokens per second)
llama_print_timings: prompt eval time =   20660.26 ms /    91 tokens (  227.04
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =    9317.20 ms /    34 runs   (  274.04
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   30126.97 ms /   125 tokens
No. of rows:  45%| | 572/1258 [5:15:17<6:24:19, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.52 ms /    44 runs   (    0.26
ms per token,  3819.78 tokens per second)
llama_print_timings: prompt eval time =   24003.80 ms /   110 tokens (  218.22
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =   11818.95 ms /    43 runs   (  274.86
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   35993.78 ms /   153 tokens
No. of rows:  46%| | 573/1258 [5:15:53<6:31:53, 34Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.56 ms /    25 runs   (    0.26
ms per token,  3809.81 tokens per second)
llama_print_timings: prompt eval time =   19293.30 ms /    88 tokens (  219.24
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    6534.86 ms /    24 runs   (  272.29
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   25923.57 ms /   112 tokens
No. of rows:  46%| | 574/1258 [5:16:19<6:02:38, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.42 ms /    50 runs   (    0.27
ms per token,  3726.34 tokens per second)
llama_print_timings: prompt eval time =   29499.76 ms /   131 tokens (  225.19
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =   13810.77 ms /    49 runs   (  281.85
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   43514.35 ms /   180 tokens
No. of rows:  46%| | 575/1258 [5:17:02<6:42:05, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.39 ms /    27 runs   (    0.27
ms per token,  3654.08 tokens per second)
llama_print_timings: prompt eval time =   19071.12 ms /    88 tokens (  216.72
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    8683.01 ms /    26 runs   (  333.96
ms per token,     2.99 tokens per second)
llama_print_timings:       total time =   27861.57 ms /   114 tokens
No. of rows:  46%| | 576/1258 [5:17:30<6:16:05, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.93 ms /    27 runs   (    0.29
ms per token,  3406.08 tokens per second)
llama_print_timings: prompt eval time =   16770.18 ms /    78 tokens (  215.00
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    7308.04 ms /    26 runs   (  281.08
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   24186.72 ms /   104 tokens
No. of rows:  46%| | 577/1258 [5:17:54<5:45:15, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.46 ms /    34 runs   (    0.28
```

```
ms per token,  3592.18 tokens per second)
llama_print_timings: prompt eval time =   21737.32 ms /   101 tokens (  215.22
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9052.87 ms /    33 runs   (  274.33
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   30925.84 ms /   134 tokens
No. of rows:  46%| | 578/1258 [5:18:25<5:46:30, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.08 ms /    30 runs   (    0.30
ms per token,  3305.42 tokens per second)
llama_print_timings: prompt eval time =   19626.90 ms /    89 tokens (  220.53
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =   10247.49 ms /    29 runs   (  353.36
ms per token,     2.83 tokens per second)
llama_print_timings:       total time =   30006.17 ms /   118 tokens
No. of rows:  46%| | 579/1258 [5:18:55<5:44:03, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.79 ms /    26 runs   (    0.26
ms per token,  3830.29 tokens per second)
llama_print_timings: prompt eval time =   20438.50 ms /    95 tokens (  215.14
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    6797.89 ms /    25 runs   (  271.92
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   27337.75 ms /   120 tokens
No. of rows:  46%| | 580/1258 [5:19:23<5:33:12, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.71 ms /    36 runs   (    0.27
ms per token,  3709.05 tokens per second)
llama_print_timings: prompt eval time =   23158.70 ms /   106 tokens (  218.48
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    9657.21 ms /    35 runs   (  275.92
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   32955.90 ms /   141 tokens
No. of rows:  46%| | 581/1258 [5:19:56<5:44:27, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.12 ms /    49 runs   (    0.27
ms per token,  3733.90 tokens per second)
llama_print_timings: prompt eval time =   25771.33 ms /   113 tokens (  228.06
ms per token,     4.38 tokens per second)
llama_print_timings:        eval time =   13050.50 ms /    48 runs   (  271.89
```

```
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   39010.35 ms /   161 tokens
No. of rows:  46%| | 582/1258 [5:20:35<6:12:40, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.74 ms /    29 runs   (    0.27
ms per token,  3746.77 tokens per second)
llama_print_timings: prompt eval time =   19946.80 ms /    83 tokens (  240.32
ms per token,     4.16 tokens per second)
llama_print_timings:        eval time =    7710.64 ms /    28 runs   (  275.38
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   27773.65 ms /   111 tokens
No. of rows:  46%| | 583/1258 [5:21:02<5:54:14, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.93 ms /    42 runs   (    0.26
ms per token,  3842.28 tokens per second)
llama_print_timings: prompt eval time =   22043.24 ms /   103 tokens (  214.01
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   11213.77 ms /    41 runs   (  273.51
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   33424.53 ms /   144 tokens
No. of rows:  46%| | 584/1258 [5:21:36<6:00:16, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.84 ms /    45 runs   (    0.26
ms per token,  3801.96 tokens per second)
llama_print_timings: prompt eval time =   23806.71 ms /   110 tokens (  216.42
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   12228.49 ms /    44 runs   (  277.92
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   36212.31 ms /   154 tokens
No. of rows:  47%| | 585/1258 [5:22:12<6:13:41, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.03 ms /    50 runs   (    0.28
ms per token,  3564.55 tokens per second)
llama_print_timings: prompt eval time =   25737.83 ms /   120 tokens (  214.48
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13657.21 ms /    49 runs   (  278.72
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   39598.61 ms /   169 tokens
No. of rows:  47%| | 586/1258 [5:22:52<6:34:16, 35Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.10 ms /    39 runs   (    0.28
ms per token,  3513.51 tokens per second)
llama_print_timings: prompt eval time =   26088.59 ms /   119 tokens (  219.23
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   10470.21 ms /    38 runs   (  275.53
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   36712.75 ms /   157 tokens
No. of rows:  47%| | 587/1258 [5:23:28<6:38:46, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.04 ms /    33 runs   (    0.27
ms per token,  3650.04 tokens per second)
llama_print_timings: prompt eval time =   19487.54 ms /    89 tokens (  218.96
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    8687.67 ms /    32 runs   (  271.49
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   28306.55 ms /   121 tokens
No. of rows:  47%| | 588/1258 [5:23:57<6:13:34, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.89 ms /    50 runs   (    0.26
ms per token,  3879.28 tokens per second)
llama_print_timings: prompt eval time =   26295.69 ms /   123 tokens (  213.79
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13378.36 ms /    49 runs   (  273.03
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   39869.35 ms /   172 tokens
No. of rows:  47%| | 589/1258 [5:24:37<6:34:30, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.37 ms /    50 runs   (    0.27
ms per token,  3740.56 tokens per second)
llama_print_timings: prompt eval time =   26780.15 ms /   117 tokens (  228.89
ms per token,     4.37 tokens per second)
llama_print_timings:        eval time =   13815.77 ms /    49 runs   (  281.95
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   40796.41 ms /   166 tokens
No. of rows:  47%| | 590/1258 [5:25:17<6:51:58, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.30 ms /    50 runs   (    0.27
ms per token,  3760.25 tokens per second)
```

```
llama_print_timings: prompt eval time =   29677.25 ms /   139 tokens (  213.51
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   15168.90 ms /    49 runs   (  309.57
ms per token,     3.23 tokens per second)
llama_print_timings:       total time =   45043.50 ms /   188 tokens
No. of rows:  47%| | 591/1258 [5:26:02<7:18:14, 39Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.75 ms /    26 runs   (    0.26
ms per token,  3852.42 tokens per second)
llama_print_timings: prompt eval time =   19663.87 ms /    89 tokens (  220.94
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    6750.59 ms /    25 runs   (  270.02
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   26515.42 ms /   114 tokens
No. of rows:  47%| | 592/1258 [5:26:29<6:34:37, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.66 ms /    32 runs   (    0.27
ms per token,  3694.30 tokens per second)
llama_print_timings: prompt eval time =   19395.98 ms /    89 tokens (  217.93
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    8800.05 ms /    31 runs   (  283.87
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   28323.93 ms /   120 tokens
No. of rows:  47%| | 593/1258 [5:26:57<6:10:01, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.55 ms /    24 runs   (    0.27
ms per token,  3663.56 tokens per second)
llama_print_timings: prompt eval time =   22240.34 ms /    96 tokens (  231.67
ms per token,     4.32 tokens per second)
llama_print_timings:        eval time =    6369.72 ms /    23 runs   (  276.94
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   28703.00 ms /   119 tokens
No. of rows:  47%| | 594/1258 [5:27:26<5:53:55, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.58 ms /    39 runs   (    0.27
ms per token,  3686.55 tokens per second)
llama_print_timings: prompt eval time =   21319.29 ms /    99 tokens (  215.35
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   10274.76 ms /    38 runs   (  270.39
ms per token,     3.70 tokens per second)
```

```
llama_print_timings:       total time =   31748.21 ms /   137 tokens
No. of rows:  47%| | 595/1258 [5:27:58<5:52:37, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.62 ms /    36 runs   (    0.27
ms per token,  3743.37 tokens per second)
llama_print_timings: prompt eval time =   20143.62 ms /    93 tokens (  216.60
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    9463.80 ms /    35 runs   (  270.39
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   29749.26 ms /   128 tokens
No. of rows:  47%| | 596/1258 [5:28:27<5:45:00, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.18 ms /    38 runs   (    0.27
ms per token,  3733.18 tokens per second)
llama_print_timings: prompt eval time =   20859.70 ms /    90 tokens (  231.77
ms per token,     4.31 tokens per second)
llama_print_timings:        eval time =   10280.14 ms /    37 runs   (  277.84
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   31294.04 ms /   127 tokens
No. of rows:  47%| | 597/1258 [5:28:59<5:44:35, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.38 ms /    36 runs   (    0.26
ms per token,  3838.77 tokens per second)
llama_print_timings: prompt eval time =   22564.44 ms /   104 tokens (  216.97
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    9695.08 ms /    35 runs   (  277.00
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   32405.35 ms /   139 tokens
No. of rows:  48%| | 598/1258 [5:29:31<5:47:49, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       4.02 ms /    16 runs   (    0.25
ms per token,  3981.09 tokens per second)
llama_print_timings: prompt eval time =   23481.98 ms /   110 tokens (  213.47
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    4176.30 ms /    15 runs   (  278.42
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   27719.21 ms /   125 tokens
No. of rows:  48%| | 599/1258 [5:29:59<5:34:27, 30Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.39 ms /    39 runs   (    0.27
ms per token,  3754.69 tokens per second)
llama_print_timings: prompt eval time =   23383.66 ms /   108 tokens (  216.52
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   10354.42 ms /    38 runs   (  272.48
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   33889.79 ms /   146 tokens
No. of rows:  48%| | 600/1258 [5:30:33<5:45:17, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.81 ms /    30 runs   (    0.26
ms per token,  3840.25 tokens per second)
llama_print_timings: prompt eval time =   21183.08 ms /    97 tokens (  218.38
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    8011.06 ms /    29 runs   (  276.24
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   29309.96 ms /   126 tokens
No. of rows:  48%| | 601/1258 [5:31:02<5:37:38, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.48 ms /    50 runs   (    0.27
ms per token,  3708.65 tokens per second)
llama_print_timings: prompt eval time =   34403.59 ms /   155 tokens (  221.96
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =   13454.43 ms /    49 runs   (  274.58
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   48058.51 ms /   204 tokens
No. of rows:  48%| | 602/1258 [5:31:50<6:33:36, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.84 ms /    37 runs   (    0.27
ms per token,  3761.31 tokens per second)
llama_print_timings: prompt eval time =   24224.88 ms /   110 tokens (  220.23
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    9719.17 ms /    36 runs   (  269.98
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   34087.65 ms /   146 tokens
No. of rows:  48%| | 603/1258 [5:32:24<6:26:48, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.68 ms /    43 runs   (    0.27
ms per token,  3682.45 tokens per second)
llama_print_timings: prompt eval time =   27929.48 ms /   123 tokens (  227.07
```

```
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =   11644.63 ms /    42 runs   (  277.25
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   39744.20 ms /   165 tokens
No. of rows:  48%| | 604/1258 [5:33:04<6:40:19, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.52 ms /    25 runs   (    0.26
ms per token,  3837.30 tokens per second)
llama_print_timings: prompt eval time =   20753.02 ms /    95 tokens (  218.45
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    6542.94 ms /    24 runs   (  272.62
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   27394.40 ms /   119 tokens
No. of rows:  48%| | 605/1258 [5:33:31<6:09:17, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.82 ms /    48 runs   (    0.27
ms per token,  3743.86 tokens per second)
llama_print_timings: prompt eval time =   27527.50 ms /   128 tokens (  215.06
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   14475.03 ms /    47 runs   (  307.98
ms per token,     3.25 tokens per second)
llama_print_timings:       total time =   42188.01 ms /   175 tokens
No. of rows:  48%| | 606/1258 [5:34:14<6:35:39, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.37 ms /    50 runs   (    0.27
ms per token,  3739.72 tokens per second)
llama_print_timings: prompt eval time =   24237.81 ms /   112 tokens (  216.41
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   13358.13 ms /    49 runs   (  272.61
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   37795.57 ms /   161 tokens
No. of rows:  48%| | 607/1258 [5:34:51<6:39:35, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.06 ms /    39 runs   (    0.28
ms per token,  3526.54 tokens per second)
llama_print_timings: prompt eval time =   25340.43 ms /   115 tokens (  220.35
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =   10930.95 ms /    38 runs   (  287.66
ms per token,     3.48 tokens per second)
llama_print_timings:       total time =   36429.72 ms /   153 tokens
```

```
No. of rows:  48%| | 608/1258 [5:35:28<6:37:41, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.05 ms /    44 runs   (    0.27
ms per token,  3651.45 tokens per second)
llama_print_timings: prompt eval time =   22805.87 ms /   105 tokens (  217.20
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =   11769.51 ms /    43 runs   (  273.71
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   34755.43 ms /   148 tokens
No. of rows:  48%| | 609/1258 [5:36:03<6:30:45, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.96 ms /    37 runs   (    0.27
ms per token,  3715.98 tokens per second)
llama_print_timings: prompt eval time =   22169.97 ms /   103 tokens (  215.24
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9886.26 ms /    36 runs   (  274.62
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   32199.96 ms /   139 tokens
No. of rows:  48%| | 610/1258 [5:36:35<6:17:24, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.91 ms /    39 runs   (    0.28
ms per token,  3575.03 tokens per second)
llama_print_timings: prompt eval time =   21780.31 ms /    99 tokens (  220.00
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   10457.27 ms /    38 runs   (  275.19
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   32393.74 ms /   137 tokens
No. of rows:  49%| | 611/1258 [5:37:07<6:08:39, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.16 ms /    50 runs   (    0.28
ms per token,  3532.07 tokens per second)
llama_print_timings: prompt eval time =   29943.20 ms /   138 tokens (  216.98
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   13423.37 ms /    49 runs   (  273.95
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   43569.75 ms /   187 tokens
No. of rows:  49%| | 612/1258 [5:37:51<6:38:24, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =       5.95 ms /    23 runs   (    0.26
ms per token,  3863.60 tokens per second)
llama_print_timings: prompt eval time =   18060.32 ms /    81 tokens (  222.97
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =    5974.57 ms /    22 runs   (  271.57
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   24123.71 ms /   103 tokens
No. of rows:  49%| | 613/1258 [5:38:15<5:56:16, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.58 ms /    29 runs   (    0.36
ms per token,  2741.02 tokens per second)
llama_print_timings: prompt eval time =   20353.14 ms /    86 tokens (  236.66
ms per token,     4.23 tokens per second)
llama_print_timings:        eval time =    7704.68 ms /    28 runs   (  275.17
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   28173.57 ms /   114 tokens
No. of rows:  49%| | 614/1258 [5:38:43<5:39:45, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.79 ms /    50 runs   (    0.28
ms per token,  3626.87 tokens per second)
llama_print_timings: prompt eval time =   23511.89 ms /   101 tokens (  232.79
ms per token,     4.30 tokens per second)
llama_print_timings:        eval time =   13711.27 ms /    49 runs   (  279.82
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   37420.38 ms /   150 tokens
No. of rows:  49%| | 615/1258 [5:39:21<5:57:47, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.11 ms /    43 runs   (    0.26
ms per token,  3872.13 tokens per second)
llama_print_timings: prompt eval time =   17257.67 ms /    77 tokens (  224.13
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =   11483.07 ms /    42 runs   (  273.41
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   28909.13 ms /   119 tokens
No. of rows:  49%| | 616/1258 [5:39:49<5:42:51, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.75 ms /    50 runs   (    0.26
ms per token,  3920.95 tokens per second)
llama_print_timings: prompt eval time =   32018.13 ms /   142 tokens (  225.48
ms per token,     4.43 tokens per second)
```

```
llama_print_timings:        eval time =   13416.77 ms /    49 runs   (  273.81
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   45633.87 ms /   191 tokens
No. of rows:  49%| | 617/1258 [5:40:35<6:25:55, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.70 ms /    32 runs   (    0.27
ms per token,  3678.16 tokens per second)
llama_print_timings: prompt eval time =   19373.16 ms /    81 tokens (  239.17
ms per token,     4.18 tokens per second)
llama_print_timings:        eval time =    8531.33 ms /    31 runs   (  275.20
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   28031.07 ms /   112 tokens
No. of rows:  49%| | 618/1258 [5:41:03<5:59:25, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.56 ms /    50 runs   (    0.27
ms per token,  3687.59 tokens per second)
llama_print_timings: prompt eval time =   29896.14 ms /   133 tokens (  224.78
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   13241.08 ms /    49 runs   (  270.23
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   43331.08 ms /   182 tokens
No. of rows:  49%| | 619/1258 [5:41:46<6:29:40, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.01 ms /    23 runs   (    0.26
ms per token,  3826.96 tokens per second)
llama_print_timings: prompt eval time =   15286.17 ms /    68 tokens (  224.80
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =    5856.39 ms /    22 runs   (  266.20
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   21232.91 ms /    90 tokens
No. of rows:  49%| | 620/1258 [5:42:08<5:40:04, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.05 ms /    50 runs   (    0.28
ms per token,  3558.21 tokens per second)
llama_print_timings: prompt eval time =   30262.11 ms /   136 tokens (  222.52
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =   13593.61 ms /    49 runs   (  277.42
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   44060.48 ms /   185 tokens
No. of rows:  49%| | 621/1258 [5:42:52<6:18:04, 35Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.05 ms /    38 runs   (    0.26
ms per token,  3779.21 tokens per second)
llama_print_timings: prompt eval time =   26041.33 ms /   112 tokens (  232.51
ms per token,     4.30 tokens per second)
llama_print_timings:        eval time =   10023.49 ms /    37 runs   (  270.91
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   36213.19 ms /   149 tokens
No. of rows:  49%| | 622/1258 [5:43:28<6:19:24, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.12 ms /    30 runs   (    0.27
ms per token,  3693.67 tokens per second)
llama_print_timings: prompt eval time =   18417.80 ms /    85 tokens (  216.68
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7862.60 ms /    29 runs   (  271.12
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   26397.93 ms /   114 tokens
No. of rows:  50%| | 623/1258 [5:43:54<5:49:00, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.27 ms /    31 runs   (    0.27
ms per token,  3750.30 tokens per second)
llama_print_timings: prompt eval time =   18641.94 ms /    83 tokens (  224.60
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =    8227.80 ms /    30 runs   (  274.26
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   26993.11 ms /   113 tokens
No. of rows:  50%| | 624/1258 [5:44:21<5:29:28, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.28 ms /    41 runs   (    0.28
ms per token,  3636.04 tokens per second)
llama_print_timings: prompt eval time =   28625.16 ms /   128 tokens (  223.63
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   11183.89 ms /    40 runs   (  279.60
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   39970.33 ms /   168 tokens
No. of rows:  50%| | 625/1258 [5:45:01<5:56:50, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.94 ms /    48 runs   (    0.31
```

```
ms per token,  3213.28 tokens per second)
llama_print_timings: prompt eval time =   24755.93 ms /   106 tokens (  233.55
ms per token,     4.28 tokens per second)
llama_print_timings:        eval time =   14961.56 ms /    47 runs   (  318.33
ms per token,     3.14 tokens per second)
llama_print_timings:       total time =   39937.28 ms /   153 tokens
No. of rows:  50%| | 626/1258 [5:45:41<6:15:34, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      15.85 ms /    50 runs   (    0.32
ms per token,  3154.18 tokens per second)
llama_print_timings: prompt eval time =   23982.72 ms /   104 tokens (  230.60
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =   18652.85 ms /    49 runs   (  380.67
ms per token,     2.63 tokens per second)
llama_print_timings:       total time =   42879.93 ms /   153 tokens
No. of rows:  50%| | 627/1258 [5:46:24<6:37:50, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.41 ms /    42 runs   (    0.34
ms per token,  2914.84 tokens per second)
llama_print_timings: prompt eval time =   29325.29 ms /   106 tokens (  276.65
ms per token,     3.61 tokens per second)
llama_print_timings:        eval time =   14281.72 ms /    41 runs   (  348.33
ms per token,     2.87 tokens per second)
llama_print_timings:       total time =   43814.89 ms /   147 tokens
No. of rows:  50%| | 628/1258 [5:47:08<6:56:05, 39Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      17.28 ms /    50 runs   (    0.35
ms per token,  2893.52 tokens per second)
llama_print_timings: prompt eval time =   26787.39 ms /   102 tokens (  262.62
ms per token,     3.81 tokens per second)
llama_print_timings:        eval time =   17531.50 ms /    49 runs   (  357.79
ms per token,     2.79 tokens per second)
llama_print_timings:       total time =   44581.39 ms /   151 tokens
No. of rows:  50%| | 629/1258 [5:47:53<7:11:01, 41Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.42 ms /    42 runs   (    0.27
ms per token,  3677.76 tokens per second)
llama_print_timings: prompt eval time =   24499.71 ms /    98 tokens (  250.00
ms per token,     4.00 tokens per second)
llama_print_timings:        eval time =   11667.54 ms /    41 runs   (  284.57
```

```
ms per token,     3.51 tokens per second)
llama_print_timings:       total time =   36336.71 ms /   139 tokens
No. of rows:  50%| | 630/1258 [5:48:29<6:55:21, 39Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.48 ms /    50 runs   (    0.27
ms per token,  3710.02 tokens per second)
llama_print_timings: prompt eval time =   23279.96 ms /   104 tokens (  223.85
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   14322.14 ms /    49 runs   (  292.29
ms per token,     3.42 tokens per second)
llama_print_timings:       total time =   37805.59 ms /   153 tokens
No. of rows:  50%| | 631/1258 [5:49:07<6:48:47, 39Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.33 ms /    50 runs   (    0.27
ms per token,  3751.22 tokens per second)
llama_print_timings: prompt eval time =   24369.66 ms /   108 tokens (  225.65
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   13695.81 ms /    49 runs   (  279.51
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   38263.10 ms /   157 tokens
No. of rows:  50%| | 632/1258 [5:49:45<6:45:30, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.20 ms /    29 runs   (    0.28
ms per token,  3537.45 tokens per second)
llama_print_timings: prompt eval time =   24090.53 ms /   109 tokens (  221.01
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    8042.81 ms /    28 runs   (  287.24
ms per token,     3.48 tokens per second)
llama_print_timings:       total time =   32259.55 ms /   137 tokens
No. of rows:  50%| | 633/1258 [5:50:17<6:24:15, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.64 ms /    26 runs   (    0.26
ms per token,  3916.25 tokens per second)
llama_print_timings: prompt eval time =   18548.87 ms /    83 tokens (  223.48
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =    6799.85 ms /    25 runs   (  271.99
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   25447.88 ms /   108 tokens
No. of rows:  50%| | 634/1258 [5:50:43<5:47:56, 33Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.33 ms /    50 runs   (    0.27
ms per token,  3752.35 tokens per second)
llama_print_timings: prompt eval time =   28312.17 ms /   121 tokens (  233.98
ms per token,     4.27 tokens per second)
llama_print_timings:        eval time =   13646.01 ms /    49 runs   (  278.49
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   42153.96 ms /   170 tokens
No. of rows:  50%| | 635/1258 [5:51:25<6:14:31, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.85 ms /    50 runs   (    0.30
ms per token,  3366.55 tokens per second)
llama_print_timings: prompt eval time =   24784.25 ms /   109 tokens (  227.38
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =   14686.71 ms /    49 runs   (  299.73
ms per token,     3.34 tokens per second)
llama_print_timings:       total time =   39684.23 ms /   158 tokens
No. of rows:  51%| | 636/1258 [5:52:05<6:25:11, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.46 ms /    30 runs   (    0.35
ms per token,  2867.79 tokens per second)
llama_print_timings: prompt eval time =   27434.58 ms /    90 tokens (  304.83
ms per token,     3.28 tokens per second)
llama_print_timings:        eval time =    9265.95 ms /    29 runs   (  319.52
ms per token,     3.13 tokens per second)
llama_print_timings:       total time =   36845.59 ms /   119 tokens
No. of rows:  51%| | 637/1258 [5:52:41<6:23:38, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.65 ms /    50 runs   (    0.27
ms per token,  3663.00 tokens per second)
llama_print_timings: prompt eval time =   25761.77 ms /   109 tokens (  236.35
ms per token,     4.23 tokens per second)
llama_print_timings:        eval time =   13303.00 ms /    49 runs   (  271.49
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   39272.13 ms /   158 tokens
No. of rows:  51%| | 638/1258 [5:53:21<6:29:52, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.01 ms /    21 runs   (    0.29
ms per token,  3493.60 tokens per second)
```

```
llama_print_timings: prompt eval time =   19355.79 ms /    86 tokens (  225.07
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =    5929.18 ms /    20 runs   (  296.46
ms per token,     3.37 tokens per second)
llama_print_timings:       total time =   25374.81 ms /   106 tokens
No. of rows:  51%| | 639/1258 [5:53:46<5:51:01, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.38 ms /    45 runs   (    0.28
ms per token,  3634.01 tokens per second)
llama_print_timings: prompt eval time =   24474.15 ms /   107 tokens (  228.73
ms per token,     4.37 tokens per second)
llama_print_timings:        eval time =   13651.05 ms /    44 runs   (  310.25
ms per token,     3.22 tokens per second)
llama_print_timings:       total time =   38306.92 ms /   151 tokens
No. of rows:  51%| | 640/1258 [5:54:24<6:03:42, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      22.29 ms /    50 runs   (    0.45
ms per token,  2243.36 tokens per second)
llama_print_timings: prompt eval time =   37823.54 ms /   157 tokens (  240.91
ms per token,     4.15 tokens per second)
llama_print_timings:        eval time =   24345.82 ms /    49 runs   (  496.85
ms per token,     2.01 tokens per second)
llama_print_timings:       total time =   62503.91 ms /   206 tokens
No. of rows:  51%| | 641/1258 [5:55:27<7:27:02, 43Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.77 ms /    21 runs   (    0.32
ms per token,  3103.75 tokens per second)
llama_print_timings: prompt eval time =   19172.70 ms /    76 tokens (  252.27
ms per token,     3.96 tokens per second)
llama_print_timings:        eval time =    5696.24 ms /    20 runs   (  284.81
ms per token,     3.51 tokens per second)
llama_print_timings:       total time =   24956.52 ms /    96 tokens
No. of rows:  51%| | 642/1258 [5:55:52<6:29:16, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.99 ms /    42 runs   (    0.29
ms per token,  3502.33 tokens per second)
llama_print_timings: prompt eval time =   31138.09 ms /   124 tokens (  251.11
ms per token,     3.98 tokens per second)
llama_print_timings:        eval time =   12862.68 ms /    41 runs   (  313.72
ms per token,     3.19 tokens per second)
```

```
llama_print_timings:       total time =   44185.30 ms /   165 tokens
No. of rows:  51%| | 643/1258 [5:56:36<6:47:57, 39Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.00 ms /    35 runs   (    0.29
ms per token,  3500.00 tokens per second)
llama_print_timings: prompt eval time =   25230.11 ms /   111 tokens (  227.30
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =    9795.02 ms /    34 runs   (  288.09
ms per token,     3.47 tokens per second)
llama_print_timings:       total time =   35167.16 ms /   145 tokens
No. of rows:  51%| | 644/1258 [5:57:11<6:33:06, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.59 ms /    24 runs   (    0.36
ms per token,  2793.30 tokens per second)
llama_print_timings: prompt eval time =   20268.75 ms /    76 tokens (  266.69
ms per token,     3.75 tokens per second)
llama_print_timings:        eval time =    7811.10 ms /    23 runs   (  339.61
ms per token,     2.94 tokens per second)
llama_print_timings:       total time =   28193.41 ms /    99 tokens
No. of rows:  51%| | 645/1258 [5:57:39<6:01:09, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.75 ms /    50 runs   (    0.28
ms per token,  3635.83 tokens per second)
llama_print_timings: prompt eval time =   46244.61 ms /   185 tokens (  249.97
ms per token,     4.00 tokens per second)
llama_print_timings:        eval time =   15601.42 ms /    49 runs   (  318.40
ms per token,     3.14 tokens per second)
llama_print_timings:       total time =   62195.76 ms /   234 tokens
No. of rows:  51%| | 646/1258 [5:58:42<7:22:43, 43Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.12 ms /    33 runs   (    0.31
ms per token,  3261.51 tokens per second)
llama_print_timings: prompt eval time =   21291.86 ms /    92 tokens (  231.43
ms per token,     4.32 tokens per second)
llama_print_timings:        eval time =   10366.00 ms /    32 runs   (  323.94
ms per token,     3.09 tokens per second)
llama_print_timings:       total time =   31804.73 ms /   124 tokens
No. of rows:  51%| | 647/1258 [5:59:13<6:46:35, 39Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.50 ms /    23 runs   (    0.28
ms per token,  3540.10 tokens per second)
llama_print_timings: prompt eval time =   20709.24 ms /    78 tokens (  265.50
ms per token,     3.77 tokens per second)
llama_print_timings:        eval time =    6770.20 ms /    22 runs   (  307.74
ms per token,     3.25 tokens per second)
llama_print_timings:       total time =   27575.81 ms /   100 tokens
No. of rows:  52%| | 648/1258 [5:59:41<6:08:15, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.74 ms /    34 runs   (    0.29
ms per token,  3489.68 tokens per second)
llama_print_timings: prompt eval time =   27140.53 ms /   115 tokens (  236.00
ms per token,     4.24 tokens per second)
llama_print_timings:        eval time =    9736.57 ms /    33 runs   (  295.05
ms per token,     3.39 tokens per second)
llama_print_timings:       total time =   37021.75 ms /   148 tokens
No. of rows:  52%| | 649/1258 [6:00:18<6:10:06, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.17 ms /    50 runs   (    0.26
ms per token,  3796.80 tokens per second)
llama_print_timings: prompt eval time =   27887.43 ms /   122 tokens (  228.59
ms per token,     4.37 tokens per second)
llama_print_timings:        eval time =   14207.40 ms /    49 runs   (  289.95
ms per token,     3.45 tokens per second)
llama_print_timings:       total time =   42296.32 ms /   171 tokens
No. of rows:  52%| | 650/1258 [6:01:00<6:27:16, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.19 ms /    38 runs   (    0.27
ms per token,  3730.98 tokens per second)
llama_print_timings: prompt eval time =   26533.60 ms /   111 tokens (  239.04
ms per token,     4.18 tokens per second)
llama_print_timings:        eval time =   12186.76 ms /    37 runs   (  329.37
ms per token,     3.04 tokens per second)
llama_print_timings:       total time =   38868.83 ms /   148 tokens
No. of rows:  52%| | 651/1258 [6:01:39<6:28:36, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.28 ms /    26 runs   (    0.28
ms per token,  3570.45 tokens per second)
llama_print_timings: prompt eval time =   19371.60 ms /    89 tokens (  217.66
```

```
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    7064.15 ms /    25 runs   (  282.57
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   26542.98 ms /   114 tokens
No. of rows:  52%| | 652/1258 [6:02:06<5:52:04, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.26 ms /    32 runs   (    0.26
ms per token,  3875.97 tokens per second)
llama_print_timings: prompt eval time =   20603.76 ms /    93 tokens (  221.55
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    8625.20 ms /    31 runs   (  278.23
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   29352.41 ms /   124 tokens
No. of rows:  52%| | 653/1258 [6:02:35<5:34:51, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.60 ms /    31 runs   (    0.28
ms per token,  3603.39 tokens per second)
llama_print_timings: prompt eval time =   18396.66 ms /    85 tokens (  216.43
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    8525.85 ms /    30 runs   (  284.20
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   27046.98 ms /   115 tokens
No. of rows:  52%| | 654/1258 [6:03:02<5:15:43, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.02 ms /    47 runs   (    0.26
ms per token,  3909.50 tokens per second)
llama_print_timings: prompt eval time =   19298.36 ms /    83 tokens (  232.51
ms per token,     4.30 tokens per second)
llama_print_timings:        eval time =   13096.63 ms /    46 runs   (  284.71
ms per token,     3.51 tokens per second)
llama_print_timings:       total time =   32579.97 ms /   129 tokens
No. of rows:  52%| | 655/1258 [6:03:35<5:18:50, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.86 ms /    34 runs   (    0.26
ms per token,  3837.47 tokens per second)
llama_print_timings: prompt eval time =   21069.92 ms /    90 tokens (  234.11
ms per token,     4.27 tokens per second)
llama_print_timings:        eval time =    9166.76 ms /    33 runs   (  277.78
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   30368.86 ms /   123 tokens
```

```
No. of rows:  52%| | 656/1258 [6:04:05<5:14:16, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.41 ms /    49 runs   (    0.27
ms per token,  3653.72 tokens per second)
llama_print_timings: prompt eval time =   24131.51 ms /   112 tokens (  215.46
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   13538.17 ms /    48 runs   (  282.05
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   37867.12 ms /   160 tokens
No. of rows:  52%| | 657/1258 [6:04:43<5:33:26, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.84 ms /    37 runs   (    0.27
ms per token,  3760.16 tokens per second)
llama_print_timings: prompt eval time =   19122.01 ms /    89 tokens (  214.85
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   11827.40 ms /    36 runs   (  328.54
ms per token,     3.04 tokens per second)
llama_print_timings:       total time =   31091.26 ms /   125 tokens
No. of rows:  52%| | 658/1258 [6:05:14<5:26:18, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.95 ms /    50 runs   (    0.26
ms per token,  3860.11 tokens per second)
llama_print_timings: prompt eval time =   28302.37 ms /   131 tokens (  216.05
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   13581.38 ms /    49 runs   (  277.17
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   42079.76 ms /   180 tokens
No. of rows:  52%| | 659/1258 [6:05:56<5:54:05, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.44 ms /    50 runs   (    0.27
ms per token,  3721.35 tokens per second)
llama_print_timings: prompt eval time =   26033.04 ms /   122 tokens (  213.39
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   15104.75 ms /    49 runs   (  308.26
ms per token,     3.24 tokens per second)
llama_print_timings:       total time =   41331.41 ms /   171 tokens
No. of rows:  52%| | 660/1258 [6:06:38<6:11:01, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      12.84 ms /    49 runs   (    0.26
ms per token,  3815.61 tokens per second)
llama_print_timings: prompt eval time =   22940.13 ms /   107 tokens (  214.39
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13362.22 ms /    48 runs   (  278.38
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   36493.42 ms /   155 tokens
No. of rows:  53%| | 661/1258 [6:07:14<6:08:14, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.10 ms /    32 runs   (    0.25
ms per token,  3952.57 tokens per second)
llama_print_timings: prompt eval time =   17388.22 ms /    73 tokens (  238.19
ms per token,     4.20 tokens per second)
llama_print_timings:        eval time =    8465.18 ms /    31 runs   (  273.07
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   25975.11 ms /   104 tokens
No. of rows:  53%| | 662/1258 [6:07:40<5:34:45, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.38 ms /    38 runs   (    0.27
ms per token,  3659.48 tokens per second)
llama_print_timings: prompt eval time =   20346.10 ms /    96 tokens (  211.94
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   10018.32 ms /    37 runs   (  270.77
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   30512.46 ms /   133 tokens
No. of rows:  53%| | 663/1258 [6:08:11<5:24:44, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.63 ms /    50 runs   (    0.25
ms per token,  3959.46 tokens per second)
llama_print_timings: prompt eval time =   25749.38 ms /   121 tokens (  212.80
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   13493.97 ms /    49 runs   (  275.39
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   39436.35 ms /   170 tokens
No. of rows:  53%| | 664/1258 [6:08:50<5:44:05, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.25 ms /    50 runs   (    0.27
ms per token,  3773.30 tokens per second)
llama_print_timings: prompt eval time =   28719.11 ms /   127 tokens (  226.13
ms per token,     4.42 tokens per second)
```

```
llama_print_timings:        eval time =   13680.30 ms /    49 runs   (  279.19
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   42594.59 ms /   176 tokens
No. of rows:  53%| | 665/1258 [6:09:33<6:06:47, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.87 ms /    26 runs   (    0.26
ms per token,  3784.02 tokens per second)
llama_print_timings: prompt eval time =   17946.78 ms /    83 tokens (  216.23
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    6931.22 ms /    25 runs   (  277.25
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   24977.77 ms /   108 tokens
No. of rows:  53%| | 666/1258 [6:09:58<5:30:16, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.81 ms /    39 runs   (    0.28
ms per token,  3606.44 tokens per second)
llama_print_timings: prompt eval time =   19346.28 ms /    89 tokens (  217.37
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =   11196.84 ms /    38 runs   (  294.65
ms per token,     3.39 tokens per second)
llama_print_timings:       total time =   30703.12 ms /   127 tokens
No. of rows:  53%| | 667/1258 [6:10:28<5:21:33, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.10 ms /    46 runs   (    0.28
ms per token,  3510.38 tokens per second)
llama_print_timings: prompt eval time =   20155.98 ms /    94 tokens (  214.43
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   12656.09 ms /    45 runs   (  281.25
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   33001.91 ms /   139 tokens
No. of rows:  53%| | 668/1258 [6:11:01<5:22:05, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.47 ms /    37 runs   (    0.26
ms per token,  3905.01 tokens per second)
llama_print_timings: prompt eval time =   22420.83 ms /   105 tokens (  213.53
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    9900.26 ms /    36 runs   (  275.01
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   32461.22 ms /   141 tokens
No. of rows:  53%| | 669/1258 [6:11:34<5:20:40, 32Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.33 ms /    39 runs   (    0.26
ms per token,  3774.32 tokens per second)
llama_print_timings: prompt eval time =   25498.40 ms /   119 tokens (  214.27
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   10272.88 ms /    38 runs   (  270.34
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   35921.49 ms /   157 tokens
No. of rows:  53%| | 670/1258 [6:12:10<5:29:45, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.70 ms /    47 runs   (    0.27
ms per token,  3699.91 tokens per second)
llama_print_timings: prompt eval time =   24282.60 ms /   113 tokens (  214.89
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   12583.04 ms /    46 runs   (  273.54
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   37047.44 ms /   159 tokens
No. of rows:  53%| | 671/1258 [6:12:47<5:39:10, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.72 ms /    34 runs   (    0.26
ms per token,  3897.74 tokens per second)
llama_print_timings: prompt eval time =   21949.29 ms /    97 tokens (  226.28
ms per token,     4.42 tokens per second)
llama_print_timings:        eval time =    9081.29 ms /    33 runs   (  275.19
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   31163.59 ms /   130 tokens
No. of rows:  53%| | 672/1258 [6:13:18<5:28:20, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.16 ms /    38 runs   (    0.27
ms per token,  3741.63 tokens per second)
llama_print_timings: prompt eval time =   24828.84 ms /   108 tokens (  229.90
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =   10573.52 ms /    37 runs   (  285.77
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   35554.19 ms /   145 tokens
No. of rows:  53%| | 673/1258 [6:13:54<5:33:31, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.77 ms /    50 runs   (    0.26
```

```
ms per token,  3914.20 tokens per second)
llama_print_timings: prompt eval time =   24169.87 ms /   112 tokens (  215.80
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   13864.93 ms /    49 runs   (  282.96
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   38234.79 ms /   161 tokens
No. of rows:  54%| | 674/1258 [6:14:32<5:44:43, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.65 ms /    50 runs   (    0.27
ms per token,  3662.74 tokens per second)
llama_print_timings: prompt eval time =   25450.65 ms /   118 tokens (  215.68
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   13872.38 ms /    49 runs   (  283.11
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   39518.92 ms /   167 tokens
No. of rows:  54%| | 675/1258 [6:15:11<5:56:05, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.17 ms /    23 runs   (    0.27
ms per token,  3724.70 tokens per second)
llama_print_timings: prompt eval time =   18404.45 ms /    84 tokens (  219.10
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    6127.53 ms /    22 runs   (  278.52
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   24620.90 ms /   106 tokens
No. of rows:  54%| | 676/1258 [6:15:36<5:20:29, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.89 ms /    25 runs   (    0.28
ms per token,  3630.55 tokens per second)
llama_print_timings: prompt eval time =   20457.87 ms /    86 tokens (  237.88
ms per token,     4.20 tokens per second)
llama_print_timings:        eval time =    6735.83 ms /    24 runs   (  280.66
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   27293.54 ms /   110 tokens
No. of rows:  54%| | 677/1258 [6:16:03<5:03:16, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.50 ms /    28 runs   (    0.27
ms per token,  3732.84 tokens per second)
llama_print_timings: prompt eval time =   17348.19 ms /    80 tokens (  216.85
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    7433.68 ms /    27 runs   (  275.32
```

```
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   24889.75 ms /   107 tokens
No. of rows:  54%| | 678/1258 [6:16:28<4:44:08, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.04 ms /    48 runs   (    0.27
ms per token,  3680.98 tokens per second)
llama_print_timings: prompt eval time =   20499.17 ms /    94 tokens (  218.08
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =   15151.00 ms /    47 runs   (  322.36
ms per token,     3.10 tokens per second)
llama_print_timings:       total time =   35841.53 ms /   141 tokens
No. of rows:  54%| | 679/1258 [6:17:04<5:02:21, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.92 ms /    23 runs   (    0.26
ms per token,  3887.76 tokens per second)
llama_print_timings: prompt eval time =   25217.02 ms /   116 tokens (  217.39
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    6171.56 ms /    22 runs   (  280.53
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   31476.86 ms /   138 tokens
No. of rows:  54%| | 680/1258 [6:17:35<5:02:15, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.67 ms /    50 runs   (    0.27
ms per token,  3657.91 tokens per second)
llama_print_timings: prompt eval time =   22758.66 ms /    98 tokens (  232.23
ms per token,     4.31 tokens per second)
llama_print_timings:        eval time =   13717.58 ms /    49 runs   (  279.95
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   36671.36 ms /   147 tokens
No. of rows:  54%| | 681/1258 [6:18:12<5:17:02, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.21 ms /    35 runs   (    0.26
ms per token,  3802.28 tokens per second)
llama_print_timings: prompt eval time =   22727.09 ms /   104 tokens (  218.53
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    9563.56 ms /    34 runs   (  281.28
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   32427.30 ms /   138 tokens
No. of rows:  54%| | 682/1258 [6:18:45<5:14:56, 32Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.52 ms /    35 runs   (    0.27
ms per token,  3676.47 tokens per second)
llama_print_timings: prompt eval time =   26870.97 ms /   116 tokens (  231.65
ms per token,     4.32 tokens per second)
llama_print_timings:        eval time =    9543.47 ms /    34 runs   (  280.69
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   36551.70 ms /   150 tokens
No. of rows:  54%| | 683/1258 [6:19:21<5:25:11, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.92 ms /    35 runs   (    0.28
ms per token,  3528.58 tokens per second)
llama_print_timings: prompt eval time =   23076.30 ms /   104 tokens (  221.89
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    9457.17 ms /    34 runs   (  278.15
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   32674.55 ms /   138 tokens
No. of rows:  54%| | 684/1258 [6:19:54<5:21:01, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.53 ms /    23 runs   (    0.28
ms per token,  3524.36 tokens per second)
llama_print_timings: prompt eval time =   20625.75 ms /    93 tokens (  221.78
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    6858.48 ms /    22 runs   (  311.75
ms per token,     3.21 tokens per second)
llama_print_timings:       total time =   27583.70 ms /   115 tokens
No. of rows:  54%| | 685/1258 [6:20:21<5:03:22, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.17 ms /    40 runs   (    0.28
ms per token,  3581.98 tokens per second)
llama_print_timings: prompt eval time =   25237.66 ms /   112 tokens (  225.34
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =   10925.70 ms /    39 runs   (  280.15
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   36322.56 ms /   151 tokens
No. of rows:  55%| | 686/1258 [6:20:58<5:15:55, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.75 ms /    36 runs   (    0.27
ms per token,  3693.44 tokens per second)
```

```
llama_print_timings: prompt eval time =   20804.10 ms /    87 tokens (  239.13
ms per token,     4.18 tokens per second)
llama_print_timings:        eval time =    9702.69 ms /    35 runs   (  277.22
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   30648.34 ms /   122 tokens
No. of rows:  55%| | 687/1258 [6:21:28<5:08:14, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.71 ms /    30 runs   (    0.26
ms per token,  3892.06 tokens per second)
llama_print_timings: prompt eval time =   19633.29 ms /    89 tokens (  220.60
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    8169.96 ms /    29 runs   (  281.72
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   27917.62 ms /   118 tokens
No. of rows:  55%| | 688/1258 [6:21:56<4:54:59, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.97 ms /    40 runs   (    0.27
ms per token,  3644.98 tokens per second)
llama_print_timings: prompt eval time =   25666.67 ms /   111 tokens (  231.23
ms per token,     4.32 tokens per second)
llama_print_timings:        eval time =   12722.73 ms /    39 runs   (  326.22
ms per token,     3.07 tokens per second)
llama_print_timings:       total time =   38547.62 ms /   150 tokens
No. of rows:  55%| | 689/1258 [6:22:35<5:15:50, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.64 ms /    31 runs   (    0.28
ms per token,  3588.79 tokens per second)
llama_print_timings: prompt eval time =   22126.01 ms /   101 tokens (  219.07
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    8496.31 ms /    30 runs   (  283.21
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   30745.65 ms /   131 tokens
No. of rows:  55%| | 690/1258 [6:23:06<5:08:02, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.41 ms /    38 runs   (    0.27
ms per token,  3649.99 tokens per second)
llama_print_timings: prompt eval time =   25268.35 ms /   109 tokens (  231.82
ms per token,     4.31 tokens per second)
llama_print_timings:        eval time =   10373.52 ms /    37 runs   (  280.37
ms per token,     3.57 tokens per second)
```

```
llama_print_timings:       total time =   35790.70 ms /   146 tokens
No. of rows:  55%| | 691/1258 [6:23:41<5:16:42, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.50 ms /    44 runs   (    0.26
ms per token,  3826.75 tokens per second)
llama_print_timings: prompt eval time =   25861.27 ms /   117 tokens (  221.04
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =   11902.57 ms /    43 runs   (  276.80
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   37936.02 ms /   160 tokens
No. of rows:  55%| | 692/1258 [6:24:19<5:28:42, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.64 ms /    50 runs   (    0.27
ms per token,  3665.15 tokens per second)
llama_print_timings: prompt eval time =   26689.36 ms /   123 tokens (  216.99
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   15399.59 ms /    49 runs   (  314.28
ms per token,     3.18 tokens per second)
llama_print_timings:       total time =   42285.05 ms /   172 tokens
No. of rows:  55%| | 693/1258 [6:25:02<5:49:07, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.68 ms /    40 runs   (    0.27
ms per token,  3744.62 tokens per second)
llama_print_timings: prompt eval time =   20289.23 ms /    91 tokens (  222.96
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =   11026.67 ms /    39 runs   (  282.74
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   31472.82 ms /   130 tokens
No. of rows:  55%| | 694/1258 [6:25:33<5:32:44, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.77 ms /    30 runs   (    0.26
ms per token,  3860.01 tokens per second)
llama_print_timings: prompt eval time =   19871.38 ms /    90 tokens (  220.79
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    7970.55 ms /    29 runs   (  274.85
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   27958.81 ms /   119 tokens
No. of rows:  55%| | 695/1258 [6:26:01<5:11:14, 33Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.58 ms /    33 runs   (    0.26
ms per token,  3847.50 tokens per second)
llama_print_timings: prompt eval time =   19255.77 ms /    87 tokens (  221.33
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    9067.19 ms /    32 runs   (  283.35
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   28453.33 ms /   119 tokens
No. of rows:  55%| | 696/1258 [6:26:30<4:57:25, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.78 ms /    31 runs   (    0.28
ms per token,  3529.95 tokens per second)
llama_print_timings: prompt eval time =   22912.69 ms /   104 tokens (  220.31
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    8404.79 ms /    30 runs   (  280.16
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   31445.15 ms /   134 tokens
No. of rows:  55%| | 697/1258 [6:27:01<4:56:03, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.25 ms /    40 runs   (    0.26
ms per token,  3903.96 tokens per second)
llama_print_timings: prompt eval time =   27374.73 ms /   117 tokens (  233.97
ms per token,     4.27 tokens per second)
llama_print_timings:        eval time =   11225.44 ms /    39 runs   (  287.83
ms per token,     3.47 tokens per second)
llama_print_timings:       total time =   38758.52 ms /   156 tokens
No. of rows:  55%| | 698/1258 [6:27:40<5:15:25, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.79 ms /    50 runs   (    0.28
ms per token,  3625.29 tokens per second)
llama_print_timings: prompt eval time =   22969.53 ms /   106 tokens (  216.69
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   13830.31 ms /    49 runs   (  282.25
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   37007.62 ms /   155 tokens
No. of rows:  56%| | 699/1258 [6:28:17<5:23:51, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.96 ms /    49 runs   (    0.26
ms per token,  3779.70 tokens per second)
llama_print_timings: prompt eval time =   26620.72 ms /   122 tokens (  218.20
```

```
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =   13406.63 ms /    48 runs   (  279.30
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   40220.39 ms /   170 tokens
No. of rows:  56%| | 700/1258 [6:28:57<5:38:32, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.23 ms /    34 runs   (    0.27
ms per token,  3683.24 tokens per second)
llama_print_timings: prompt eval time =   23163.67 ms /    98 tokens (  236.36
ms per token,     4.23 tokens per second)
llama_print_timings:        eval time =    9121.33 ms /    33 runs   (  276.40
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   32420.64 ms /   131 tokens
No. of rows:  56%| | 701/1258 [6:29:29<5:26:50, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.70 ms /    38 runs   (    0.28
ms per token,  3550.41 tokens per second)
llama_print_timings: prompt eval time =   21195.68 ms /    90 tokens (  235.51
ms per token,     4.25 tokens per second)
llama_print_timings:        eval time =   11933.48 ms /    37 runs   (  322.53
ms per token,     3.10 tokens per second)
llama_print_timings:       total time =   33282.00 ms /   127 tokens
No. of rows:  56%| | 702/1258 [6:30:03<5:20:56, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.33 ms /    33 runs   (    0.28
ms per token,  3535.46 tokens per second)
llama_print_timings: prompt eval time =   21070.69 ms /    96 tokens (  219.49
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    9260.51 ms /    32 runs   (  289.39
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   30463.15 ms /   128 tokens
No. of rows:  56%| | 703/1258 [6:30:33<5:08:46, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.15 ms /    30 runs   (    0.27
ms per token,  3679.18 tokens per second)
llama_print_timings: prompt eval time =   17479.31 ms /    80 tokens (  218.49
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    8059.30 ms /    29 runs   (  277.91
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   25658.14 ms /   109 tokens
```

```
No. of rows:  56%| | 704/1258 [6:30:59<4:46:53, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.18 ms /    27 runs   (    0.27
ms per token,  3761.49 tokens per second)
llama_print_timings: prompt eval time =   21051.16 ms /    95 tokens (  221.59
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    7391.71 ms /    26 runs   (  284.30
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   28546.76 ms /   121 tokens
No. of rows:  56%| | 705/1258 [6:31:27<4:39:23, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.92 ms /    36 runs   (    0.28
ms per token,  3629.40 tokens per second)
llama_print_timings: prompt eval time =   23405.60 ms /   109 tokens (  214.73
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    9676.77 ms /    35 runs   (  276.48
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   33224.20 ms /   144 tokens
No. of rows:  56%| | 706/1258 [6:32:01<4:46:56, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.11 ms /    50 runs   (    0.26
ms per token,  3814.17 tokens per second)
llama_print_timings: prompt eval time =   26145.15 ms /   112 tokens (  233.44
ms per token,     4.28 tokens per second)
llama_print_timings:        eval time =   13831.08 ms /    49 runs   (  282.27
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   40173.90 ms /   161 tokens
No. of rows:  56%| | 707/1258 [6:32:41<5:11:12, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.85 ms /    31 runs   (    0.29
ms per token,  3503.22 tokens per second)
llama_print_timings: prompt eval time =   20454.54 ms /    93 tokens (  219.94
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =    8796.79 ms /    30 runs   (  293.23
ms per token,     3.41 tokens per second)
llama_print_timings:       total time =   29375.03 ms /   123 tokens
No. of rows:  56%| | 708/1258 [6:33:10<4:58:14, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =       7.25 ms /    28 runs   (    0.26
ms per token,  3864.20 tokens per second)
llama_print_timings: prompt eval time =   17364.70 ms /    79 tokens (  219.81
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =    7547.52 ms /    27 runs   (  279.54
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   25021.46 ms /   106 tokens
No. of rows:  56%| | 709/1258 [6:33:35<4:37:04, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.43 ms /    36 runs   (    0.26
ms per token,  3815.98 tokens per second)
llama_print_timings: prompt eval time =   20093.56 ms /    85 tokens (  236.39
ms per token,     4.23 tokens per second)
llama_print_timings:        eval time =    9760.69 ms /    35 runs   (  278.88
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   29992.52 ms /   120 tokens
No. of rows:  56%| | 710/1258 [6:34:05<4:35:50, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.39 ms /    35 runs   (    0.27
ms per token,  3726.97 tokens per second)
llama_print_timings: prompt eval time =   18914.24 ms /    85 tokens (  222.52
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =    9660.58 ms /    34 runs   (  284.13
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   28711.94 ms /   119 tokens
No. of rows:  57%| | 711/1258 [6:34:34<4:31:15, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.10 ms /    37 runs   (    0.27
ms per token,  3662.28 tokens per second)
llama_print_timings: prompt eval time =   19702.91 ms /    89 tokens (  221.38
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =   11867.56 ms /    36 runs   (  329.65
ms per token,     3.03 tokens per second)
llama_print_timings:       total time =   31714.38 ms /   125 tokens
No. of rows:  57%| | 712/1258 [6:35:06<4:36:08, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.91 ms /    50 runs   (    0.28
ms per token,  3595.05 tokens per second)
llama_print_timings: prompt eval time =   22675.48 ms /   102 tokens (  222.31
ms per token,     4.50 tokens per second)
```

```
llama_print_timings:        eval time =   13756.55 ms /    49 runs   (  280.75
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   36631.95 ms /   151 tokens
No. of rows:  57%| | 713/1258 [6:35:42<4:52:46, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.67 ms /    28 runs   (    0.27
ms per token,  3649.16 tokens per second)
llama_print_timings: prompt eval time =   20407.18 ms /    85 tokens (  240.08
ms per token,     4.17 tokens per second)
llama_print_timings:        eval time =    7457.72 ms /    27 runs   (  276.21
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   27974.06 ms /   112 tokens
No. of rows:  57%| | 714/1258 [6:36:10<4:40:41, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.70 ms /    50 runs   (    0.25
ms per token,  3935.77 tokens per second)
llama_print_timings: prompt eval time =   30361.57 ms /   132 tokens (  230.01
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =   13877.90 ms /    49 runs   (  283.22
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   44437.98 ms /   181 tokens
No. of rows:  57%| | 715/1258 [6:36:55<5:16:46, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.46 ms /    30 runs   (    0.28
ms per token,  3547.78 tokens per second)
llama_print_timings: prompt eval time =   21746.66 ms /    98 tokens (  221.90
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    8101.83 ms /    29 runs   (  279.37
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   29967.49 ms /   127 tokens
No. of rows:  57%| | 716/1258 [6:37:25<5:02:33, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.98 ms /    33 runs   (    0.27
ms per token,  3672.79 tokens per second)
llama_print_timings: prompt eval time =   25313.66 ms /   109 tokens (  232.24
ms per token,     4.31 tokens per second)
llama_print_timings:        eval time =    9057.36 ms /    32 runs   (  283.04
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   34499.71 ms /   141 tokens
No. of rows:  57%| | 717/1258 [6:37:59<5:04:43, 33Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.89 ms /    50 runs   (    0.28
ms per token,  3599.19 tokens per second)
llama_print_timings: prompt eval time =   29072.74 ms /   125 tokens (  232.58
ms per token,     4.30 tokens per second)
llama_print_timings:        eval time =   14015.77 ms /    49 runs   (  286.04
ms per token,     3.50 tokens per second)
llama_print_timings:       total time =   43288.95 ms /   174 tokens
No. of rows:  57%| | 718/1258 [6:38:43<5:29:51, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.46 ms /    50 runs   (    0.27
ms per token,  3714.99 tokens per second)
llama_print_timings: prompt eval time =   25781.30 ms /   117 tokens (  220.35
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =   13832.36 ms /    49 runs   (  282.29
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   39810.70 ms /   166 tokens
No. of rows:  57%| | 719/1258 [6:39:22<5:37:46, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.78 ms /    24 runs   (    0.28
ms per token,  3541.39 tokens per second)
llama_print_timings: prompt eval time =   18791.69 ms /    85 tokens (  221.08
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    6440.25 ms /    23 runs   (  280.01
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   25327.92 ms /   108 tokens
No. of rows:  57%| | 720/1258 [6:39:48<5:04:07, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.45 ms /    31 runs   (    0.27
ms per token,  3669.51 tokens per second)
llama_print_timings: prompt eval time =   18810.51 ms /    85 tokens (  221.30
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    8466.78 ms /    30 runs   (  282.23
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   27401.50 ms /   115 tokens
No. of rows:  57%| | 721/1258 [6:40:15<4:46:06, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.78 ms /    36 runs   (    0.27
```

```
ms per token,  3682.49 tokens per second)
llama_print_timings: prompt eval time =   18668.17 ms /    84 tokens (  222.24
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =    9873.24 ms /    35 runs   (  282.09
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   28687.41 ms /   119 tokens
No. of rows:  57%| | 722/1258 [6:40:44<4:36:47, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.87 ms /    42 runs   (    0.28
ms per token,  3537.74 tokens per second)
llama_print_timings: prompt eval time =   20911.96 ms /    95 tokens (  220.13
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =   11841.73 ms /    41 runs   (  288.82
ms per token,     3.46 tokens per second)
llama_print_timings:       total time =   32923.87 ms /   136 tokens
No. of rows:  57%| | 723/1258 [6:41:17<4:41:29, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.72 ms /    34 runs   (    0.26
ms per token,  3897.29 tokens per second)
llama_print_timings: prompt eval time =   21800.76 ms /    92 tokens (  236.96
ms per token,     4.22 tokens per second)
llama_print_timings:        eval time =    9267.56 ms /    33 runs   (  280.84
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   31199.87 ms /   125 tokens
No. of rows:  58%| | 724/1258 [6:41:48<4:39:58, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.32 ms /    31 runs   (    0.27
ms per token,  3726.41 tokens per second)
llama_print_timings: prompt eval time =   18769.96 ms /    85 tokens (  220.82
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    8382.17 ms /    30 runs   (  279.41
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   27273.38 ms /   115 tokens
No. of rows:  58%| | 725/1258 [6:42:15<4:28:20, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.60 ms /    50 runs   (    0.27
ms per token,  3677.55 tokens per second)
llama_print_timings: prompt eval time =   33054.60 ms /   140 tokens (  236.10
ms per token,     4.24 tokens per second)
llama_print_timings:        eval time =   13845.97 ms /    49 runs   (  282.57
```

```
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   47099.67 ms /   189 tokens
No. of rows:  58%| | 726/1258 [6:43:02<5:12:47, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.06 ms /    37 runs   (    0.27
ms per token,  3678.30 tokens per second)
llama_print_timings: prompt eval time =   18139.22 ms /    81 tokens (  223.94
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   10175.56 ms /    36 runs   (  282.65
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   28463.69 ms /   117 tokens
No. of rows:  58%| | 727/1258 [6:43:31<4:54:07, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.04 ms /    38 runs   (    0.26
ms per token,  3784.11 tokens per second)
llama_print_timings: prompt eval time =   22602.65 ms /   103 tokens (  219.44
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   10247.68 ms /    37 runs   (  276.96
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   32999.97 ms /   140 tokens
No. of rows:  58%| | 728/1258 [6:44:04<4:52:58, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.80 ms /    34 runs   (    0.26
ms per token,  3862.32 tokens per second)
llama_print_timings: prompt eval time =   20820.89 ms /    95 tokens (  219.17
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    9268.76 ms /    33 runs   (  280.87
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   30221.80 ms /   128 tokens
No. of rows:  58%| | 729/1258 [6:44:34<4:44:38, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.08 ms /    26 runs   (    0.27
ms per token,  3672.32 tokens per second)
llama_print_timings: prompt eval time =   19040.90 ms /    86 tokens (  221.41
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    6935.30 ms /    25 runs   (  277.41
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   26078.09 ms /   111 tokens
No. of rows:  58%| | 730/1258 [6:45:00<4:27:44, 30Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.85 ms /    25 runs   (    0.27
ms per token,  3648.57 tokens per second)
llama_print_timings: prompt eval time =   19521.24 ms /    88 tokens (  221.83
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    6878.89 ms /    24 runs   (  286.62
ms per token,     3.49 tokens per second)
llama_print_timings:       total time =   26499.64 ms /   112 tokens
No. of rows:  58%| | 731/1258 [6:45:27<4:16:52, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.55 ms /    50 runs   (    0.27
ms per token,  3690.04 tokens per second)
llama_print_timings: prompt eval time =   18173.86 ms /    79 tokens (  230.05
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =   13698.46 ms /    49 runs   (  279.56
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   32071.07 ms /   128 tokens
No. of rows:  58%| | 732/1258 [6:45:59<4:23:50, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.79 ms /    50 runs   (    0.26
ms per token,  3910.53 tokens per second)
llama_print_timings: prompt eval time =   28256.18 ms /   122 tokens (  231.61
ms per token,     4.32 tokens per second)
llama_print_timings:        eval time =   13650.56 ms /    49 runs   (  278.58
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   42105.96 ms /   171 tokens
No. of rows:  58%| | 733/1258 [6:46:41<4:54:54, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.24 ms /    50 runs   (    0.26
ms per token,  3776.15 tokens per second)
llama_print_timings: prompt eval time =   32599.88 ms /   151 tokens (  215.89
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   13664.43 ms /    49 runs   (  278.87
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   46460.65 ms /   200 tokens
No. of rows:  58%| | 734/1258 [6:47:27<5:27:47, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.02 ms /    31 runs   (    0.26
ms per token,  3867.27 tokens per second)
```

```
llama_print_timings: prompt eval time =   17702.12 ms /    81 tokens (  218.54
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =   10048.08 ms /    30 runs   (  334.94
ms per token,     2.99 tokens per second)
llama_print_timings:       total time =   27871.04 ms /   111 tokens
No. of rows:  58%| | 735/1258 [6:47:55<5:01:56, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.90 ms /    27 runs   (    0.26
ms per token,  3913.61 tokens per second)
llama_print_timings: prompt eval time =   19993.95 ms /    93 tokens (  214.99
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    7172.71 ms /    26 runs   (  275.87
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   27272.06 ms /   119 tokens
No. of rows:  59%| | 736/1258 [6:48:22<4:42:07, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.22 ms /    50 runs   (    0.26
ms per token,  3782.15 tokens per second)
llama_print_timings: prompt eval time =   26883.86 ms /   125 tokens (  215.07
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13556.82 ms /    49 runs   (  276.67
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   40639.31 ms /   174 tokens
No. of rows:  59%| | 737/1258 [6:49:03<5:03:00, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      16.01 ms /    50 runs   (    0.32
ms per token,  3123.83 tokens per second)
llama_print_timings: prompt eval time =   30946.67 ms /   108 tokens (  286.54
ms per token,     3.49 tokens per second)
llama_print_timings:        eval time =   15625.86 ms /    49 runs   (  318.90
ms per token,     3.14 tokens per second)
llama_print_timings:       total time =   46800.04 ms /   157 tokens
No. of rows:  59%| | 738/1258 [6:49:50<5:33:22, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.56 ms /    37 runs   (    0.29
ms per token,  3504.78 tokens per second)
llama_print_timings: prompt eval time =   27397.61 ms /   113 tokens (  242.46
ms per token,     4.12 tokens per second)
llama_print_timings:        eval time =    9967.09 ms /    36 runs   (  276.86
ms per token,     3.61 tokens per second)
```

```
llama_print_timings:       total time =   37512.93 ms /   149 tokens
No. of rows:  59%| | 739/1258 [6:50:27<5:30:18, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.80 ms /    41 runs   (    0.26
ms per token,  3795.59 tokens per second)
llama_print_timings: prompt eval time =   22357.59 ms /   108 tokens (  207.01
ms per token,     4.83 tokens per second)
llama_print_timings:        eval time =   10586.79 ms /    40 runs   (  264.67
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   33105.64 ms /   148 tokens
No. of rows:  59%| | 740/1258 [6:51:01<5:16:31, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.89 ms /    26 runs   (    0.26
ms per token,  3774.13 tokens per second)
llama_print_timings: prompt eval time =   18897.59 ms /    92 tokens (  205.41
ms per token,     4.87 tokens per second)
llama_print_timings:        eval time =    7071.19 ms /    25 runs   (  282.85
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   26071.77 ms /   117 tokens
No. of rows:  59%| | 741/1258 [6:51:27<4:48:33, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.17 ms /    50 runs   (    0.26
ms per token,  3797.08 tokens per second)
llama_print_timings: prompt eval time =   23050.22 ms /   114 tokens (  202.19
ms per token,     4.95 tokens per second)
llama_print_timings:        eval time =   12749.94 ms /    49 runs   (  260.20
ms per token,     3.84 tokens per second)
llama_print_timings:       total time =   36000.20 ms /   163 tokens
No. of rows:  59%| | 742/1258 [6:52:03<4:54:28, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.67 ms /    45 runs   (    0.26
ms per token,  3856.37 tokens per second)
llama_print_timings: prompt eval time =   21009.10 ms /   103 tokens (  203.97
ms per token,     4.90 tokens per second)
llama_print_timings:        eval time =   11699.17 ms /    44 runs   (  265.89
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   32883.45 ms /   147 tokens
No. of rows:  59%| | 743/1258 [6:52:36<4:50:25, 33Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.21 ms /    50 runs   (    0.26
ms per token,  3784.15 tokens per second)
llama_print_timings: prompt eval time =   19464.84 ms /    94 tokens (  207.07
ms per token,     4.83 tokens per second)
llama_print_timings:        eval time =   12781.47 ms /    49 runs   (  260.85
ms per token,     3.83 tokens per second)
llama_print_timings:       total time =   32443.93 ms /   143 tokens
No. of rows:  59%| | 744/1258 [6:53:08<4:46:18, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.74 ms /    40 runs   (    0.27
ms per token,  3724.74 tokens per second)
llama_print_timings: prompt eval time =   20427.04 ms /    98 tokens (  208.44
ms per token,     4.80 tokens per second)
llama_print_timings:        eval time =   10610.62 ms /    39 runs   (  272.07
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   31197.08 ms /   137 tokens
No. of rows:  59%| | 745/1258 [6:53:39<4:40:05, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.38 ms /    30 runs   (    0.28
ms per token,  3577.82 tokens per second)
llama_print_timings: prompt eval time =   19774.45 ms /    98 tokens (  201.78
ms per token,     4.96 tokens per second)
llama_print_timings:        eval time =    7673.41 ms /    29 runs   (  264.60
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   27567.39 ms /   127 tokens
No. of rows:  59%| | 746/1258 [6:54:07<4:26:15, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.75 ms /    41 runs   (    0.26
ms per token,  3815.02 tokens per second)
llama_print_timings: prompt eval time =   20491.39 ms /   101 tokens (  202.89
ms per token,     4.93 tokens per second)
llama_print_timings:        eval time =   10721.93 ms /    40 runs   (  268.05
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   31373.63 ms /   141 tokens
No. of rows:  59%| | 747/1258 [6:54:38<4:26:10, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.94 ms /    33 runs   (    0.27
ms per token,  3691.28 tokens per second)
llama_print_timings: prompt eval time =   21416.15 ms /   104 tokens (  205.92
```

```
ms per token,     4.86 tokens per second)
llama_print_timings:        eval time =    8284.33 ms /    32 runs   (  258.89
ms per token,     3.86 tokens per second)
llama_print_timings:       total time =   29828.16 ms /   136 tokens
No. of rows:  59%| | 748/1258 [6:55:08<4:22:03, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.22 ms /    40 runs   (    0.26
ms per token,  3913.89 tokens per second)
llama_print_timings: prompt eval time =   20526.97 ms /    99 tokens (  207.34
ms per token,     4.82 tokens per second)
llama_print_timings:        eval time =   10401.88 ms /    39 runs   (  266.71
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   31082.12 ms /   138 tokens
No. of rows:  60%| | 749/1258 [6:55:39<4:22:10, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.87 ms /    39 runs   (    0.25
ms per token,  3949.77 tokens per second)
llama_print_timings: prompt eval time =   20091.87 ms /    98 tokens (  205.02
ms per token,     4.88 tokens per second)
llama_print_timings:        eval time =    9850.52 ms /    38 runs   (  259.22
ms per token,     3.86 tokens per second)
llama_print_timings:       total time =   30090.46 ms /   136 tokens
No. of rows:  60%| | 750/1258 [6:56:09<4:19:36, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.81 ms /    33 runs   (    0.27
ms per token,  3744.04 tokens per second)
llama_print_timings: prompt eval time =   20019.92 ms /    92 tokens (  217.61
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    8296.06 ms /    32 runs   (  259.25
ms per token,     3.86 tokens per second)
llama_print_timings:       total time =   28444.16 ms /   124 tokens
No. of rows:  60%| | 751/1258 [6:56:38<4:13:31, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.86 ms /    23 runs   (    0.25
ms per token,  3924.25 tokens per second)
llama_print_timings: prompt eval time =   16123.57 ms /    73 tokens (  220.87
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    6335.25 ms /    22 runs   (  287.97
ms per token,     3.47 tokens per second)
llama_print_timings:       total time =   22548.24 ms /    95 tokens
```

```
No. of rows:  60%| | 752/1258 [6:57:00<3:54:09, 27Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.54 ms /    32 runs   (    0.27
ms per token,  3746.63 tokens per second)
llama_print_timings: prompt eval time =   19837.97 ms /    97 tokens (  204.52
ms per token,     4.89 tokens per second)
llama_print_timings:        eval time =    8139.51 ms /    31 runs   (  262.56
ms per token,     3.81 tokens per second)
llama_print_timings:       total time =   28102.01 ms /   128 tokens
No. of rows:  60%| | 753/1258 [6:57:28<3:54:34, 27Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.07 ms /    34 runs   (    0.27
ms per token,  3747.80 tokens per second)
llama_print_timings: prompt eval time =   22114.00 ms /   111 tokens (  199.23
ms per token,     5.02 tokens per second)
llama_print_timings:        eval time =    8690.42 ms /    33 runs   (  263.35
ms per token,     3.80 tokens per second)
llama_print_timings:       total time =   30936.07 ms /   144 tokens
No. of rows:  60%| | 754/1258 [6:57:59<4:01:52, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.85 ms /    39 runs   (    0.28
ms per token,  3593.48 tokens per second)
llama_print_timings: prompt eval time =   23191.28 ms /   107 tokens (  216.74
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   10105.45 ms /    38 runs   (  265.93
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   33455.34 ms /   145 tokens
No. of rows:  60%| | 755/1258 [6:58:33<4:13:06, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.97 ms /    50 runs   (    0.26
ms per token,  3854.75 tokens per second)
llama_print_timings: prompt eval time =   30651.84 ms /   145 tokens (  211.39
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   13171.14 ms /    49 runs   (  268.80
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   44021.98 ms /   194 tokens
No. of rows:  60%| | 756/1258 [6:59:17<4:47:21, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =       7.62 ms /    29 runs   (    0.26
ms per token,  3805.77 tokens per second)
llama_print_timings: prompt eval time =   18201.42 ms /    88 tokens (  206.83
ms per token,     4.83 tokens per second)
llama_print_timings:        eval time =    7141.69 ms /    28 runs   (  255.06
ms per token,     3.92 tokens per second)
llama_print_timings:       total time =   25452.86 ms /   116 tokens
No. of rows:  60%| | 757/1258 [6:59:42<4:24:29, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.98 ms /    50 runs   (    0.26
ms per token,  3851.19 tokens per second)
llama_print_timings: prompt eval time =   21681.45 ms /   100 tokens (  216.81
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   12763.09 ms /    49 runs   (  260.47
ms per token,     3.84 tokens per second)
llama_print_timings:       total time =   34640.90 ms /   149 tokens
No. of rows:  60%| | 758/1258 [7:00:17<4:31:25, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.76 ms /    29 runs   (    0.27
ms per token,  3737.11 tokens per second)
llama_print_timings: prompt eval time =   16816.02 ms /    83 tokens (  202.60
ms per token,     4.94 tokens per second)
llama_print_timings:        eval time =    7273.33 ms /    28 runs   (  259.76
ms per token,     3.85 tokens per second)
llama_print_timings:       total time =   24198.88 ms /   111 tokens
No. of rows:  60%| | 759/1258 [7:00:41<4:09:59, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.59 ms /    36 runs   (    0.27
ms per token,  3754.30 tokens per second)
llama_print_timings: prompt eval time =   17644.76 ms /    86 tokens (  205.17
ms per token,     4.87 tokens per second)
llama_print_timings:        eval time =    9040.57 ms /    35 runs   (  258.30
ms per token,     3.87 tokens per second)
llama_print_timings:       total time =   26821.74 ms /   121 tokens
No. of rows:  60%| | 760/1258 [7:01:08<4:01:26, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.44 ms /    50 runs   (    0.27
ms per token,  3720.51 tokens per second)
llama_print_timings: prompt eval time =   28876.17 ms /   138 tokens (  209.25
ms per token,     4.78 tokens per second)
```

```
llama_print_timings:        eval time =   12626.28 ms /    49 runs   (  257.68
ms per token,     3.88 tokens per second)
llama_print_timings:       total time =   41697.02 ms /   187 tokens
No. of rows:  60%| | 761/1258 [7:01:50<4:32:21, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.69 ms /    45 runs   (    0.26
ms per token,  3851.09 tokens per second)
llama_print_timings: prompt eval time =   17762.08 ms /    86 tokens (  206.54
ms per token,     4.84 tokens per second)
llama_print_timings:        eval time =   11262.94 ms /    44 runs   (  255.98
ms per token,     3.91 tokens per second)
llama_print_timings:       total time =   29196.98 ms /   130 tokens
No. of rows:  61%| | 762/1258 [7:02:19<4:22:40, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.66 ms /    39 runs   (    0.27
ms per token,  3658.19 tokens per second)
llama_print_timings: prompt eval time =   22029.24 ms /   102 tokens (  215.97
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    9984.29 ms /    38 runs   (  262.74
ms per token,     3.81 tokens per second)
llama_print_timings:       total time =   32163.44 ms /   140 tokens
No. of rows:  61%| | 763/1258 [7:02:51<4:23:07, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.20 ms /    34 runs   (    0.27
ms per token,  3697.66 tokens per second)
llama_print_timings: prompt eval time =   20674.40 ms /   101 tokens (  204.70
ms per token,     4.89 tokens per second)
llama_print_timings:        eval time =    8757.87 ms /    33 runs   (  265.39
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   29566.41 ms /   134 tokens
No. of rows:  61%| | 764/1258 [7:03:21<4:16:52, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.71 ms /    50 runs   (    0.27
ms per token,  3648.30 tokens per second)
llama_print_timings: prompt eval time =   25728.80 ms /   125 tokens (  205.83
ms per token,     4.86 tokens per second)
llama_print_timings:        eval time =   12824.26 ms /    49 runs   (  261.72
ms per token,     3.82 tokens per second)
llama_print_timings:       total time =   38747.09 ms /   174 tokens
No. of rows:  61%| | 765/1258 [7:03:59<4:34:58, 33Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.75 ms /    42 runs   (    0.26
ms per token,  3906.61 tokens per second)
llama_print_timings: prompt eval time =   18203.56 ms /    91 tokens (  200.04
ms per token,     5.00 tokens per second)
llama_print_timings:        eval time =   10799.63 ms /    41 runs   (  263.41
ms per token,     3.80 tokens per second)
llama_print_timings:       total time =   29164.63 ms /   132 tokens
No. of rows:  61%| | 766/1258 [7:04:28<4:23:50, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.01 ms /    35 runs   (    0.26
ms per token,  3884.57 tokens per second)
llama_print_timings: prompt eval time =   20117.03 ms /    93 tokens (  216.31
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    9034.76 ms /    34 runs   (  265.73
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   29286.41 ms /   127 tokens
No. of rows:  61%| | 767/1258 [7:04:58<4:16:14, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.94 ms /    50 runs   (    0.26
ms per token,  3863.39 tokens per second)
llama_print_timings: prompt eval time =   24212.67 ms /   116 tokens (  208.73
ms per token,     4.79 tokens per second)
llama_print_timings:        eval time =   14700.87 ms /    49 runs   (  300.02
ms per token,     3.33 tokens per second)
llama_print_timings:       total time =   39113.02 ms /   165 tokens
No. of rows:  61%| | 768/1258 [7:05:37<4:34:49, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.01 ms /    41 runs   (    0.27
ms per token,  3722.87 tokens per second)
llama_print_timings: prompt eval time =   21222.44 ms /   106 tokens (  200.21
ms per token,     4.99 tokens per second)
llama_print_timings:        eval time =   10423.56 ms /    40 runs   (  260.59
ms per token,     3.84 tokens per second)
llama_print_timings:       total time =   31804.97 ms /   146 tokens
No. of rows:  61%| | 769/1258 [7:06:09<4:29:45, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.42 ms /    43 runs   (    0.27
```

```
ms per token,  3764.01 tokens per second)
llama_print_timings: prompt eval time =   24360.66 ms /   113 tokens (  215.58
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   10723.22 ms /    42 runs   (  255.31
ms per token,     3.92 tokens per second)
llama_print_timings:       total time =   35250.79 ms /   155 tokens
No. of rows:  61%| | 770/1258 [7:06:44<4:34:29, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.61 ms /    28 runs   (    0.27
ms per token,  3677.92 tokens per second)
llama_print_timings: prompt eval time =   18617.75 ms /    91 tokens (  204.59
ms per token,     4.89 tokens per second)
llama_print_timings:        eval time =    7084.30 ms /    27 runs   (  262.38
ms per token,     3.81 tokens per second)
llama_print_timings:       total time =   25811.72 ms /   118 tokens
No. of rows:  61%| | 771/1258 [7:07:10<4:14:36, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.90 ms /    35 runs   (    0.25
ms per token,  3932.14 tokens per second)
llama_print_timings: prompt eval time =   18910.81 ms /    92 tokens (  205.55
ms per token,     4.86 tokens per second)
llama_print_timings:        eval time =    8791.80 ms /    34 runs   (  258.58
ms per token,     3.87 tokens per second)
llama_print_timings:       total time =   27833.94 ms /   126 tokens
No. of rows:  61%| | 772/1258 [7:07:38<4:05:31, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.22 ms /    50 runs   (    0.26
ms per token,  3781.58 tokens per second)
llama_print_timings: prompt eval time =   30640.93 ms /   154 tokens (  198.97
ms per token,     5.03 tokens per second)
llama_print_timings:        eval time =   12602.40 ms /    49 runs   (  257.19
ms per token,     3.89 tokens per second)
llama_print_timings:       total time =   43439.77 ms /   203 tokens
No. of rows:  61%| | 773/1258 [7:08:21<4:36:53, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.47 ms /    39 runs   (    0.27
ms per token,  3724.57 tokens per second)
llama_print_timings: prompt eval time =   17523.38 ms /    85 tokens (  206.16
ms per token,     4.85 tokens per second)
llama_print_timings:        eval time =    9727.12 ms /    38 runs   (  255.98
```

```
ms per token,     3.91 tokens per second)
llama_print_timings:       total time =   27399.89 ms /   123 tokens
No. of rows:  62%| | 774/1258 [7:08:48<4:19:44, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.05 ms /    33 runs   (    0.30
ms per token,  3282.60 tokens per second)
llama_print_timings: prompt eval time =   18897.22 ms /    87 tokens (  217.21
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    9948.70 ms /    32 runs   (  310.90
ms per token,     3.22 tokens per second)
llama_print_timings:       total time =   28989.04 ms /   119 tokens
No. of rows:  62%| | 775/1258 [7:09:17<4:11:28, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.86 ms /    29 runs   (    0.27
ms per token,  3690.04 tokens per second)
llama_print_timings: prompt eval time =   18781.42 ms /    90 tokens (  208.68
ms per token,     4.79 tokens per second)
llama_print_timings:        eval time =    7435.35 ms /    28 runs   (  265.55
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   26332.83 ms /   118 tokens
No. of rows:  62%| | 776/1258 [7:09:44<3:59:07, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.11 ms /    40 runs   (    0.28
ms per token,  3600.36 tokens per second)
llama_print_timings: prompt eval time =   22415.28 ms /   111 tokens (  201.94
ms per token,     4.95 tokens per second)
llama_print_timings:        eval time =   10601.50 ms /    39 runs   (  271.83
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   33178.44 ms /   150 tokens
No. of rows:  62%| | 777/1258 [7:10:17<4:06:51, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.65 ms /    50 runs   (    0.27
ms per token,  3662.74 tokens per second)
llama_print_timings: prompt eval time =   24820.43 ms /   113 tokens (  219.65
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   14819.61 ms /    49 runs   (  302.44
ms per token,     3.31 tokens per second)
llama_print_timings:       total time =   39837.08 ms /   162 tokens
No. of rows:  62%| | 778/1258 [7:10:57<4:28:05, 33Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.15 ms /    34 runs   (    0.27
ms per token,  3716.66 tokens per second)
llama_print_timings: prompt eval time =   20848.42 ms /   100 tokens (  208.48
ms per token,     4.80 tokens per second)
llama_print_timings:        eval time =    8873.99 ms /    33 runs   (  268.91
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   29857.13 ms /   133 tokens
No. of rows:  62%| | 779/1258 [7:11:27<4:18:48, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.66 ms /    45 runs   (    0.26
ms per token,  3859.68 tokens per second)
llama_print_timings: prompt eval time =   21403.90 ms /   104 tokens (  205.81
ms per token,     4.86 tokens per second)
llama_print_timings:        eval time =   13414.60 ms /    44 runs   (  304.88
ms per token,     3.28 tokens per second)
llama_print_timings:       total time =   34990.97 ms /   148 tokens
No. of rows:  62%| | 780/1258 [7:12:02<4:24:25, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.88 ms /    50 runs   (    0.28
ms per token,  3601.79 tokens per second)
llama_print_timings: prompt eval time =   26701.40 ms /   131 tokens (  203.83
ms per token,     4.91 tokens per second)
llama_print_timings:        eval time =   13177.95 ms /    49 runs   (  268.94
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   40077.09 ms /   180 tokens
No. of rows:  62%| | 781/1258 [7:12:42<4:40:17, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.90 ms /    22 runs   (    0.27
ms per token,  3726.92 tokens per second)
llama_print_timings: prompt eval time =   16719.44 ms /    78 tokens (  214.35
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    5606.11 ms /    21 runs   (  266.96
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   22409.53 ms /    99 tokens
No. of rows:  62%| | 782/1258 [7:13:04<4:09:09, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.51 ms /    39 runs   (    0.27
ms per token,  3712.52 tokens per second)
```

```
llama_print_timings: prompt eval time =   24231.91 ms /   114 tokens (  212.56
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   10188.69 ms /    38 runs   (  268.12
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   34571.99 ms /   152 tokens
No. of rows:  62%| | 783/1258 [7:13:39<4:16:09, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.90 ms /    22 runs   (    0.27
ms per token,  3730.71 tokens per second)
llama_print_timings: prompt eval time =   19670.09 ms /    92 tokens (  213.81
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    5503.37 ms /    21 runs   (  262.07
ms per token,     3.82 tokens per second)
llama_print_timings:       total time =   25259.06 ms /   113 tokens
No. of rows:  62%| | 784/1258 [7:14:04<3:58:49, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.64 ms /    50 runs   (    0.27
ms per token,  3666.50 tokens per second)
llama_print_timings: prompt eval time =   26703.55 ms /   129 tokens (  207.00
ms per token,     4.83 tokens per second)
llama_print_timings:        eval time =   14845.70 ms /    49 runs   (  302.97
ms per token,     3.30 tokens per second)
llama_print_timings:       total time =   41744.05 ms /   178 tokens
No. of rows:  62%| | 785/1258 [7:14:46<4:25:34, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.63 ms /    50 runs   (    0.25
ms per token,  3959.46 tokens per second)
llama_print_timings: prompt eval time =   29416.27 ms /   139 tokens (  211.63
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   13380.48 ms /    49 runs   (  273.07
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   42990.35 ms /   188 tokens
No. of rows:  62%| | 786/1258 [7:15:29<4:46:58, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.69 ms /    32 runs   (    0.27
ms per token,  3683.67 tokens per second)
llama_print_timings: prompt eval time =   21179.88 ms /    97 tokens (  218.35
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =    8286.09 ms /    31 runs   (  267.29
ms per token,     3.74 tokens per second)
```

```
llama_print_timings:       total time =   29589.90 ms /   128 tokens
No. of rows:  63%| | 787/1258 [7:15:58<4:30:09, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.74 ms /    50 runs   (    0.27
ms per token,  3639.01 tokens per second)
llama_print_timings: prompt eval time =   24947.86 ms /   113 tokens (  220.78
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =   14934.35 ms /    49 runs   (  304.78
ms per token,     3.28 tokens per second)
llama_print_timings:       total time =   40087.77 ms /   162 tokens
No. of rows:  63%| | 788/1258 [7:16:38<4:42:56, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.87 ms /    30 runs   (    0.30
ms per token,  3381.81 tokens per second)
llama_print_timings: prompt eval time =   18807.97 ms /    88 tokens (  213.73
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    8219.23 ms /    29 runs   (  283.42
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   27150.69 ms /   117 tokens
No. of rows:  63%| | 789/1258 [7:17:06<4:21:19, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.91 ms /    48 runs   (    0.27
ms per token,  3718.91 tokens per second)
llama_print_timings: prompt eval time =   28663.67 ms /   135 tokens (  212.32
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   12591.05 ms /    47 runs   (  267.89
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   41443.96 ms /   182 tokens
No. of rows:  63%| | 790/1258 [7:17:47<4:39:31, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.26 ms /    32 runs   (    0.26
ms per token,  3876.44 tokens per second)
llama_print_timings: prompt eval time =   18601.42 ms /    86 tokens (  216.30
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    8225.77 ms /    31 runs   (  265.35
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   26951.44 ms /   117 tokens
No. of rows:  63%| | 791/1258 [7:18:14<4:18:12, 33Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.27 ms /    35 runs   (    0.26
ms per token,  3777.66 tokens per second)
llama_print_timings: prompt eval time =   22728.86 ms /   108 tokens (  210.45
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =    9148.80 ms /    34 runs   (  269.08
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   32016.13 ms /   142 tokens
No. of rows:  63%| | 792/1258 [7:18:46<4:14:57, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.88 ms /    50 runs   (    0.28
ms per token,  3601.27 tokens per second)
llama_print_timings: prompt eval time =   25185.71 ms /   121 tokens (  208.15
ms per token,     4.80 tokens per second)
llama_print_timings:        eval time =   13542.31 ms /    49 runs   (  276.37
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   38934.03 ms /   170 tokens
No. of rows:  63%| | 793/1258 [7:19:25<4:28:37, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.47 ms /    28 runs   (    0.27
ms per token,  3748.83 tokens per second)
llama_print_timings: prompt eval time =   18821.60 ms /    88 tokens (  213.88
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    7266.47 ms /    27 runs   (  269.13
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   26197.90 ms /   115 tokens
No. of rows:  63%| | 794/1258 [7:19:51<4:08:25, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.17 ms /    30 runs   (    0.27
ms per token,  3673.77 tokens per second)
llama_print_timings: prompt eval time =   19765.53 ms /    88 tokens (  224.61
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =    7941.67 ms /    29 runs   (  273.85
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   27829.77 ms /   117 tokens
No. of rows:  63%| | 795/1258 [7:20:19<3:57:58, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.81 ms /    33 runs   (    0.27
ms per token,  3744.47 tokens per second)
llama_print_timings: prompt eval time =   19748.38 ms /    93 tokens (  212.35
```

```
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =    8512.13 ms /    32 runs   (  266.00
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   28391.67 ms /   125 tokens
No. of rows:  63%| | 796/1258 [7:20:47<3:51:48, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.63 ms /    45 runs   (    0.28
ms per token,  3561.54 tokens per second)
llama_print_timings: prompt eval time =   19830.68 ms /    93 tokens (  213.23
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   12040.40 ms /    44 runs   (  273.65
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   32051.26 ms /   137 tokens
No. of rows:  63%| | 797/1258 [7:21:19<3:55:49, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.96 ms /    42 runs   (    0.33
ms per token,  3008.81 tokens per second)
llama_print_timings: prompt eval time =   24785.39 ms /   117 tokens (  211.84
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   11165.95 ms /    41 runs   (  272.34
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   36120.93 ms /   158 tokens
No. of rows:  63%| | 798/1258 [7:21:56<4:07:48, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.84 ms /    48 runs   (    0.27
ms per token,  3738.90 tokens per second)
llama_print_timings: prompt eval time =   22180.27 ms /    98 tokens (  226.33
ms per token,     4.42 tokens per second)
llama_print_timings:        eval time =   14727.03 ms /    47 runs   (  313.34
ms per token,     3.19 tokens per second)
llama_print_timings:       total time =   37100.39 ms /   145 tokens
No. of rows:  64%| | 799/1258 [7:22:33<4:18:15, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.59 ms /    25 runs   (    0.26
ms per token,  3792.48 tokens per second)
llama_print_timings: prompt eval time =   20016.06 ms /    95 tokens (  210.70
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =    6438.19 ms /    24 runs   (  268.26
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   26552.25 ms /   119 tokens
```

```
No. of rows:  64%| | 800/1258 [7:22:59<4:01:12, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.07 ms /    33 runs   (    0.27
ms per token,  3639.57 tokens per second)
llama_print_timings: prompt eval time =   21186.87 ms /    92 tokens (  230.29
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =    8707.50 ms /    32 runs   (  272.11
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   30025.35 ms /   124 tokens
No. of rows:  64%| | 801/1258 [7:23:29<3:57:06, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.15 ms /    50 runs   (    0.26
ms per token,  3802.86 tokens per second)
llama_print_timings: prompt eval time =   26611.33 ms /   126 tokens (  211.20
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   13071.58 ms /    49 runs   (  266.77
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   39878.16 ms /   175 tokens
No. of rows:  64%| | 802/1258 [7:24:09<4:16:32, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.61 ms /    31 runs   (    0.28
ms per token,  3600.46 tokens per second)
llama_print_timings: prompt eval time =   19156.56 ms /    90 tokens (  212.85
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =    8006.56 ms /    30 runs   (  266.89
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   27283.83 ms /   120 tokens
No. of rows:  64%| | 803/1258 [7:24:36<4:01:15, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.00 ms /    29 runs   (    0.28
ms per token,  3626.36 tokens per second)
llama_print_timings: prompt eval time =   20888.16 ms /    96 tokens (  217.58
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    7593.97 ms /    28 runs   (  271.21
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   28593.82 ms /   124 tokens
No. of rows:  64%| | 804/1258 [7:25:05<3:53:27, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      11.41 ms /    42 runs   (    0.27
ms per token,  3682.27 tokens per second)
llama_print_timings: prompt eval time =   22584.28 ms /    98 tokens (  230.45
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =   11083.11 ms /    41 runs   (  270.32
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   33834.54 ms /   139 tokens
No. of rows:  64%| | 805/1258 [7:25:39<3:59:42, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.69 ms /    38 runs   (    0.25
ms per token,  3923.19 tokens per second)
llama_print_timings: prompt eval time =   21720.76 ms /   101 tokens (  215.06
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9947.15 ms /    37 runs   (  268.84
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   31811.71 ms /   138 tokens
No. of rows:  64%| | 806/1258 [7:26:11<3:59:20, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.81 ms /    37 runs   (    0.27
ms per token,  3770.89 tokens per second)
llama_print_timings: prompt eval time =   23709.14 ms /   107 tokens (  221.58
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    9733.07 ms /    36 runs   (  270.36
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   33587.37 ms /   143 tokens
No. of rows:  64%| | 807/1258 [7:26:44<4:02:54, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.97 ms /    50 runs   (    0.26
ms per token,  3854.75 tokens per second)
llama_print_timings: prompt eval time =   27289.72 ms /   122 tokens (  223.69
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   13681.54 ms /    49 runs   (  279.22
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   41171.58 ms /   171 tokens
No. of rows:  64%| | 808/1258 [7:27:26<4:22:17, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.69 ms /    33 runs   (    0.26
ms per token,  3797.91 tokens per second)
llama_print_timings: prompt eval time =   20622.64 ms /    94 tokens (  219.39
ms per token,     4.56 tokens per second)
```

```
llama_print_timings:        eval time =    8793.01 ms /    32 runs   (  274.78
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   29542.76 ms /   126 tokens
No. of rows:  64%| | 809/1258 [7:27:55<4:09:33, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.74 ms /    49 runs   (    0.26
ms per token,  3845.55 tokens per second)
llama_print_timings: prompt eval time =   24435.70 ms /   115 tokens (  212.48
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   12787.30 ms /    48 runs   (  266.40
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   37421.28 ms /   163 tokens
No. of rows:  64%| | 810/1258 [7:28:33<4:18:07, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.94 ms /    23 runs   (    0.26
ms per token,  3869.45 tokens per second)
llama_print_timings: prompt eval time =   21545.66 ms /    99 tokens (  217.63
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    6044.56 ms /    22 runs   (  274.75
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   27678.13 ms /   121 tokens
No. of rows:  64%| | 811/1258 [7:29:00<4:02:10, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.81 ms /    33 runs   (    0.27
ms per token,  3746.17 tokens per second)
llama_print_timings: prompt eval time =   26409.93 ms /   117 tokens (  225.73
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =    8793.78 ms /    32 runs   (  274.81
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   35336.45 ms /   149 tokens
No. of rows:  65%| | 812/1258 [7:29:36<4:07:57, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.19 ms /    40 runs   (    0.28
ms per token,  3573.66 tokens per second)
llama_print_timings: prompt eval time =   19910.54 ms /    94 tokens (  211.81
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   10613.76 ms /    39 runs   (  272.15
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   30685.54 ms /   133 tokens
No. of rows:  65%| | 813/1258 [7:30:06<4:01:27, 32Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.64 ms /    50 runs   (    0.27
ms per token,  3665.42 tokens per second)
llama_print_timings: prompt eval time =   25821.89 ms /   123 tokens (  209.93
ms per token,     4.76 tokens per second)
llama_print_timings:        eval time =   13165.04 ms /    49 runs   (  268.67
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   39186.60 ms /   172 tokens
No. of rows:  65%| | 814/1258 [7:30:45<4:15:40, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.78 ms /    50 runs   (    0.28
ms per token,  3628.97 tokens per second)
llama_print_timings: prompt eval time =   22466.40 ms /   100 tokens (  224.66
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   13709.61 ms /    49 runs   (  279.79
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   36379.84 ms /   149 tokens
No. of rows:  65%| | 815/1258 [7:31:22<4:19:08, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.36 ms /    31 runs   (    0.27
ms per token,  3708.58 tokens per second)
llama_print_timings: prompt eval time =   22697.91 ms /   105 tokens (  216.17
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    8237.24 ms /    30 runs   (  274.57
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   31060.86 ms /   135 tokens
No. of rows:  65%| | 816/1258 [7:31:53<4:09:40, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.61 ms /    50 runs   (    0.27
ms per token,  3675.12 tokens per second)
llama_print_timings: prompt eval time =   27915.85 ms /   131 tokens (  213.10
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   15098.87 ms /    49 runs   (  308.14
ms per token,     3.25 tokens per second)
llama_print_timings:       total time =   43216.38 ms /   180 tokens
No. of rows:  65%| | 817/1258 [7:32:36<4:29:41, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.89 ms /    26 runs   (    0.26
```

```
ms per token,  3775.78 tokens per second)
llama_print_timings: prompt eval time =   17626.21 ms /    81 tokens (  217.61
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    6651.92 ms /    25 runs   (  266.08
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   24382.95 ms /   106 tokens
No. of rows:  65%| | 818/1258 [7:33:00<4:01:59, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.07 ms /    29 runs   (    0.28
ms per token,  3591.78 tokens per second)
llama_print_timings: prompt eval time =   20917.88 ms /    91 tokens (  229.87
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =    7619.22 ms /    28 runs   (  272.12
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   28652.14 ms /   119 tokens
No. of rows:  65%| | 819/1258 [7:33:29<3:51:56, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.22 ms /    50 runs   (    0.26
ms per token,  3783.29 tokens per second)
llama_print_timings: prompt eval time =   21974.61 ms /   102 tokens (  215.44
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   13248.03 ms /    49 runs   (  270.37
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   35420.75 ms /   151 tokens
No. of rows:  65%| | 820/1258 [7:34:05<3:59:34, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.62 ms /    36 runs   (    0.30
ms per token,  3388.55 tokens per second)
llama_print_timings: prompt eval time =   26668.85 ms /   127 tokens (  209.99
ms per token,     4.76 tokens per second)
llama_print_timings:        eval time =    9505.72 ms /    35 runs   (  271.59
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   36321.05 ms /   162 tokens
No. of rows:  65%| | 821/1258 [7:34:41<4:06:41, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.18 ms /    35 runs   (    0.26
ms per token,  3812.64 tokens per second)
llama_print_timings: prompt eval time =   20789.90 ms /    92 tokens (  225.98
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =    9157.04 ms /    34 runs   (  269.32
```

```
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   30084.59 ms /   126 tokens
No. of rows:  65%| | 822/1258 [7:35:11<3:57:54, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.02 ms /    39 runs   (    0.28
ms per token,  3539.66 tokens per second)
llama_print_timings: prompt eval time =   22090.00 ms /   100 tokens (  220.90
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =   10401.39 ms /    38 runs   (  273.72
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   32646.99 ms /   138 tokens
No. of rows:  65%| | 823/1258 [7:35:44<3:57:10, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.04 ms /    50 runs   (    0.28
ms per token,  3561.76 tokens per second)
llama_print_timings: prompt eval time =   25106.35 ms /   120 tokens (  209.22
ms per token,     4.78 tokens per second)
llama_print_timings:        eval time =   13000.98 ms /    49 runs   (  265.33
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   38307.03 ms /   169 tokens
No. of rows:  66%| | 824/1258 [7:36:22<4:08:47, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.45 ms /    40 runs   (    0.29
ms per token,  3492.53 tokens per second)
llama_print_timings: prompt eval time =   23683.85 ms /   108 tokens (  219.29
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   11044.69 ms /    39 runs   (  283.20
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   34890.44 ms /   147 tokens
No. of rows:  66%| | 825/1258 [7:36:57<4:09:18, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.03 ms /    33 runs   (    0.30
ms per token,  3290.46 tokens per second)
llama_print_timings: prompt eval time =   17349.22 ms /    81 tokens (  214.19
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    9159.02 ms /    32 runs   (  286.22
ms per token,     3.49 tokens per second)
llama_print_timings:       total time =   26649.58 ms /   113 tokens
No. of rows:  66%| | 826/1258 [7:37:24<3:51:41, 32Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.38 ms /    45 runs   (    0.28
ms per token,  3634.60 tokens per second)
llama_print_timings: prompt eval time =   24173.12 ms /   114 tokens (  212.04
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   11981.07 ms /    44 runs   (  272.30
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   36332.22 ms /   158 tokens
No. of rows:  66%| | 827/1258 [7:38:00<4:00:05, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.25 ms /    35 runs   (    0.26
ms per token,  3782.97 tokens per second)
llama_print_timings: prompt eval time =   19114.22 ms /    90 tokens (  212.38
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   10695.44 ms /    34 runs   (  314.57
ms per token,     3.18 tokens per second)
llama_print_timings:       total time =   29947.08 ms /   124 tokens
No. of rows:  66%| | 828/1258 [7:38:30<3:52:05, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.54 ms /    50 runs   (    0.27
ms per token,  3691.40 tokens per second)
llama_print_timings: prompt eval time =   27405.22 ms /   130 tokens (  210.81
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =   13139.87 ms /    49 runs   (  268.16
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   40742.47 ms /   179 tokens
No. of rows:  66%| | 829/1258 [7:39:11<4:09:29, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.92 ms /    50 runs   (    0.28
ms per token,  3593.24 tokens per second)
llama_print_timings: prompt eval time =   33417.09 ms /   152 tokens (  219.85
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   13283.21 ms /    49 runs   (  271.09
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   46905.53 ms /   201 tokens
No. of rows:  66%| | 830/1258 [7:39:57<4:34:38, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.40 ms /    31 runs   (    0.27
ms per token,  3688.72 tokens per second)
```

```
llama_print_timings: prompt eval time =   18505.18 ms /    86 tokens (  215.18
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    8098.39 ms /    30 runs   (  269.95
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   26724.88 ms /   116 tokens
No. of rows:  66%| | 831/1258 [7:40:24<4:08:52, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.82 ms /    29 runs   (    0.27
ms per token,  3709.39 tokens per second)
llama_print_timings: prompt eval time =   23405.05 ms /   111 tokens (  210.86
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =    7464.50 ms /    28 runs   (  266.59
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   30985.04 ms /   139 tokens
No. of rows:  66%| | 832/1258 [7:40:55<3:59:48, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.18 ms /    29 runs   (    0.28
ms per token,  3544.37 tokens per second)
llama_print_timings: prompt eval time =   21983.37 ms /    99 tokens (  222.05
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =    8050.58 ms /    28 runs   (  287.52
ms per token,     3.48 tokens per second)
llama_print_timings:       total time =   30151.47 ms /   127 tokens
No. of rows:  66%| | 833/1258 [7:41:25<3:51:32, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.78 ms /    29 runs   (    0.27
ms per token,  3729.42 tokens per second)
llama_print_timings: prompt eval time =   19925.01 ms /    95 tokens (  209.74
ms per token,     4.77 tokens per second)
llama_print_timings:        eval time =    7594.08 ms /    28 runs   (  271.22
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   27632.68 ms /   123 tokens
No. of rows:  66%| | 834/1258 [7:41:53<3:40:17, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.57 ms /    24 runs   (    0.27
ms per token,  3653.52 tokens per second)
llama_print_timings: prompt eval time =   18547.92 ms /    87 tokens (  213.19
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    6100.49 ms /    23 runs   (  265.24
ms per token,     3.77 tokens per second)
```

```
llama_print_timings:       total time =   24743.81 ms /   110 tokens
No. of rows:  66%| | 835/1258 [7:42:18<3:26:12, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.57 ms /    39 runs   (    0.27
ms per token,  3690.39 tokens per second)
llama_print_timings: prompt eval time =   21524.22 ms /   102 tokens (  211.02
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =   10270.59 ms /    38 runs   (  270.28
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   31951.76 ms /   140 tokens
No. of rows:  66%| | 836/1258 [7:42:50<3:31:26, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.13 ms /    50 runs   (    0.28
ms per token,  3538.32 tokens per second)
llama_print_timings: prompt eval time =   20681.36 ms /    91 tokens (  227.27
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =   15178.69 ms /    49 runs   (  309.77
ms per token,     3.23 tokens per second)
llama_print_timings:       total time =   36068.26 ms /   140 tokens
No. of rows:  67%| | 837/1258 [7:43:26<3:43:35, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.37 ms /    28 runs   (    0.26
ms per token,  3798.16 tokens per second)
llama_print_timings: prompt eval time =   19149.09 ms /    89 tokens (  215.16
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    7436.61 ms /    27 runs   (  275.43
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   26693.01 ms /   116 tokens
No. of rows:  67%| | 838/1258 [7:43:52<3:32:11, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.53 ms /    50 runs   (    0.27
ms per token,  3696.86 tokens per second)
llama_print_timings: prompt eval time =   22260.59 ms /   105 tokens (  212.01
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   13177.92 ms /    49 runs   (  268.94
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   35635.60 ms /   154 tokens
No. of rows:  67%| | 839/1258 [7:44:28<3:42:50, 31Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.29 ms /    26 runs   (    0.28
ms per token,  3565.06 tokens per second)
llama_print_timings: prompt eval time =   19708.07 ms /    92 tokens (  214.22
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    6711.92 ms /    25 runs   (  268.48
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   26521.56 ms /   117 tokens
No. of rows:  67%| | 840/1258 [7:44:55<3:31:05, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.17 ms /    40 runs   (    0.28
ms per token,  3581.02 tokens per second)
llama_print_timings: prompt eval time =   24378.03 ms /   115 tokens (  211.98
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   11427.29 ms /    39 runs   (  293.01
ms per token,     3.41 tokens per second)
llama_print_timings:       total time =   35969.21 ms /   154 tokens
No. of rows:  67%| | 841/1258 [7:45:31<3:42:25, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.54 ms /    50 runs   (    0.27
ms per token,  3692.76 tokens per second)
llama_print_timings: prompt eval time =   24767.33 ms /   115 tokens (  215.37
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   13073.64 ms /    49 runs   (  266.81
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   38035.70 ms /   164 tokens
No. of rows:  67%| | 842/1258 [7:46:09<3:54:26, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.62 ms /    31 runs   (    0.28
ms per token,  3597.12 tokens per second)
llama_print_timings: prompt eval time =   20066.92 ms /    94 tokens (  213.48
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    8163.81 ms /    30 runs   (  272.13
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   28350.27 ms /   124 tokens
No. of rows:  67%| | 843/1258 [7:46:37<3:42:32, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.19 ms /    43 runs   (    0.28
ms per token,  3528.35 tokens per second)
llama_print_timings: prompt eval time =   21774.14 ms /   103 tokens (  211.40
```

```
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   11569.96 ms /    42 runs   (  275.48
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   33518.60 ms /   145 tokens
No. of rows:  67%| | 844/1258 [7:47:11<3:44:49, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.31 ms /    38 runs   (    0.27
ms per token,  3685.03 tokens per second)
llama_print_timings: prompt eval time =   21459.91 ms /    98 tokens (  218.98
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =   10032.83 ms /    37 runs   (  271.16
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   31642.43 ms /   135 tokens
No. of rows:  67%| | 845/1258 [7:47:42<3:42:19, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.34 ms /    41 runs   (    0.28
ms per token,  3616.80 tokens per second)
llama_print_timings: prompt eval time =   25474.77 ms /   122 tokens (  208.81
ms per token,     4.79 tokens per second)
llama_print_timings:        eval time =   10738.80 ms /    40 runs   (  268.47
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   36375.17 ms /   162 tokens
No. of rows:  67%| | 846/1258 [7:48:19<3:50:13, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.24 ms /    34 runs   (    0.27
ms per token,  3681.25 tokens per second)
llama_print_timings: prompt eval time =   18661.96 ms /    87 tokens (  214.51
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    8767.99 ms /    33 runs   (  265.70
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   27563.59 ms /   120 tokens
No. of rows:  67%| | 847/1258 [7:48:46<3:37:23, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.13 ms /    37 runs   (    0.27
ms per token,  3652.88 tokens per second)
llama_print_timings: prompt eval time =   19876.50 ms /    93 tokens (  213.73
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    9767.05 ms /    36 runs   (  271.31
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   29788.32 ms /   129 tokens
```

```
No. of rows:  67%| | 848/1258 [7:49:16<3:32:53, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.66 ms /    43 runs   (    0.27
ms per token,  3687.82 tokens per second)
llama_print_timings: prompt eval time =   19910.67 ms /    90 tokens (  221.23
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =   11204.25 ms /    42 runs   (  266.77
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   31284.09 ms /   132 tokens
No. of rows:  67%| | 849/1258 [7:49:47<3:32:40, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.65 ms /    50 runs   (    0.27
ms per token,  3663.54 tokens per second)
llama_print_timings: prompt eval time =   23658.00 ms /   112 tokens (  211.23
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   13283.18 ms /    49 runs   (  271.09
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   37146.99 ms /   161 tokens
No. of rows:  68%| | 850/1258 [7:50:24<3:44:17, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.44 ms /    50 runs   (    0.29
ms per token,  3463.32 tokens per second)
llama_print_timings: prompt eval time =   26612.46 ms /   121 tokens (  219.94
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   14965.28 ms /    49 runs   (  305.41
ms per token,     3.27 tokens per second)
llama_print_timings:       total time =   41775.48 ms /   170 tokens
No. of rows:  68%| | 851/1258 [7:51:06<4:01:39, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.59 ms /    50 runs   (    0.27
ms per token,  3679.45 tokens per second)
llama_print_timings: prompt eval time =   20621.91 ms /    95 tokens (  217.07
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   13405.15 ms /    49 runs   (  273.57
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   34228.84 ms /   144 tokens
No. of rows:  68%| | 852/1258 [7:51:40<3:58:13, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      12.44 ms /    45 runs   (    0.28
ms per token,  3616.20 tokens per second)
llama_print_timings: prompt eval time =   18804.06 ms /    86 tokens (  218.65
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =   11984.94 ms /    44 runs   (  272.39
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   30969.43 ms /   130 tokens
No. of rows:  68%| | 853/1258 [7:52:11<3:49:05, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.18 ms /    33 runs   (    0.28
ms per token,  3596.34 tokens per second)
llama_print_timings: prompt eval time =   18809.98 ms /    89 tokens (  211.35
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =    8539.36 ms /    32 runs   (  266.85
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   27478.95 ms /   121 tokens
No. of rows:  68%| | 854/1258 [7:52:39<3:35:29, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.91 ms /    31 runs   (    0.26
ms per token,  3917.11 tokens per second)
llama_print_timings: prompt eval time =   20366.08 ms /    96 tokens (  212.15
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =    8005.76 ms /    30 runs   (  266.86
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   28495.77 ms /   126 tokens
No. of rows:  68%| | 855/1258 [7:53:07<3:27:53, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.32 ms /    50 runs   (    0.29
ms per token,  3491.62 tokens per second)
llama_print_timings: prompt eval time =   20231.17 ms /    94 tokens (  215.23
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13572.37 ms /    49 runs   (  276.99
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   34004.82 ms /   143 tokens
No. of rows:  68%| | 856/1258 [7:53:41<3:33:32, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.83 ms /    28 runs   (    0.28
ms per token,  3577.82 tokens per second)
llama_print_timings: prompt eval time =   17262.54 ms /    81 tokens (  213.12
ms per token,     4.69 tokens per second)
```

```
llama_print_timings:        eval time =    7173.18 ms /    27 runs   (  265.67
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   24547.49 ms /   108 tokens
No. of rows:  68%| | 857/1258 [7:54:06<3:18:19, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.44 ms /    23 runs   (    0.28
ms per token,  3574.20 tokens per second)
llama_print_timings: prompt eval time =   19632.15 ms /    92 tokens (  213.39
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    5963.24 ms /    22 runs   (  271.06
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   25689.08 ms /   114 tokens
No. of rows:  68%| | 858/1258 [7:54:32<3:09:52, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.79 ms /    28 runs   (    0.28
ms per token,  3595.27 tokens per second)
llama_print_timings: prompt eval time =   18815.36 ms /    87 tokens (  216.27
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7339.97 ms /    27 runs   (  271.85
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   26266.33 ms /   114 tokens
No. of rows:  68%| | 859/1258 [7:54:58<3:05:00, 27Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.79 ms /    32 runs   (    0.27
ms per token,  3640.91 tokens per second)
llama_print_timings: prompt eval time =   18858.15 ms /    88 tokens (  214.30
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    8534.14 ms /    31 runs   (  275.29
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   27519.70 ms /   119 tokens
No. of rows:  68%| | 860/1258 [7:55:25<3:03:56, 27Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.46 ms /    50 runs   (    0.27
ms per token,  3715.26 tokens per second)
llama_print_timings: prompt eval time =   23575.76 ms /   110 tokens (  214.33
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13252.22 ms /    49 runs   (  270.45
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   37025.46 ms /   159 tokens
No. of rows:  68%| | 861/1258 [7:56:03<3:21:58, 30Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.93 ms /    35 runs   (    0.28
ms per token,  3523.25 tokens per second)
llama_print_timings: prompt eval time =   22170.33 ms /   105 tokens (  211.15
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =    9197.68 ms /    34 runs   (  270.52
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   31510.23 ms /   139 tokens
No. of rows:  69%| | 862/1258 [7:56:34<3:23:25, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.32 ms /    39 runs   (    0.29
ms per token,  3446.45 tokens per second)
llama_print_timings: prompt eval time =   21855.90 ms /    98 tokens (  223.02
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =   10276.88 ms /    38 runs   (  270.44
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   32289.32 ms /   136 tokens
No. of rows:  69%| | 863/1258 [7:57:06<3:25:49, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.30 ms /    24 runs   (    0.26
ms per token,  3808.92 tokens per second)
llama_print_timings: prompt eval time =   17062.29 ms /    78 tokens (  218.75
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    6267.73 ms /    23 runs   (  272.51
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   23422.07 ms /   101 tokens
No. of rows:  69%| | 864/1258 [7:57:30<3:09:52, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.75 ms /    50 runs   (    0.27
ms per token,  3636.63 tokens per second)
llama_print_timings: prompt eval time =   22151.18 ms /    98 tokens (  226.03
ms per token,     4.42 tokens per second)
llama_print_timings:        eval time =   13515.66 ms /    49 runs   (  275.83
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   35868.12 ms /   147 tokens
No. of rows:  69%| | 865/1258 [7:58:06<3:23:03, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.22 ms /    33 runs   (    0.25
```

```
ms per token,  4015.09 tokens per second)
llama_print_timings: prompt eval time =   22363.79 ms /   104 tokens (  215.04
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    8691.06 ms /    32 runs   (  271.60
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   31182.76 ms /   136 tokens
No. of rows:  69%| | 866/1258 [7:58:37<3:22:54, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.01 ms /    46 runs   (    0.28
ms per token,  3537.10 tokens per second)
llama_print_timings: prompt eval time =   19652.05 ms /    92 tokens (  213.61
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   12475.34 ms /    45 runs   (  277.23
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   32319.39 ms /   137 tokens
No. of rows:  69%| | 867/1258 [7:59:09<3:24:52, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.65 ms /    50 runs   (    0.27
ms per token,  3662.20 tokens per second)
llama_print_timings: prompt eval time =   20924.79 ms /    91 tokens (  229.94
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =   13572.78 ms /    49 runs   (  277.00
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   34696.45 ms /   140 tokens
No. of rows:  69%| | 868/1258 [7:59:44<3:30:42, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.67 ms /    49 runs   (    0.28
ms per token,  3584.75 tokens per second)
llama_print_timings: prompt eval time =   22133.46 ms /   103 tokens (  214.89
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13117.55 ms /    48 runs   (  273.28
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   35449.24 ms /   151 tokens
No. of rows:  69%| | 869/1258 [8:00:19<3:36:04, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.01 ms /    41 runs   (    0.27
ms per token,  3723.55 tokens per second)
llama_print_timings: prompt eval time =   22370.99 ms /   103 tokens (  217.19
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =   10804.03 ms /    40 runs   (  270.10
```

```
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   33334.97 ms /   143 tokens
No. of rows:  69%| | 870/1258 [8:00:53<3:35:34, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.58 ms /    50 runs   (    0.27
ms per token,  3682.97 tokens per second)
llama_print_timings: prompt eval time =   23745.62 ms /   112 tokens (  212.01
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   13485.31 ms /    49 runs   (  275.21
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   37429.33 ms /   161 tokens
No. of rows:  69%| | 871/1258 [8:01:30<3:42:57, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.48 ms /    50 runs   (    0.27
ms per token,  3707.82 tokens per second)
llama_print_timings: prompt eval time =   25034.41 ms /   112 tokens (  223.52
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   13238.79 ms /    49 runs   (  270.18
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   38473.80 ms /   161 tokens
No. of rows:  69%| | 872/1258 [8:02:09<3:49:54, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.05 ms /    34 runs   (    0.27
ms per token,  3757.32 tokens per second)
llama_print_timings: prompt eval time =   23052.70 ms /   109 tokens (  211.49
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =    8851.86 ms /    33 runs   (  268.24
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   32036.65 ms /   142 tokens
No. of rows:  69%| | 873/1258 [8:02:41<3:42:12, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.09 ms /    38 runs   (    0.27
ms per token,  3767.60 tokens per second)
llama_print_timings: prompt eval time =   19648.22 ms /    86 tokens (  228.47
ms per token,     4.38 tokens per second)
llama_print_timings:        eval time =   11639.28 ms /    37 runs   (  314.58
ms per token,     3.18 tokens per second)
llama_print_timings:       total time =   31440.95 ms /   123 tokens
No. of rows:  69%| | 874/1258 [8:03:12<3:35:32, 33Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.37 ms /    39 runs   (    0.27
ms per token,  3762.30 tokens per second)
llama_print_timings: prompt eval time =   21801.24 ms /   102 tokens (  213.74
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   10518.78 ms /    38 runs   (  276.81
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   32479.55 ms /   140 tokens
No. of rows:  70%| | 875/1258 [8:03:45<3:32:41, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.96 ms /    38 runs   (    0.26
ms per token,  3815.26 tokens per second)
llama_print_timings: prompt eval time =   21202.16 ms /    98 tokens (  216.35
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    9761.44 ms /    37 runs   (  263.82
ms per token,     3.79 tokens per second)
llama_print_timings:       total time =   31110.06 ms /   135 tokens
No. of rows:  70%| | 876/1258 [8:04:16<3:27:56, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.88 ms /    37 runs   (    0.27
ms per token,  3745.32 tokens per second)
llama_print_timings: prompt eval time =   21604.27 ms /   103 tokens (  209.75
ms per token,     4.77 tokens per second)
llama_print_timings:        eval time =    9485.08 ms /    36 runs   (  263.47
ms per token,     3.80 tokens per second)
llama_print_timings:       total time =   31231.69 ms /   139 tokens
No. of rows:  70%| | 877/1258 [8:04:47<3:24:40, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.50 ms /    50 runs   (    0.27
ms per token,  3705.08 tokens per second)
llama_print_timings: prompt eval time =   19273.29 ms /    85 tokens (  226.74
ms per token,     4.41 tokens per second)
llama_print_timings:        eval time =   13158.05 ms /    49 runs   (  268.53
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   32630.94 ms /   134 tokens
No. of rows:  70%| | 878/1258 [8:05:20<3:24:55, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.15 ms /    27 runs   (    0.30
ms per token,  3314.10 tokens per second)
```

```
llama_print_timings: prompt eval time =   22205.15 ms /   102 tokens (  217.70
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    7113.19 ms /    26 runs   (  273.58
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   29429.42 ms /   128 tokens
No. of rows:  70%| | 879/1258 [8:05:49<3:18:50, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.83 ms /    43 runs   (    0.28
ms per token,  3633.91 tokens per second)
llama_print_timings: prompt eval time =   24677.80 ms /   117 tokens (  210.92
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =   11264.83 ms /    42 runs   (  268.21
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   36117.18 ms /   159 tokens
No. of rows:  70%| | 880/1258 [8:06:25<3:27:06, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.77 ms /    40 runs   (    0.27
ms per token,  3714.37 tokens per second)
llama_print_timings: prompt eval time =   23936.18 ms /   113 tokens (  211.82
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   10929.42 ms /    39 runs   (  280.24
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   35024.81 ms /   152 tokens
No. of rows:  70%| | 881/1258 [8:07:00<3:30:37, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.18 ms /    34 runs   (    0.27
ms per token,  3703.70 tokens per second)
llama_print_timings: prompt eval time =   19673.44 ms /    93 tokens (  211.54
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =    9086.37 ms /    33 runs   (  275.34
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   28900.49 ms /   126 tokens
No. of rows:  70%| | 882/1258 [8:07:29<3:21:23, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.79 ms /    34 runs   (    0.26
ms per token,  3866.71 tokens per second)
llama_print_timings: prompt eval time =   20571.91 ms /    98 tokens (  209.92
ms per token,     4.76 tokens per second)
llama_print_timings:        eval time =    8854.00 ms /    33 runs   (  268.30
ms per token,     3.73 tokens per second)
```

```
llama_print_timings:       total time =   29557.79 ms /   131 tokens
No. of rows:  70%| | 883/1258 [8:07:59<3:16:01, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.99 ms /    37 runs   (    0.27
ms per token,  3702.22 tokens per second)
llama_print_timings: prompt eval time =   23427.22 ms /   111 tokens (  211.06
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =    9657.53 ms /    36 runs   (  268.26
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   33231.13 ms /   147 tokens
No. of rows:  70%| | 884/1258 [8:08:32<3:19:00, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.98 ms /    50 runs   (    0.28
ms per token,  3576.79 tokens per second)
llama_print_timings: prompt eval time =   21684.00 ms /   103 tokens (  210.52
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =   13141.34 ms /    49 runs   (  268.19
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   35023.02 ms /   152 tokens
No. of rows:  70%| | 885/1258 [8:09:07<3:24:16, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.15 ms /    50 runs   (    0.26
ms per token,  3802.86 tokens per second)
llama_print_timings: prompt eval time =   33953.40 ms /   156 tokens (  217.65
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =   13672.03 ms /    49 runs   (  279.02
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   47827.87 ms /   205 tokens
No. of rows:  70%| | 886/1258 [8:09:55<3:51:35, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.66 ms /    50 runs   (    0.27
ms per token,  3660.59 tokens per second)
llama_print_timings: prompt eval time =   30624.28 ms /   143 tokens (  214.16
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13348.92 ms /    49 runs   (  272.43
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   44177.52 ms /   192 tokens
No. of rows:  71%| | 887/1258 [8:10:39<4:03:38, 39Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.88 ms /    50 runs   (    0.28
ms per token,  3602.57 tokens per second)
llama_print_timings: prompt eval time =   26431.80 ms /   125 tokens (  211.45
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   13129.19 ms /    49 runs   (  267.94
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   39758.50 ms /   174 tokens
No. of rows:  71%| | 888/1258 [8:11:19<4:03:37, 39Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.95 ms /    22 runs   (    0.27
ms per token,  3698.10 tokens per second)
llama_print_timings: prompt eval time =   18534.22 ms /    79 tokens (  234.61
ms per token,     4.26 tokens per second)
llama_print_timings:        eval time =    5643.46 ms /    21 runs   (  268.74
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   24263.90 ms /   100 tokens
No. of rows:  71%| | 889/1258 [8:11:43<3:34:52, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.81 ms /    29 runs   (    0.27
ms per token,  3714.62 tokens per second)
llama_print_timings: prompt eval time =   16979.89 ms /    78 tokens (  217.69
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    7509.55 ms /    28 runs   (  268.20
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   24600.89 ms /   106 tokens
No. of rows:  71%| | 890/1258 [8:12:08<3:15:17, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.69 ms /    28 runs   (    0.27
ms per token,  3642.04 tokens per second)
llama_print_timings: prompt eval time =   19090.55 ms /    88 tokens (  216.94
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    7362.82 ms /    27 runs   (  272.70
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   26563.45 ms /   115 tokens
No. of rows:  71%| | 891/1258 [8:12:34<3:05:04, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.14 ms /    49 runs   (    0.27
ms per token,  3729.64 tokens per second)
llama_print_timings: prompt eval time =   23273.92 ms /   105 tokens (  221.66
```

```
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =   12777.93 ms /    48 runs   (  266.21
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   36247.39 ms /   153 tokens
No. of rows:  71%| | 892/1258 [8:13:10<3:15:33, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.34 ms /    34 runs   (    0.27
ms per token,  3639.09 tokens per second)
llama_print_timings: prompt eval time =   20337.57 ms /    94 tokens (  216.36
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    8789.54 ms /    33 runs   (  266.35
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   29261.39 ms /   127 tokens
No. of rows:  71%| | 893/1258 [8:13:40<3:09:55, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.23 ms /    25 runs   (    0.29
ms per token,  3456.38 tokens per second)
llama_print_timings: prompt eval time =   19427.05 ms /    91 tokens (  213.48
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    6585.68 ms /    24 runs   (  274.40
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   26118.67 ms /   115 tokens
No. of rows:  71%| | 894/1258 [8:14:06<3:00:08, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.42 ms /    46 runs   (    0.27
ms per token,  3704.60 tokens per second)
llama_print_timings: prompt eval time =   24961.75 ms /   119 tokens (  209.76
ms per token,     4.77 tokens per second)
llama_print_timings:        eval time =   12162.77 ms /    45 runs   (  270.28
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   37306.27 ms /   164 tokens
No. of rows:  71%| | 895/1258 [8:14:43<3:13:28, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.88 ms /    43 runs   (    0.28
ms per token,  3619.53 tokens per second)
llama_print_timings: prompt eval time =   25140.61 ms /   118 tokens (  213.06
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   11519.83 ms /    42 runs   (  274.28
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   36835.56 ms /   160 tokens
```

```
No. of rows:  71%| | 896/1258 [8:15:20<3:21:42, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.54 ms /    32 runs   (    0.27
ms per token,  3745.32 tokens per second)
llama_print_timings: prompt eval time =   19483.61 ms /    88 tokens (  221.40
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    8232.22 ms /    31 runs   (  265.56
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   27841.16 ms /   119 tokens
No. of rows:  71%| | 897/1258 [8:15:48<3:11:05, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.54 ms /    50 runs   (    0.27
ms per token,  3693.85 tokens per second)
llama_print_timings: prompt eval time =   19517.45 ms /    90 tokens (  216.86
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   13216.35 ms /    49 runs   (  269.72
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   32938.66 ms /   139 tokens
No. of rows:  71%| | 898/1258 [8:16:21<3:12:41, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.74 ms /    26 runs   (    0.26
ms per token,  3859.28 tokens per second)
llama_print_timings: prompt eval time =   18618.30 ms /    85 tokens (  219.04
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    6775.44 ms /    25 runs   (  271.02
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   25498.49 ms /   110 tokens
No. of rows:  71%| | 899/1258 [8:16:46<3:00:18, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.39 ms /    31 runs   (    0.27
ms per token,  3693.11 tokens per second)
llama_print_timings: prompt eval time =   20994.10 ms /   100 tokens (  209.94
ms per token,     4.76 tokens per second)
llama_print_timings:        eval time =    8076.28 ms /    30 runs   (  269.21
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   29198.36 ms /   130 tokens
No. of rows:  72%| | 900/1258 [8:17:15<2:58:08, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      13.86 ms /    50 runs   (    0.28
ms per token,  3607.24 tokens per second)
llama_print_timings: prompt eval time =   20712.50 ms /    96 tokens (  215.76
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   13216.47 ms /    49 runs   (  269.72
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   34126.42 ms /   145 tokens
No. of rows:  72%| | 901/1258 [8:17:50<3:05:16, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.27 ms /    43 runs   (    0.26
ms per token,  3814.09 tokens per second)
llama_print_timings: prompt eval time =   22509.32 ms /   105 tokens (  214.37
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   11326.04 ms /    42 runs   (  269.67
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   34001.31 ms /   147 tokens
No. of rows:  72%| | 902/1258 [8:18:24<3:09:51, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.09 ms /    34 runs   (    0.27
ms per token,  3742.02 tokens per second)
llama_print_timings: prompt eval time =   23372.87 ms /   109 tokens (  214.43
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   10620.18 ms /    33 runs   (  321.82
ms per token,     3.11 tokens per second)
llama_print_timings:       total time =   34128.86 ms /   142 tokens
No. of rows:  72%| | 903/1258 [8:18:58<3:13:07, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.11 ms /    34 runs   (    0.27
ms per token,  3732.57 tokens per second)
llama_print_timings: prompt eval time =   21448.51 ms /   101 tokens (  212.36
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =    9118.83 ms /    33 runs   (  276.33
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   30702.01 ms /   134 tokens
No. of rows:  72%| | 904/1258 [8:19:28<3:09:09, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.21 ms /    44 runs   (    0.28
ms per token,  3603.31 tokens per second)
llama_print_timings: prompt eval time =   20430.75 ms /    96 tokens (  212.82
ms per token,     4.70 tokens per second)
```

```
llama_print_timings:        eval time =   12203.07 ms /    43 runs   (  283.79
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   32813.29 ms /   139 tokens
No. of rows:  72%| | 905/1258 [8:20:01<3:09:56, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.99 ms /    40 runs   (    0.27
ms per token,  3639.67 tokens per second)
llama_print_timings: prompt eval time =   24423.64 ms /   115 tokens (  212.38
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   10636.83 ms /    39 runs   (  272.74
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   35221.71 ms /   154 tokens
No. of rows:  72%| | 906/1258 [8:20:36<3:14:35, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.61 ms /    34 runs   (    0.28
ms per token,  3537.61 tokens per second)
llama_print_timings: prompt eval time =   23694.50 ms /   111 tokens (  213.46
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    8864.29 ms /    33 runs   (  268.61
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   32693.94 ms /   144 tokens
No. of rows:  72%| | 907/1258 [8:21:09<3:13:13, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.67 ms /    38 runs   (    0.28
ms per token,  3561.72 tokens per second)
llama_print_timings: prompt eval time =   23515.85 ms /   104 tokens (  226.11
ms per token,     4.42 tokens per second)
llama_print_timings:        eval time =   11696.89 ms /    37 runs   (  316.13
ms per token,     3.16 tokens per second)
llama_print_timings:       total time =   35363.32 ms /   141 tokens
No. of rows:  72%| | 908/1258 [8:21:45<3:16:46, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.98 ms /    29 runs   (    0.28
ms per token,  3632.72 tokens per second)
llama_print_timings: prompt eval time =   19623.14 ms /    89 tokens (  220.48
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    7658.07 ms /    28 runs   (  273.50
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   27396.55 ms /   117 tokens
No. of rows:  72%| | 909/1258 [8:22:12<3:05:10, 31Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.12 ms /    35 runs   (    0.26
ms per token,  3835.62 tokens per second)
llama_print_timings: prompt eval time =   21038.92 ms /    97 tokens (  216.90
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    9213.31 ms /    34 runs   (  270.98
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   30387.87 ms /   131 tokens
No. of rows:  72%| | 910/1258 [8:22:42<3:02:07, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.34 ms /    38 runs   (    0.27
ms per token,  3673.63 tokens per second)
llama_print_timings: prompt eval time =   20996.51 ms /    97 tokens (  216.46
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   10074.43 ms /    37 runs   (  272.28
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   31223.09 ms /   134 tokens
No. of rows:  72%| | 911/1258 [8:23:14<3:01:17, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.41 ms /    24 runs   (    0.27
ms per token,  3741.81 tokens per second)
llama_print_timings: prompt eval time =   18283.96 ms /    83 tokens (  220.29
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    6247.59 ms /    23 runs   (  271.63
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   24626.41 ms /   106 tokens
No. of rows:  72%| | 912/1258 [8:23:38<2:49:09, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.16 ms /    25 runs   (    0.29
ms per token,  3493.08 tokens per second)
llama_print_timings: prompt eval time =   17589.10 ms /    81 tokens (  217.15
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    6617.23 ms /    24 runs   (  275.72
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   24308.12 ms /   105 tokens
No. of rows:  73%| | 913/1258 [8:24:03<2:40:01, 27Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.83 ms /    50 runs   (    0.28
```

```
ms per token,  3615.59 tokens per second)
llama_print_timings: prompt eval time =   26984.12 ms /   128 tokens (  210.81
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =   13025.78 ms /    49 runs   (  265.83
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   40208.04 ms /   177 tokens
No. of rows:  73%| | 914/1258 [8:24:43<3:00:51, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.26 ms /    24 runs   (    0.26
ms per token,  3833.25 tokens per second)
llama_print_timings: prompt eval time =   19703.33 ms /    92 tokens (  214.17
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    6196.04 ms /    23 runs   (  269.39
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   25992.34 ms /   115 tokens
No. of rows:  73%| | 915/1258 [8:25:09<2:50:49, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.91 ms /    30 runs   (    0.26
ms per token,  3795.07 tokens per second)
llama_print_timings: prompt eval time =   19274.83 ms /    89 tokens (  216.57
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7837.59 ms /    29 runs   (  270.26
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   27232.49 ms /   118 tokens
No. of rows:  73%| | 916/1258 [8:25:36<2:45:48, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.82 ms /    41 runs   (    0.26
ms per token,  3787.88 tokens per second)
llama_print_timings: prompt eval time =   20092.95 ms /    94 tokens (  213.75
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   11111.77 ms /    40 runs   (  277.79
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   31369.14 ms /   134 tokens
No. of rows:  73%| | 917/1258 [8:26:07<2:49:13, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.18 ms /    50 runs   (    0.28
ms per token,  3526.59 tokens per second)
llama_print_timings: prompt eval time =   25602.43 ms /   120 tokens (  213.35
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   13213.20 ms /    49 runs   (  269.66
```

```
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   39014.66 ms /   169 tokens
No. of rows:  73%| | 918/1258 [8:26:46<3:04:26, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.54 ms /    35 runs   (    0.27
ms per token,  3667.99 tokens per second)
llama_print_timings: prompt eval time =   20273.60 ms /    93 tokens (  218.00
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    9209.11 ms /    34 runs   (  270.86
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   29626.13 ms /   127 tokens
No. of rows:  73%| | 919/1258 [8:27:16<2:58:56, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.98 ms /    30 runs   (    0.27
ms per token,  3759.40 tokens per second)
llama_print_timings: prompt eval time =   28875.23 ms /   135 tokens (  213.89
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    7812.98 ms /    29 runs   (  269.41
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   36804.03 ms /   164 tokens
No. of rows:  73%| | 920/1258 [8:27:53<3:07:06, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.34 ms /    50 runs   (    0.27
ms per token,  3748.41 tokens per second)
llama_print_timings: prompt eval time =   33567.44 ms /   153 tokens (  219.40
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   13164.57 ms /    49 runs   (  268.66
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   46932.68 ms /   202 tokens
No. of rows:  73%| | 921/1258 [8:28:40<3:29:40, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.41 ms /    26 runs   (    0.29
ms per token,  3506.88 tokens per second)
llama_print_timings: prompt eval time =   16957.90 ms /    79 tokens (  214.66
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    6990.04 ms /    25 runs   (  279.60
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   24052.82 ms /   104 tokens
No. of rows:  73%| | 922/1258 [8:29:04<3:06:45, 33Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.33 ms /    49 runs   (    0.29
ms per token,  3418.68 tokens per second)
llama_print_timings: prompt eval time =   19434.74 ms /    91 tokens (  213.57
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   15119.91 ms /    48 runs   (  315.00
ms per token,     3.17 tokens per second)
llama_print_timings:       total time =   34752.18 ms /   139 tokens
No. of rows:  73%| | 923/1258 [8:29:39<3:08:34, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.90 ms /    26 runs   (    0.27
ms per token,  3767.57 tokens per second)
llama_print_timings: prompt eval time =   19079.12 ms /    88 tokens (  216.81
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    6709.32 ms /    25 runs   (  268.37
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   25890.20 ms /   113 tokens
No. of rows:  73%| | 924/1258 [8:30:04<2:54:51, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.76 ms /    33 runs   (    0.27
ms per token,  3765.40 tokens per second)
llama_print_timings: prompt eval time =   22059.78 ms /    97 tokens (  227.42
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =    8876.93 ms /    32 runs   (  277.40
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   31065.93 ms /   129 tokens
No. of rows:  74%| | 925/1258 [8:30:36<2:53:46, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.19 ms /    34 runs   (    0.27
ms per token,  3699.67 tokens per second)
llama_print_timings: prompt eval time =   18157.46 ms /    85 tokens (  213.62
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    8793.48 ms /    33 runs   (  266.47
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   27084.64 ms /   118 tokens
No. of rows:  74%| | 926/1258 [8:31:03<2:46:14, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.47 ms /    47 runs   (    0.29
ms per token,  3488.20 tokens per second)
```

```
llama_print_timings: prompt eval time =   23825.57 ms /   109 tokens (  218.58
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =   14450.88 ms /    46 runs   (  314.15
ms per token,     3.18 tokens per second)
llama_print_timings:       total time =   38467.99 ms /   155 tokens
No. of rows:  74%| | 927/1258 [8:31:41<2:59:41, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.45 ms /    50 runs   (    0.29
ms per token,  3460.93 tokens per second)
llama_print_timings: prompt eval time =   31313.83 ms /   148 tokens (  211.58
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   13204.26 ms /    49 runs   (  269.47
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   44716.66 ms /   197 tokens
No. of rows:  74%| | 928/1258 [8:32:26<3:19:11, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.55 ms /    35 runs   (    0.27
ms per token,  3665.31 tokens per second)
llama_print_timings: prompt eval time =   22047.44 ms /   102 tokens (  216.15
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    9297.01 ms /    34 runs   (  273.44
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   31484.44 ms /   136 tokens
No. of rows:  74%| | 929/1258 [8:32:57<3:10:49, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.05 ms /    47 runs   (    0.30
ms per token,  3346.15 tokens per second)
llama_print_timings: prompt eval time =   24118.42 ms /   114 tokens (  211.57
ms per token,     4.73 tokens per second)
llama_print_timings:        eval time =   12885.88 ms /    46 runs   (  280.13
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   37203.41 ms /   160 tokens
No. of rows:  74%| | 930/1258 [8:33:35<3:14:12, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.84 ms /    50 runs   (    0.28
ms per token,  3613.24 tokens per second)
llama_print_timings: prompt eval time =   28713.89 ms /   137 tokens (  209.59
ms per token,     4.77 tokens per second)
llama_print_timings:        eval time =   13533.52 ms /    49 runs   (  276.19
ms per token,     3.62 tokens per second)
```

```
llama_print_timings:       total time =   42448.14 ms /   186 tokens
No. of rows:  74%| | 931/1258 [8:34:17<3:24:55, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.79 ms /    47 runs   (    0.27
ms per token,  3675.32 tokens per second)
llama_print_timings: prompt eval time =   20384.66 ms /    93 tokens (  219.19
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   12338.89 ms /    46 runs   (  268.24
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   32910.37 ms /   139 tokens
No. of rows:  74%| | 932/1258 [8:34:50<3:16:39, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.95 ms /    50 runs   (    0.28
ms per token,  3585.00 tokens per second)
llama_print_timings: prompt eval time =   22761.02 ms /   107 tokens (  212.72
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   13437.44 ms /    49 runs   (  274.23
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   36401.24 ms /   156 tokens
No. of rows:  74%| | 933/1258 [8:35:26<3:16:24, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.82 ms /    25 runs   (    0.27
ms per token,  3667.30 tokens per second)
llama_print_timings: prompt eval time =   19597.84 ms /    89 tokens (  220.20
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    6392.30 ms /    24 runs   (  266.35
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   26088.68 ms /   113 tokens
No. of rows:  74%| | 934/1258 [8:35:52<2:59:21, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.30 ms /    22 runs   (    0.29
ms per token,  3490.96 tokens per second)
llama_print_timings: prompt eval time =   17514.39 ms /    81 tokens (  216.23
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    5875.91 ms /    21 runs   (  279.81
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   23482.74 ms /   102 tokens
No. of rows:  74%| | 935/1258 [8:36:16<2:43:05, 30Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.24 ms /    44 runs   (    0.28
ms per token,  3595.65 tokens per second)
llama_print_timings: prompt eval time =   20981.89 ms /    98 tokens (  214.10
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   11600.20 ms /    43 runs   (  269.77
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   32754.59 ms /   141 tokens
No. of rows:  74%| | 936/1258 [8:36:49<2:46:34, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.58 ms /    43 runs   (    0.29
ms per token,  3418.40 tokens per second)
llama_print_timings: prompt eval time =   23662.98 ms /   111 tokens (  213.18
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   11109.85 ms /    42 runs   (  264.52
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   34944.22 ms /   153 tokens
No. of rows:  74%| | 937/1258 [8:37:24<2:52:19, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.25 ms /    38 runs   (    0.27
ms per token,  3706.96 tokens per second)
llama_print_timings: prompt eval time =   20245.97 ms /    94 tokens (  215.38
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    9845.54 ms /    37 runs   (  266.10
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   30243.14 ms /   131 tokens
No. of rows:  75%| | 938/1258 [8:37:54<2:48:39, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.68 ms /    43 runs   (    0.27
ms per token,  3683.08 tokens per second)
llama_print_timings: prompt eval time =   24819.52 ms /   110 tokens (  225.63
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   11177.41 ms /    42 runs   (  266.13
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   36165.67 ms /   152 tokens
No. of rows:  75%| | 939/1258 [8:38:30<2:55:22, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.22 ms /    50 runs   (    0.26
ms per token,  3781.86 tokens per second)
llama_print_timings: prompt eval time =   22217.35 ms /   104 tokens (  213.63
```

```
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13152.77 ms /    49 runs   (  268.42
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   35565.04 ms /   153 tokens
No. of rows:  75%| | 940/1258 [8:39:06<2:58:56, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.37 ms /    50 runs   (    0.29
ms per token,  3479.71 tokens per second)
llama_print_timings: prompt eval time =   22626.46 ms /   102 tokens (  221.83
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =   13445.97 ms /    49 runs   (  274.41
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   36270.17 ms /   151 tokens
No. of rows:  75%| | 941/1258 [8:39:42<3:02:21, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      18.55 ms /    50 runs   (    0.37
ms per token,  2695.42 tokens per second)
llama_print_timings: prompt eval time =   22110.13 ms /   102 tokens (  216.77
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   13386.75 ms /    49 runs   (  273.20
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   35705.38 ms /   151 tokens
No. of rows:  75%| | 942/1258 [8:40:18<3:03:41, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.41 ms /    32 runs   (    0.26
ms per token,  3804.54 tokens per second)
llama_print_timings: prompt eval time =   21494.44 ms /    98 tokens (  219.33
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    8338.83 ms /    31 runs   (  268.99
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   29960.63 ms /   129 tokens
No. of rows:  75%| | 943/1258 [8:40:48<2:55:22, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.03 ms /    30 runs   (    0.27
ms per token,  3738.32 tokens per second)
llama_print_timings: prompt eval time =   16006.33 ms /    74 tokens (  216.30
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7781.66 ms /    29 runs   (  268.33
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   23908.15 ms /   103 tokens
```

```
No. of rows:  75%| | 944/1258 [8:41:11<2:39:53, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.39 ms /    42 runs   (    0.27
ms per token,  3687.12 tokens per second)
llama_print_timings: prompt eval time =   21900.47 ms /    96 tokens (  228.13
ms per token,     4.38 tokens per second)
llama_print_timings:        eval time =   11021.63 ms /    41 runs   (  268.82
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   33087.18 ms /   137 tokens
No. of rows:  75%| | 945/1258 [8:41:45<2:43:22, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.14 ms /    34 runs   (    0.27
ms per token,  3717.88 tokens per second)
llama_print_timings: prompt eval time =   20405.38 ms /    94 tokens (  217.08
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   10654.68 ms /    33 runs   (  322.87
ms per token,     3.10 tokens per second)
llama_print_timings:       total time =   31193.82 ms /   127 tokens
No. of rows:  75%| | 946/1258 [8:42:16<2:42:39, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.13 ms /    31 runs   (    0.29
ms per token,  3395.03 tokens per second)
llama_print_timings: prompt eval time =   19652.85 ms /    92 tokens (  213.62
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    8123.77 ms /    30 runs   (  270.79
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   27906.25 ms /   122 tokens
No. of rows:  75%| | 947/1258 [8:42:44<2:36:53, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.92 ms /    37 runs   (    0.27
ms per token,  3730.59 tokens per second)
llama_print_timings: prompt eval time =   22308.57 ms /   106 tokens (  210.46
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =    9642.76 ms /    36 runs   (  267.85
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   32100.75 ms /   142 tokens
No. of rows:  75%| | 948/1258 [8:43:16<2:39:14, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =       9.20 ms /    34 runs   (    0.27
ms per token,  3695.25 tokens per second)
llama_print_timings: prompt eval time =   20503.78 ms /    96 tokens (  213.58
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    8976.66 ms /    33 runs   (  272.02
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   29617.61 ms /   129 tokens
No. of rows:  75%| | 949/1258 [8:43:45<2:36:52, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.05 ms /    45 runs   (    0.27
ms per token,  3733.82 tokens per second)
llama_print_timings: prompt eval time =   21705.99 ms /   101 tokens (  214.91
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   11848.31 ms /    44 runs   (  269.28
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   33730.77 ms /   145 tokens
No. of rows:  76%| | 950/1258 [8:44:19<2:41:26, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.52 ms /    27 runs   (    0.28
ms per token,  3590.43 tokens per second)
llama_print_timings: prompt eval time =   17293.08 ms /    80 tokens (  216.16
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    6898.26 ms /    26 runs   (  265.32
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   24296.87 ms /   106 tokens
No. of rows:  76%| | 951/1258 [8:44:43<2:29:55, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.85 ms /    40 runs   (    0.27
ms per token,  3686.98 tokens per second)
llama_print_timings: prompt eval time =   19135.53 ms /    90 tokens (  212.62
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   10435.74 ms /    39 runs   (  267.58
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   29731.39 ms /   129 tokens
No. of rows:  76%| | 952/1258 [8:45:13<2:30:07, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.99 ms /    26 runs   (    0.27
ms per token,  3718.00 tokens per second)
llama_print_timings: prompt eval time =   22345.80 ms /    99 tokens (  225.72
ms per token,     4.43 tokens per second)
```

```
llama_print_timings:        eval time =    6761.07 ms /    25 runs   (  270.44
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   29207.49 ms /   124 tokens
No. of rows:  76%| | 953/1258 [8:45:42<2:29:16, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.29 ms /    50 runs   (    0.27
ms per token,  3760.81 tokens per second)
llama_print_timings: prompt eval time =   29144.08 ms /   138 tokens (  211.19
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =   13198.62 ms /    49 runs   (  269.36
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   42544.97 ms /   187 tokens
No. of rows:  76%| | 954/1258 [8:46:25<2:48:50, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.84 ms /    50 runs   (    0.28
ms per token,  3614.02 tokens per second)
llama_print_timings: prompt eval time =   26587.97 ms /   126 tokens (  211.02
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =   13538.06 ms /    49 runs   (  276.29
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   40328.46 ms /   175 tokens
No. of rows:  76%| | 955/1258 [8:47:05<2:58:54, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.28 ms /    24 runs   (    0.26
ms per token,  3820.44 tokens per second)
llama_print_timings: prompt eval time =   16909.32 ms /    77 tokens (  219.60
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =    6101.85 ms /    23 runs   (  265.30
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   23104.98 ms /   100 tokens
No. of rows:  76%| | 956/1258 [8:47:28<2:39:42, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.02 ms /    27 runs   (    0.26
ms per token,  3848.35 tokens per second)
llama_print_timings: prompt eval time =   21221.07 ms /    99 tokens (  214.35
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    6988.30 ms /    26 runs   (  268.78
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   28317.91 ms /   125 tokens
No. of rows:  76%| | 957/1258 [8:47:57<2:34:03, 30Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.52 ms /    50 runs   (    0.27
ms per token,  3699.59 tokens per second)
llama_print_timings: prompt eval time =   30145.08 ms /   140 tokens (  215.32
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   13405.56 ms /    49 runs   (  273.58
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   43749.92 ms /   189 tokens
No. of rows:  76%| | 958/1258 [8:48:40<2:53:07, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.68 ms /    29 runs   (    0.26
ms per token,  3778.01 tokens per second)
llama_print_timings: prompt eval time =   18389.47 ms /    85 tokens (  216.35
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7593.10 ms /    28 runs   (  271.18
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   26100.67 ms /   113 tokens
No. of rows:  76%| | 959/1258 [8:49:07<2:39:48, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.14 ms /    44 runs   (    0.28
ms per token,  3624.38 tokens per second)
llama_print_timings: prompt eval time =   19955.64 ms /    94 tokens (  212.29
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   11744.72 ms /    43 runs   (  273.13
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   31873.61 ms /   137 tokens
No. of rows:  76%| | 960/1258 [8:49:38<2:39:00, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.57 ms /    32 runs   (    0.27
ms per token,  3735.26 tokens per second)
llama_print_timings: prompt eval time =   20227.42 ms /    94 tokens (  215.19
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    8361.89 ms /    31 runs   (  269.74
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   28713.67 ms /   125 tokens
No. of rows:  76%| | 961/1258 [8:50:07<2:33:34, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.33 ms /    50 runs   (    0.27
```

```
ms per token,  3750.66 tokens per second)
llama_print_timings: prompt eval time =   29908.04 ms /   142 tokens (  210.62
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =   13387.55 ms /    49 runs   (  273.22
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   43497.17 ms /   191 tokens
No. of rows:  76%| | 962/1258 [8:50:51<2:51:31, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.77 ms /    50 runs   (    0.28
ms per token,  3630.29 tokens per second)
llama_print_timings: prompt eval time =   19281.50 ms /    90 tokens (  214.24
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13095.94 ms /    49 runs   (  267.26
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   32575.33 ms /   139 tokens
No. of rows:  77%| | 963/1258 [8:51:23<2:47:44, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.69 ms /    50 runs   (    0.27
ms per token,  3652.57 tokens per second)
llama_print_timings: prompt eval time =   23551.14 ms /   109 tokens (  216.07
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   13220.53 ms /    49 runs   (  269.81
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   36974.49 ms /   158 tokens
No. of rows:  77%| | 964/1258 [8:52:00<2:51:21, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.64 ms /    36 runs   (    0.27
ms per token,  3735.60 tokens per second)
llama_print_timings: prompt eval time =   20402.38 ms /    88 tokens (  231.85
ms per token,     4.31 tokens per second)
llama_print_timings:        eval time =    9408.18 ms /    35 runs   (  268.81
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   29951.43 ms /   123 tokens
No. of rows:  77%| | 965/1258 [8:52:30<2:43:27, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.64 ms /    50 runs   (    0.27
ms per token,  3665.69 tokens per second)
llama_print_timings: prompt eval time =   24502.95 ms /   109 tokens (  224.80
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   13150.50 ms /    49 runs   (  268.38
```

```
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   37853.75 ms /   158 tokens
No. of rows:  77%| | 966/1258 [8:53:08<2:49:18, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.32 ms /    29 runs   (    0.29
ms per token,  3487.25 tokens per second)
llama_print_timings: prompt eval time =   22022.28 ms /   103 tokens (  213.81
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    7777.97 ms /    28 runs   (  277.78
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   29924.77 ms /   131 tokens
No. of rows:  77%| | 967/1258 [8:53:38<2:41:39, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.74 ms /    37 runs   (    0.26
ms per token,  3799.94 tokens per second)
llama_print_timings: prompt eval time =   24146.90 ms /   113 tokens (  213.69
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    9464.38 ms /    36 runs   (  262.90
ms per token,     3.80 tokens per second)
llama_print_timings:       total time =   33762.65 ms /   149 tokens
No. of rows:  77%| | 968/1258 [8:54:12<2:41:44, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.88 ms /    45 runs   (    0.26
ms per token,  3788.84 tokens per second)
llama_print_timings: prompt eval time =   21517.88 ms /   102 tokens (  210.96
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =   12087.09 ms /    44 runs   (  274.71
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   33780.71 ms /   146 tokens
No. of rows:  77%| | 969/1258 [8:54:46<2:41:39, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.66 ms /    23 runs   (    0.29
ms per token,  3455.53 tokens per second)
llama_print_timings: prompt eval time =   18637.25 ms /    85 tokens (  219.26
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    5993.69 ms /    22 runs   (  272.44
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   24724.49 ms /   107 tokens
No. of rows:  77%| | 970/1258 [8:55:10<2:28:22, 30Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.50 ms /    50 runs   (    0.27
ms per token,  3704.25 tokens per second)
llama_print_timings: prompt eval time =   23396.67 ms /   108 tokens (  216.64
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   13342.55 ms /    49 runs   (  272.30
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   36938.23 ms /   157 tokens
No. of rows:  77%| | 971/1258 [8:55:47<2:36:31, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.12 ms /    37 runs   (    0.27
ms per token,  3655.40 tokens per second)
llama_print_timings: prompt eval time =   20884.38 ms /    97 tokens (  215.30
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    9609.24 ms /    36 runs   (  266.92
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   30641.02 ms /   133 tokens
No. of rows:  77%| | 972/1258 [8:56:18<2:33:00, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.42 ms /    50 runs   (    0.27
ms per token,  3726.89 tokens per second)
llama_print_timings: prompt eval time =   23052.25 ms /   107 tokens (  215.44
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   13211.59 ms /    49 runs   (  269.62
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   36461.98 ms /   156 tokens
No. of rows:  77%| | 973/1258 [8:56:54<2:38:42, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.74 ms /    25 runs   (    0.27
ms per token,  3708.10 tokens per second)
llama_print_timings: prompt eval time =   19079.22 ms /    88 tokens (  216.81
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    6334.75 ms /    24 runs   (  263.95
ms per token,     3.79 tokens per second)
llama_print_timings:       total time =   25514.17 ms /   112 tokens
No. of rows:  77%| | 974/1258 [8:57:20<2:26:56, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.29 ms /    42 runs   (    0.29
ms per token,  3417.41 tokens per second)
```

```
llama_print_timings: prompt eval time =   26960.00 ms /   126 tokens (  213.97
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   11440.40 ms /    41 runs   (  279.03
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   38574.99 ms /   167 tokens
No. of rows:  78%| | 975/1258 [8:57:58<2:37:04, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.37 ms /    32 runs   (    0.26
ms per token,  3822.26 tokens per second)
llama_print_timings: prompt eval time =   21685.88 ms /    97 tokens (  223.57
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =    8428.81 ms /    31 runs   (  271.90
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   30241.39 ms /   128 tokens
No. of rows:  78%| | 976/1258 [8:58:29<2:32:14, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.90 ms /    29 runs   (    0.27
ms per token,  3671.35 tokens per second)
llama_print_timings: prompt eval time =   19608.56 ms /    86 tokens (  228.01
ms per token,     4.39 tokens per second)
llama_print_timings:        eval time =    7427.21 ms /    28 runs   (  265.26
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   27150.46 ms /   114 tokens
No. of rows:  78%| | 977/1258 [8:58:56<2:24:19, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.82 ms /    22 runs   (    0.26
ms per token,  3781.37 tokens per second)
llama_print_timings: prompt eval time =   20910.92 ms /    97 tokens (  215.58
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    5639.55 ms /    21 runs   (  268.55
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   26636.56 ms /   118 tokens
No. of rows:  78%| | 978/1258 [8:59:23<2:17:59, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.80 ms /    36 runs   (    0.27
ms per token,  3672.72 tokens per second)
llama_print_timings: prompt eval time =   21304.27 ms /    93 tokens (  229.08
ms per token,     4.37 tokens per second)
llama_print_timings:        eval time =    9495.17 ms /    35 runs   (  271.29
ms per token,     3.69 tokens per second)
```

```
llama_print_timings:       total time =   30940.08 ms /   128 tokens
No. of rows:  78%| | 979/1258 [8:59:53<2:19:24, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.45 ms /    41 runs   (    0.25
ms per token,  3923.44 tokens per second)
llama_print_timings: prompt eval time =   21774.53 ms /    97 tokens (  224.48
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   12541.68 ms /    40 runs   (  313.54
ms per token,     3.19 tokens per second)
llama_print_timings:       total time =   34475.02 ms /   137 tokens
No. of rows:  78%| | 980/1258 [9:00:28<2:25:09, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.20 ms /    38 runs   (    0.27
ms per token,  3726.95 tokens per second)
llama_print_timings: prompt eval time =   19682.11 ms /    89 tokens (  221.15
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    9799.25 ms /    37 runs   (  264.84
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   29630.92 ms /   126 tokens
No. of rows:  78%| | 981/1258 [9:00:58<2:22:18, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.60 ms /    34 runs   (    0.25
ms per token,  3954.41 tokens per second)
llama_print_timings: prompt eval time =   21010.34 ms /    98 tokens (  214.39
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    8787.38 ms /    33 runs   (  266.28
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   29928.44 ms /   131 tokens
No. of rows:  78%| | 982/1258 [9:01:28<2:20:34, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.52 ms /    21 runs   (    0.26
ms per token,  3805.04 tokens per second)
llama_print_timings: prompt eval time =   16687.72 ms /    75 tokens (  222.50
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =    5345.51 ms /    20 runs   (  267.28
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   22114.78 ms /    95 tokens
No. of rows:  78%| | 983/1258 [9:01:50<2:08:27, 28Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.28 ms /    48 runs   (    0.28
ms per token,  3613.64 tokens per second)
llama_print_timings: prompt eval time =   20000.85 ms /    93 tokens (  215.06
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13011.05 ms /    47 runs   (  276.83
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   33212.06 ms /   140 tokens
No. of rows:  78%| | 984/1258 [9:02:23<2:15:06, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.39 ms /    30 runs   (    0.31
ms per token,  3196.25 tokens per second)
llama_print_timings: prompt eval time =   27113.86 ms /   122 tokens (  222.24
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =    9375.82 ms /    29 runs   (  323.30
ms per token,     3.09 tokens per second)
llama_print_timings:       total time =   36624.29 ms /   151 tokens
No. of rows:  78%| | 985/1258 [9:03:00<2:24:13, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.47 ms /    27 runs   (    0.28
ms per token,  3613.01 tokens per second)
llama_print_timings: prompt eval time =   20452.61 ms /    81 tokens (  252.50
ms per token,     3.96 tokens per second)
llama_print_timings:        eval time =    7359.39 ms /    26 runs   (  283.05
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   27921.13 ms /   107 tokens
No. of rows:  78%| | 986/1258 [9:03:27<2:18:34, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.16 ms /    33 runs   (    0.25
ms per token,  4042.63 tokens per second)
llama_print_timings: prompt eval time =   19981.45 ms /    88 tokens (  227.06
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =    8421.82 ms /    32 runs   (  263.18
ms per token,     3.80 tokens per second)
llama_print_timings:       total time =   28527.72 ms /   120 tokens
No. of rows:  78%| | 987/1258 [9:03:56<2:15:18, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.38 ms /    45 runs   (    0.28
ms per token,  3634.89 tokens per second)
llama_print_timings: prompt eval time =   30475.94 ms /   150 tokens (  203.17
```

```
ms per token,     4.92 tokens per second)
llama_print_timings:        eval time =   11440.04 ms /    44 runs   (  260.00
ms per token,     3.85 tokens per second)
llama_print_timings:       total time =   42089.83 ms /   194 tokens
No. of rows:  79%| | 988/1258 [9:04:38<2:31:11, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.01 ms /    23 runs   (    0.26
ms per token,  3825.05 tokens per second)
llama_print_timings: prompt eval time =   17653.41 ms /    77 tokens (  229.27
ms per token,     4.36 tokens per second)
llama_print_timings:        eval time =    5656.90 ms /    22 runs   (  257.13
ms per token,     3.89 tokens per second)
llama_print_timings:       total time =   23397.07 ms /    99 tokens
No. of rows:  79%| | 989/1258 [9:05:01<2:16:55, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.63 ms /    33 runs   (    0.26
ms per token,  3823.43 tokens per second)
llama_print_timings: prompt eval time =   20275.99 ms /    95 tokens (  213.43
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    8485.68 ms /    32 runs   (  265.18
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   28885.88 ms /   127 tokens
No. of rows:  79%| | 990/1258 [9:05:30<2:14:12, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.87 ms /    33 runs   (    0.27
ms per token,  3719.15 tokens per second)
llama_print_timings: prompt eval time =   22071.08 ms /   103 tokens (  214.28
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    8595.52 ms /    32 runs   (  268.61
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   30794.64 ms /   135 tokens
No. of rows:  79%| | 991/1258 [9:06:01<2:14:43, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.17 ms /    32 runs   (    0.26
ms per token,  3919.17 tokens per second)
llama_print_timings: prompt eval time =   20268.37 ms /    94 tokens (  215.62
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    8477.97 ms /    31 runs   (  273.48
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   28869.74 ms /   125 tokens
```

```
No. of rows:  79%| | 992/1258 [9:06:30<2:12:21, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.38 ms /    38 runs   (    0.27
ms per token,  3660.89 tokens per second)
llama_print_timings: prompt eval time =   20219.69 ms /    94 tokens (  215.10
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   10033.70 ms /    37 runs   (  271.18
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   30405.56 ms /   131 tokens
No. of rows:  79%| | 993/1258 [9:07:00<2:12:36, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.40 ms /    50 runs   (    0.27
ms per token,  3730.51 tokens per second)
llama_print_timings: prompt eval time =   24510.86 ms /   114 tokens (  215.01
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13374.05 ms /    49 runs   (  272.94
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   38083.21 ms /   163 tokens
No. of rows:  79%| | 994/1258 [9:07:39<2:22:45, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.03 ms /    49 runs   (    0.27
ms per token,  3759.69 tokens per second)
llama_print_timings: prompt eval time =   20305.87 ms /    94 tokens (  216.02
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   12837.45 ms /    48 runs   (  267.45
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   33336.19 ms /   142 tokens
No. of rows:  79%| | 995/1258 [9:08:12<2:23:23, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.21 ms /    35 runs   (    0.26
ms per token,  3798.98 tokens per second)
llama_print_timings: prompt eval time =   21871.01 ms /   103 tokens (  212.34
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =    9077.02 ms /    34 runs   (  266.97
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   31083.71 ms /   137 tokens
No. of rows:  79%| | 996/1258 [9:08:43<2:20:43, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      14.20 ms /    50 runs   (    0.28
ms per token,  3520.38 tokens per second)
llama_print_timings: prompt eval time =   32838.64 ms /   154 tokens (  213.24
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   13484.23 ms /    49 runs   (  275.19
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   46526.70 ms /   203 tokens
No. of rows:  79%| | 997/1258 [9:09:30<2:38:51, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.16 ms /    43 runs   (    0.26
ms per token,  3853.39 tokens per second)
llama_print_timings: prompt eval time =   27149.11 ms /   127 tokens (  213.77
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   11399.91 ms /    42 runs   (  271.43
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   38721.48 ms /   169 tokens
No. of rows:  79%| | 998/1258 [9:10:08<2:41:07, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.06 ms /    46 runs   (    0.26
ms per token,  3812.68 tokens per second)
llama_print_timings: prompt eval time =   25287.86 ms /   116 tokens (  218.00
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =   12202.55 ms /    45 runs   (  271.17
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   37671.69 ms /   161 tokens
No. of rows:  79%| | 999/1258 [9:10:46<2:41:08, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.40 ms /    28 runs   (    0.26
ms per token,  3782.76 tokens per second)
llama_print_timings: prompt eval time =   21342.50 ms /    94 tokens (  227.05
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =    7165.51 ms /    27 runs   (  265.39
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   28615.97 ms /   121 tokens
No. of rows:  79%| | 1000/1258 [9:11:15<2:29:17, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.38 ms /    50 runs   (    0.27
ms per token,  3736.36 tokens per second)
llama_print_timings: prompt eval time =   24351.18 ms /   113 tokens (  215.50
ms per token,     4.64 tokens per second)
```

```
llama_print_timings:        eval time =   13326.92 ms /    49 runs   (  271.98
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   37875.16 ms /   162 tokens
No. of rows:  80%| | 1001/1258 [9:11:52<2:32:45, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.73 ms /    24 runs   (    0.28
ms per token,  3566.12 tokens per second)
llama_print_timings: prompt eval time =   20113.54 ms /    94 tokens (  213.97
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    6175.34 ms /    23 runs   (  268.49
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   26382.61 ms /   117 tokens
No. of rows:  80%| | 1002/1258 [9:12:19<2:20:18, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.34 ms /    43 runs   (    0.29
ms per token,  3484.04 tokens per second)
llama_print_timings: prompt eval time =   21244.69 ms /    95 tokens (  223.63
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   11518.02 ms /    42 runs   (  274.24
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   32939.42 ms /   137 tokens
No. of rows:  80%| | 1003/1258 [9:12:52<2:19:50, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.91 ms /    30 runs   (    0.26
ms per token,  3792.19 tokens per second)
llama_print_timings: prompt eval time =   17932.01 ms /    83 tokens (  216.05
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    7788.22 ms /    29 runs   (  268.56
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   25839.60 ms /   112 tokens
No. of rows:  80%| | 1004/1258 [9:13:18<2:10:19, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.44 ms /    39 runs   (    0.27
ms per token,  3737.06 tokens per second)
llama_print_timings: prompt eval time =   21236.47 ms /    98 tokens (  216.70
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   10390.47 ms /    38 runs   (  273.43
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   31783.11 ms /   136 tokens
No. of rows:  80%| | 1005/1258 [9:13:49<2:11:05, 3Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.80 ms /    50 runs   (    0.28
ms per token,  3623.45 tokens per second)
llama_print_timings: prompt eval time =   27915.53 ms /   131 tokens (  213.10
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   13066.57 ms /    49 runs   (  266.66
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   41180.47 ms /   180 tokens
No. of rows:  80%| | 1006/1258 [9:14:31<2:23:18, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.27 ms /    36 runs   (    0.29
ms per token,  3503.99 tokens per second)
llama_print_timings: prompt eval time =   22670.00 ms /   106 tokens (  213.87
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    9576.66 ms /    35 runs   (  273.62
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   32391.65 ms /   141 tokens
No. of rows:  80%| | 1007/1258 [9:15:03<2:20:33, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.18 ms /    26 runs   (    0.28
ms per token,  3622.68 tokens per second)
llama_print_timings: prompt eval time =   20868.54 ms /    97 tokens (  215.14
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    6826.26 ms /    25 runs   (  273.05
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   27795.85 ms /   122 tokens
No. of rows:  80%| | 1008/1258 [9:15:31<2:12:45, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.94 ms /    50 runs   (    0.26
ms per token,  3864.29 tokens per second)
llama_print_timings: prompt eval time =   24810.10 ms /   111 tokens (  223.51
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   13219.81 ms /    49 runs   (  269.79
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   38230.52 ms /   160 tokens
No. of rows:  80%| | 1009/1258 [9:16:09<2:20:09, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.44 ms /    50 runs   (    0.27
```

```
ms per token,  3721.07 tokens per second)
llama_print_timings: prompt eval time =   28546.35 ms /   130 tokens (  219.59
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   13217.48 ms /    49 runs   (  269.74
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   41960.25 ms /   179 tokens
No. of rows:  80%| | 1010/1258 [9:16:51<2:29:45, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.37 ms /    45 runs   (    0.27
ms per token,  3637.25 tokens per second)
llama_print_timings: prompt eval time =   25004.05 ms /   117 tokens (  213.71
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   11793.17 ms /    44 runs   (  268.03
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   36976.99 ms /   161 tokens
No. of rows:  80%| | 1011/1258 [9:17:28<2:30:05, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.12 ms /    44 runs   (    0.28
ms per token,  3629.17 tokens per second)
llama_print_timings: prompt eval time =   24663.88 ms /   114 tokens (  216.35
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   11603.68 ms /    43 runs   (  269.85
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   36440.88 ms /   157 tokens
No. of rows:  80%| | 1012/1258 [9:18:04<2:29:28, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.68 ms /    25 runs   (    0.27
ms per token,  3744.20 tokens per second)
llama_print_timings: prompt eval time =   18267.84 ms /    85 tokens (  214.92
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    6439.66 ms /    24 runs   (  268.32
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   24807.53 ms /   109 tokens
No. of rows:  81%| | 1013/1258 [9:18:29<2:14:35, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       4.96 ms /    18 runs   (    0.28
ms per token,  3630.50 tokens per second)
llama_print_timings: prompt eval time =   14127.28 ms /    65 tokens (  217.34
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    4565.07 ms /    17 runs   (  268.53
```

```
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   18764.30 ms /    82 tokens
No. of rows:  81%| | 1014/1258 [9:18:48<1:56:43, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.48 ms /    46 runs   (    0.27
ms per token,  3686.78 tokens per second)
llama_print_timings: prompt eval time =   21125.18 ms /    98 tokens (  215.56
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   12013.59 ms /    45 runs   (  266.97
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   33320.97 ms /   143 tokens
No. of rows:  81%| | 1015/1258 [9:19:21<2:01:52, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.21 ms /    50 runs   (    0.26
ms per token,  3784.72 tokens per second)
llama_print_timings: prompt eval time =   25491.51 ms /   118 tokens (  216.03
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   13414.28 ms /    49 runs   (  273.76
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   39103.86 ms /   167 tokens
No. of rows:  81%| | 1016/1258 [9:20:00<2:12:17, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      15.89 ms /    50 runs   (    0.32
ms per token,  3145.84 tokens per second)
llama_print_timings: prompt eval time =   32125.01 ms /   140 tokens (  229.46
ms per token,     4.36 tokens per second)
llama_print_timings:        eval time =   13189.20 ms /    49 runs   (  269.17
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   45515.00 ms /   189 tokens
No. of rows:  81%| | 1017/1258 [9:20:46<2:27:04, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.31 ms /    46 runs   (    0.27
ms per token,  3735.28 tokens per second)
llama_print_timings: prompt eval time =   24188.55 ms /   112 tokens (  215.97
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   12048.61 ms /    45 runs   (  267.75
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   36420.73 ms /   157 tokens
No. of rows:  81%| | 1018/1258 [9:21:22<2:26:14, 3Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.10 ms /    23 runs   (    0.27
ms per token,  3769.87 tokens per second)
llama_print_timings: prompt eval time =   20804.57 ms /    95 tokens (  219.00
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    6021.65 ms /    22 runs   (  273.71
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   26917.99 ms /   117 tokens
No. of rows:  81%| | 1019/1258 [9:21:49<2:14:07, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.81 ms /    50 runs   (    0.28
ms per token,  3620.83 tokens per second)
llama_print_timings: prompt eval time =   29408.71 ms /   140 tokens (  210.06
ms per token,     4.76 tokens per second)
llama_print_timings:        eval time =   13352.29 ms /    49 runs   (  272.50
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   42959.13 ms /   189 tokens
No. of rows:  81%| | 1020/1258 [9:22:32<2:24:36, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.87 ms /    40 runs   (    0.27
ms per token,  3679.51 tokens per second)
llama_print_timings: prompt eval time =   23303.45 ms /   110 tokens (  211.85
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   10774.90 ms /    39 runs   (  276.28
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   34238.32 ms /   149 tokens
No. of rows:  81%| | 1021/1258 [9:23:07<2:21:23, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.41 ms /    42 runs   (    0.27
ms per token,  3682.27 tokens per second)
llama_print_timings: prompt eval time =   23981.35 ms /   114 tokens (  210.36
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =   11396.97 ms /    41 runs   (  277.97
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   35542.40 ms /   155 tokens
No. of rows:  81%| | 1022/1258 [9:23:42<2:20:30, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.52 ms /    28 runs   (    0.27
ms per token,  3721.42 tokens per second)
```

```
llama_print_timings: prompt eval time =   22473.20 ms /   105 tokens (  214.03
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    7372.57 ms /    27 runs   (  273.06
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   29954.45 ms /   132 tokens
No. of rows:  81%| | 1023/1258 [9:24:12<2:13:07, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.63 ms /    33 runs   (    0.26
ms per token,  3822.98 tokens per second)
llama_print_timings: prompt eval time =   18321.60 ms /    83 tokens (  220.74
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    8532.99 ms /    32 runs   (  266.66
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   26985.30 ms /   115 tokens
No. of rows:  81%| | 1024/1258 [9:24:39<2:04:22, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.64 ms /    50 runs   (    0.27
ms per token,  3664.61 tokens per second)
llama_print_timings: prompt eval time =   24379.16 ms /   110 tokens (  221.63
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =   13515.16 ms /    49 runs   (  275.82
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   38093.98 ms /   159 tokens
No. of rows:  81%| | 1025/1258 [9:25:17<2:11:04, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.79 ms /    33 runs   (    0.27
ms per token,  3755.55 tokens per second)
llama_print_timings: prompt eval time =   21451.61 ms /    96 tokens (  223.45
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =   10298.53 ms /    32 runs   (  321.83
ms per token,     3.11 tokens per second)
llama_print_timings:       total time =   31880.78 ms /   128 tokens
No. of rows:  82%| | 1026/1258 [9:25:49<2:08:21, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.17 ms /    34 runs   (    0.27
ms per token,  3709.77 tokens per second)
llama_print_timings: prompt eval time =   21598.55 ms /   101 tokens (  213.85
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    8845.45 ms /    33 runs   (  268.04
ms per token,     3.73 tokens per second)
```

```
llama_print_timings:       total time =   30576.77 ms /   134 tokens
No. of rows:  82%| | 1027/1258 [9:26:20<2:04:47, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.92 ms /    33 runs   (    0.27
ms per token,  3697.89 tokens per second)
llama_print_timings: prompt eval time =   22480.09 ms /   106 tokens (  212.08
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =    8708.78 ms /    32 runs   (  272.15
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   31322.17 ms /   138 tokens
No. of rows:  82%| | 1028/1258 [9:26:51<2:02:59, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.46 ms /    29 runs   (    0.26
ms per token,  3885.84 tokens per second)
llama_print_timings: prompt eval time =   17587.61 ms /    80 tokens (  219.85
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =    7386.55 ms /    28 runs   (  263.81
ms per token,     3.79 tokens per second)
llama_print_timings:       total time =   25089.59 ms /   108 tokens
No. of rows:  82%| | 1029/1258 [9:27:16<1:54:28, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.13 ms /    34 runs   (    0.36
ms per token,  2803.89 tokens per second)
llama_print_timings: prompt eval time =   19812.97 ms /    93 tokens (  213.04
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    9113.80 ms /    33 runs   (  276.18
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   29068.78 ms /   126 tokens
No. of rows:  82%| | 1030/1258 [9:27:45<1:52:54, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.36 ms /    31 runs   (    0.27
ms per token,  3708.13 tokens per second)
llama_print_timings: prompt eval time =   22160.87 ms /   104 tokens (  213.09
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    8094.44 ms /    30 runs   (  269.81
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   30377.32 ms /   134 tokens
No. of rows:  82%| | 1031/1258 [9:28:16<1:53:11, 2Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.71 ms /    42 runs   (    0.26
ms per token,  3920.10 tokens per second)
llama_print_timings: prompt eval time =   22942.04 ms /   107 tokens (  214.41
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   10955.31 ms /    41 runs   (  267.20
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   34061.06 ms /   148 tokens
No. of rows:  82%| | 1032/1258 [9:28:50<1:57:22, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.55 ms /    35 runs   (    0.27
ms per token,  3663.77 tokens per second)
llama_print_timings: prompt eval time =   19502.78 ms /    90 tokens (  216.70
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    9602.02 ms /    34 runs   (  282.41
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   29248.56 ms /   124 tokens
No. of rows:  82%| | 1033/1258 [9:29:19<1:54:43, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.22 ms /    50 runs   (    0.26
ms per token,  3782.15 tokens per second)
llama_print_timings: prompt eval time =   31206.11 ms /   147 tokens (  212.29
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   13402.83 ms /    49 runs   (  273.53
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   44807.55 ms /   196 tokens
No. of rows:  82%| | 1034/1258 [9:30:04<2:10:07, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.39 ms /    40 runs   (    0.26
ms per token,  3850.60 tokens per second)
llama_print_timings: prompt eval time =   23370.94 ms /   109 tokens (  214.41
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   10599.59 ms /    39 runs   (  271.78
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   34132.00 ms /   148 tokens
No. of rows:  82%| | 1035/1258 [9:30:38<2:08:45, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.30 ms /    29 runs   (    0.29
ms per token,  3493.13 tokens per second)
llama_print_timings: prompt eval time =   21764.67 ms /    97 tokens (  224.38
```

```
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =    7410.54 ms /    28 runs   (  264.66
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   29289.25 ms /   125 tokens
No. of rows:  82%| | 1036/1258 [9:31:07<2:02:14, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.50 ms /    50 runs   (    0.27
ms per token,  3704.80 tokens per second)
llama_print_timings: prompt eval time =   25178.62 ms /   118 tokens (  213.38
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   13488.41 ms /    49 runs   (  275.27
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   38863.98 ms /   167 tokens
No. of rows:  82%| | 1037/1258 [9:31:46<2:08:07, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.35 ms /    38 runs   (    0.27
ms per token,  3673.27 tokens per second)
llama_print_timings: prompt eval time =   18416.03 ms /    85 tokens (  216.66
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   10043.19 ms /    37 runs   (  271.44
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   28609.88 ms /   122 tokens
No. of rows:  83%| | 1038/1258 [9:32:15<2:00:45, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.27 ms /    36 runs   (    0.26
ms per token,  3881.82 tokens per second)
llama_print_timings: prompt eval time =   21045.63 ms /    93 tokens (  226.30
ms per token,     4.42 tokens per second)
llama_print_timings:        eval time =    9250.83 ms /    35 runs   (  264.31
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   30442.33 ms /   128 tokens
No. of rows:  83%| | 1039/1258 [9:32:45<1:57:30, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.42 ms /    48 runs   (    0.26
ms per token,  3865.36 tokens per second)
llama_print_timings: prompt eval time =   24617.29 ms /   112 tokens (  219.80
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   12641.78 ms /    47 runs   (  268.97
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   37451.12 ms /   159 tokens
```

```
No. of rows:  83%| | 1040/1258 [9:33:22<2:02:42, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.95 ms /    34 runs   (    0.26
ms per token,  3798.46 tokens per second)
llama_print_timings: prompt eval time =   16847.74 ms /    77 tokens (  218.80
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    8817.39 ms /    33 runs   (  267.19
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   25801.02 ms /   110 tokens
No. of rows:  83%| | 1041/1258 [9:33:48<1:53:30, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.98 ms /    50 runs   (    0.28
ms per token,  3577.56 tokens per second)
llama_print_timings: prompt eval time =   28164.45 ms /   131 tokens (  215.00
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13442.31 ms /    49 runs   (  274.33
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   41806.45 ms /   180 tokens
No. of rows:  83%| | 1042/1258 [9:34:30<2:04:14, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.16 ms /    50 runs   (    0.26
ms per token,  3798.24 tokens per second)
llama_print_timings: prompt eval time =   21797.79 ms /   102 tokens (  213.70
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13222.70 ms /    49 runs   (  269.85
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   35217.61 ms /   151 tokens
No. of rows:  83%| | 1043/1258 [9:35:05<2:04:25, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.01 ms /    29 runs   (    0.28
ms per token,  3621.83 tokens per second)
llama_print_timings: prompt eval time =   19275.64 ms /    89 tokens (  216.58
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7824.66 ms /    28 runs   (  279.45
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   27215.73 ms /   117 tokens
No. of rows:  83%| | 1044/1258 [9:35:33<1:55:49, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      13.46 ms /    50 runs   (    0.27
ms per token,  3715.26 tokens per second)
llama_print_timings: prompt eval time =   29735.35 ms /   141 tokens (  210.89
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =   13100.53 ms /    49 runs   (  267.36
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   43033.48 ms /   190 tokens
No. of rows:  83%| | 1045/1258 [9:36:16<2:06:32, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.54 ms /    42 runs   (    0.27
ms per token,  3640.46 tokens per second)
llama_print_timings: prompt eval time =   23642.45 ms /   106 tokens (  223.04
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =   10942.13 ms /    41 runs   (  266.88
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   34753.95 ms /   147 tokens
No. of rows:  83%| | 1046/1258 [9:36:50<2:05:00, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.45 ms /    28 runs   (    0.27
ms per token,  3760.41 tokens per second)
llama_print_timings: prompt eval time =   17840.30 ms /    81 tokens (  220.25
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    7272.81 ms /    27 runs   (  269.36
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   25223.78 ms /   108 tokens
No. of rows:  83%| | 1047/1258 [9:37:16<1:53:43, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.72 ms /    50 runs   (    0.27
ms per token,  3643.52 tokens per second)
llama_print_timings: prompt eval time =   26732.64 ms /   125 tokens (  213.86
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13240.16 ms /    49 runs   (  270.21
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   40172.50 ms /   174 tokens
No. of rows:  83%| | 1048/1258 [9:37:56<2:01:24, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.11 ms /    28 runs   (    0.29
ms per token,  3454.66 tokens per second)
llama_print_timings: prompt eval time =   18189.61 ms /    84 tokens (  216.54
ms per token,     4.62 tokens per second)
```

```
llama_print_timings:        eval time =    7165.43 ms /    27 runs   (  265.39
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   25469.97 ms /   111 tokens
No. of rows:  83%| | 1049/1258 [9:38:21<1:51:12, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.15 ms /    45 runs   (    0.27
ms per token,  3704.31 tokens per second)
llama_print_timings: prompt eval time =   22715.26 ms /   102 tokens (  222.70
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =   13478.14 ms /    44 runs   (  306.32
ms per token,     3.26 tokens per second)
llama_print_timings:       total time =   36371.33 ms /   146 tokens
No. of rows:  83%| | 1050/1258 [9:38:58<1:55:18, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.63 ms /    35 runs   (    0.28
ms per token,  3635.23 tokens per second)
llama_print_timings: prompt eval time =   19939.69 ms /    94 tokens (  212.12
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =    9442.20 ms /    34 runs   (  277.71
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   29527.27 ms /   128 tokens
No. of rows:  84%| | 1051/1258 [9:39:27<1:50:53, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.24 ms /    50 runs   (    0.28
ms per token,  3510.74 tokens per second)
llama_print_timings: prompt eval time =   22414.80 ms /   105 tokens (  213.47
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13076.77 ms /    49 runs   (  266.87
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   35693.32 ms /   154 tokens
No. of rows:  84%| | 1052/1258 [9:40:03<1:54:01, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.10 ms /    37 runs   (    0.27
ms per token,  3665.18 tokens per second)
llama_print_timings: prompt eval time =   19749.69 ms /    91 tokens (  217.03
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    9703.73 ms /    36 runs   (  269.55
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   29604.36 ms /   127 tokens
No. of rows:  84%| | 1053/1258 [9:40:32<1:49:47, 3Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.82 ms /    50 runs   (    0.28
ms per token,  3618.21 tokens per second)
llama_print_timings: prompt eval time =   24501.38 ms /   115 tokens (  213.06
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   13048.85 ms /    49 runs   (  266.30
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   37748.47 ms /   164 tokens
No. of rows:  84%| | 1054/1258 [9:41:10<1:54:58, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.17 ms /    33 runs   (    0.28
ms per token,  3597.91 tokens per second)
llama_print_timings: prompt eval time =   20924.44 ms /    95 tokens (  220.26
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    8784.94 ms /    32 runs   (  274.53
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   29839.78 ms /   127 tokens
No. of rows:  84%| | 1055/1258 [9:41:40<1:50:23, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.14 ms /    50 runs   (    0.26
ms per token,  3806.33 tokens per second)
llama_print_timings: prompt eval time =   21711.30 ms /   101 tokens (  214.96
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   13166.12 ms /    49 runs   (  268.70
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   35075.12 ms /   150 tokens
No. of rows:  84%| | 1056/1258 [9:42:15<1:52:19, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.77 ms /    50 runs   (    0.28
ms per token,  3630.82 tokens per second)
llama_print_timings: prompt eval time =   28107.82 ms /   134 tokens (  209.76
ms per token,     4.77 tokens per second)
llama_print_timings:        eval time =   13327.66 ms /    49 runs   (  271.99
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   41636.69 ms /   183 tokens
No. of rows:  84%| | 1057/1258 [9:42:57<2:00:05, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.78 ms /    33 runs   (    0.27
```

```
ms per token,  3759.40 tokens per second)
llama_print_timings: prompt eval time =   16913.27 ms /    79 tokens (  214.09
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    8470.81 ms /    32 runs   (  264.71
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   25513.12 ms /   111 tokens
No. of rows:  84%| | 1058/1258 [9:43:22<1:49:09, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.97 ms /    50 runs   (    0.26
ms per token,  3854.16 tokens per second)
llama_print_timings: prompt eval time =   24178.95 ms /   112 tokens (  215.88
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   13249.94 ms /    49 runs   (  270.41
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   37627.28 ms /   161 tokens
No. of rows:  84%| | 1059/1258 [9:44:00<1:53:29, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.74 ms /    29 runs   (    0.27
ms per token,  3748.71 tokens per second)
llama_print_timings: prompt eval time =   20659.85 ms /    95 tokens (  217.47
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    7622.36 ms /    28 runs   (  272.23
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   28396.33 ms /   123 tokens
No. of rows:  84%| | 1060/1258 [9:44:28<1:47:09, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.47 ms /    28 runs   (    0.27
ms per token,  3749.33 tokens per second)
llama_print_timings: prompt eval time =   18711.38 ms /    85 tokens (  220.13
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    7212.26 ms /    27 runs   (  267.12
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   26036.90 ms /   112 tokens
No. of rows:  84%| | 1061/1258 [9:44:54<1:40:17, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.56 ms /    47 runs   (    0.29
ms per token,  3465.31 tokens per second)
llama_print_timings: prompt eval time =   26384.70 ms /   124 tokens (  212.78
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   12716.96 ms /    46 runs   (  276.46
```

```
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   39294.74 ms /   170 tokens
No. of rows:  84%| | 1062/1258 [9:45:34<1:48:21, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.51 ms /    50 runs   (    0.27
ms per token,  3701.78 tokens per second)
llama_print_timings: prompt eval time =   24179.38 ms /   110 tokens (  219.81
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   13246.67 ms /    49 runs   (  270.34
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   37621.88 ms /   159 tokens
No. of rows:  84%| | 1063/1258 [9:46:11<1:52:08, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.27 ms /    30 runs   (    0.28
ms per token,  3628.45 tokens per second)
llama_print_timings: prompt eval time =   18552.40 ms /    86 tokens (  215.73
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    7753.60 ms /    29 runs   (  267.37
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   26423.83 ms /   115 tokens
No. of rows:  85%| | 1064/1258 [9:46:38<1:43:44, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.50 ms /    45 runs   (    0.28
ms per token,  3600.00 tokens per second)
llama_print_timings: prompt eval time =   26554.78 ms /   124 tokens (  214.15
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   11680.58 ms /    44 runs   (  265.47
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   38412.76 ms /   168 tokens
No. of rows:  85%| | 1065/1258 [9:47:16<1:49:19, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.17 ms /    50 runs   (    0.26
ms per token,  3796.51 tokens per second)
llama_print_timings: prompt eval time =   26988.17 ms /   117 tokens (  230.67
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =   13523.60 ms /    49 runs   (  275.99
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   40711.30 ms /   166 tokens
No. of rows:  85%| | 1066/1258 [9:47:57<1:55:13, 3Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.26 ms /    39 runs   (    0.26
ms per token,  3801.17 tokens per second)
llama_print_timings: prompt eval time =   25956.10 ms /   122 tokens (  212.75
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   10279.04 ms /    38 runs   (  270.50
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   36388.17 ms /   160 tokens
No. of rows:  85%| | 1067/1258 [9:48:33<1:54:59, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.79 ms /    40 runs   (    0.27
ms per token,  3707.48 tokens per second)
llama_print_timings: prompt eval time =   22011.50 ms /   102 tokens (  215.80
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   10373.16 ms /    39 runs   (  265.98
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   32542.53 ms /   141 tokens
No. of rows:  85%| | 1068/1258 [9:49:06<1:50:59, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.50 ms /    50 runs   (    0.27
ms per token,  3704.80 tokens per second)
llama_print_timings: prompt eval time =   23748.64 ms /   111 tokens (  213.95
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13774.78 ms /    49 runs   (  281.12
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   37726.15 ms /   160 tokens
No. of rows:  85%| | 1069/1258 [9:49:44<1:52:56, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.85 ms /    38 runs   (    0.29
ms per token,  3502.30 tokens per second)
llama_print_timings: prompt eval time =   22091.26 ms /   102 tokens (  216.58
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   10103.50 ms /    37 runs   (  273.07
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   32345.99 ms /   139 tokens
No. of rows:  85%| | 1070/1258 [9:50:16<1:49:03, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.39 ms /    38 runs   (    0.27
ms per token,  3658.07 tokens per second)
```

```
llama_print_timings: prompt eval time =   23855.99 ms /   112 tokens (  213.00
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    9897.22 ms /    37 runs   (  267.49
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   33903.35 ms /   149 tokens
No. of rows:  85%| | 1071/1258 [9:50:50<1:47:38, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.77 ms /    50 runs   (    0.28
ms per token,  3631.87 tokens per second)
llama_print_timings: prompt eval time =   24094.15 ms /   113 tokens (  213.22
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   13010.06 ms /    49 runs   (  265.51
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   37305.23 ms /   162 tokens
No. of rows:  85%| | 1072/1258 [9:51:27<1:49:38, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.03 ms /    34 runs   (    0.27
ms per token,  3766.06 tokens per second)
llama_print_timings: prompt eval time =   19934.93 ms /    88 tokens (  226.53
ms per token,     4.41 tokens per second)
llama_print_timings:        eval time =    8894.17 ms /    33 runs   (  269.52
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   28962.21 ms /   121 tokens
No. of rows:  85%| | 1073/1258 [9:51:56<1:43:08, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.94 ms /    31 runs   (    0.26
ms per token,  3903.79 tokens per second)
llama_print_timings: prompt eval time =   21159.65 ms /    99 tokens (  213.73
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    7965.68 ms /    30 runs   (  265.52
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   29247.37 ms /   129 tokens
No. of rows:  85%| | 1074/1258 [9:52:25<1:38:42, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.13 ms /    24 runs   (    0.26
ms per token,  3913.26 tokens per second)
llama_print_timings: prompt eval time =   16808.97 ms /    78 tokens (  215.50
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    6026.46 ms /    23 runs   (  262.02
ms per token,     3.82 tokens per second)
```

```
llama_print_timings:       total time =   22930.28 ms /   101 tokens
No. of rows:  85%| | 1075/1258 [9:52:48<1:29:43, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.39 ms /    34 runs   (    0.28
ms per token,  3621.64 tokens per second)
llama_print_timings: prompt eval time =   22756.92 ms /   106 tokens (  214.69
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   10460.65 ms /    33 runs   (  316.99
ms per token,     3.15 tokens per second)
llama_print_timings:       total time =   33356.38 ms /   139 tokens
No. of rows:  86%| | 1076/1258 [9:53:22<1:32:48, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.44 ms /    50 runs   (    0.27
ms per token,  3720.24 tokens per second)
llama_print_timings: prompt eval time =   25554.46 ms /   114 tokens (  224.16
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =   13213.13 ms /    49 runs   (  269.66
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   38964.36 ms /   163 tokens
No. of rows:  86%| | 1077/1258 [9:54:01<1:39:53, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.26 ms /    27 runs   (    0.27
ms per token,  3719.52 tokens per second)
llama_print_timings: prompt eval time =   19972.69 ms /    91 tokens (  219.48
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    7006.98 ms /    26 runs   (  269.50
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   27084.40 ms /   117 tokens
No. of rows:  86%| | 1078/1258 [9:54:28<1:33:54, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.28 ms /    27 runs   (    0.27
ms per token,  3710.32 tokens per second)
llama_print_timings: prompt eval time =   19473.01 ms /    91 tokens (  213.99
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    7013.16 ms /    26 runs   (  269.74
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   26591.24 ms /   117 tokens
No. of rows:  86%| | 1079/1258 [9:54:54<1:29:10, 2Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.40 ms /    27 runs   (    0.27
ms per token,  3650.13 tokens per second)
llama_print_timings: prompt eval time =   18033.69 ms /    83 tokens (  217.27
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    7027.49 ms /    26 runs   (  270.29
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   25167.04 ms /   109 tokens
No. of rows:  86%| | 1080/1258 [9:55:20<1:24:29, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.92 ms /    49 runs   (    0.26
ms per token,  3794.04 tokens per second)
llama_print_timings: prompt eval time =   22408.20 ms /    97 tokens (  231.01
ms per token,     4.33 tokens per second)
llama_print_timings:        eval time =   13141.62 ms /    48 runs   (  273.78
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   35742.58 ms /   145 tokens
No. of rows:  86%| | 1081/1258 [9:55:55<1:30:26, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.06 ms /    34 runs   (    0.27
ms per token,  3751.93 tokens per second)
llama_print_timings: prompt eval time =   22059.48 ms /   103 tokens (  214.17
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =    8774.28 ms /    33 runs   (  265.89
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   30969.14 ms /   136 tokens
No. of rows:  86%| | 1082/1258 [9:56:26<1:30:12, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.41 ms /    50 runs   (    0.27
ms per token,  3728.00 tokens per second)
llama_print_timings: prompt eval time =   22529.50 ms /   105 tokens (  214.57
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   12991.31 ms /    49 runs   (  265.13
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   35720.32 ms /   154 tokens
No. of rows:  86%| | 1083/1258 [9:57:02<1:34:03, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.83 ms /    33 runs   (    0.27
ms per token,  3737.68 tokens per second)
llama_print_timings: prompt eval time =   19368.22 ms /    89 tokens (  217.62
```

```
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    8628.41 ms /    32 runs   (  269.64
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   28127.79 ms /   121 tokens
No. of rows:  86%| | 1084/1258 [9:57:30<1:29:55, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.57 ms /    49 runs   (    0.28
ms per token,  3611.97 tokens per second)
llama_print_timings: prompt eval time =   25960.06 ms /   119 tokens (  218.15
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =   13479.79 ms /    48 runs   (  280.83
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   39635.82 ms /   167 tokens
No. of rows:  86%| | 1085/1258 [9:58:10<1:36:53, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.89 ms /    25 runs   (    0.28
ms per token,  3626.87 tokens per second)
llama_print_timings: prompt eval time =   19023.41 ms /    86 tokens (  221.20
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    6443.62 ms /    24 runs   (  268.48
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   25566.59 ms /   110 tokens
No. of rows:  86%| | 1086/1258 [9:58:35<1:29:24, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.38 ms /    43 runs   (    0.29
ms per token,  3474.19 tokens per second)
llama_print_timings: prompt eval time =   21435.09 ms /    99 tokens (  216.52
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   11138.93 ms /    42 runs   (  265.21
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   32744.26 ms /   141 tokens
No. of rows:  86%| | 1087/1258 [9:59:08<1:30:13, 3Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.96 ms /    19 runs   (    0.31
ms per token,  3185.78 tokens per second)
llama_print_timings: prompt eval time =   15500.91 ms /    69 tokens (  224.65
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =    6806.26 ms /    18 runs   (  378.13
ms per token,     2.64 tokens per second)
llama_print_timings:       total time =   22390.83 ms /    87 tokens
```

```
No. of rows:  86%| | 1088/1258 [9:59:30<1:21:50, 2Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.78 ms /    45 runs   (    0.26
ms per token,  3819.06 tokens per second)
llama_print_timings: prompt eval time =   23883.91 ms /   107 tokens (  223.21
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =   12014.59 ms /    44 runs   (  273.06
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   36075.55 ms /   151 tokens
No. of rows:  87%| | 1089/1258 [10:00:07<1:27:26, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.16 ms /    50 runs   (    0.26
ms per token,  3799.39 tokens per second)
llama_print_timings: prompt eval time =   19616.94 ms /    90 tokens (  217.97
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =   13156.49 ms /    49 runs   (  268.50
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   32972.03 ms /   139 tokens
No. of rows:  87%| | 1090/1258 [10:00:40<1:28:33, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.07 ms /    30 runs   (    0.27
ms per token,  3715.63 tokens per second)
llama_print_timings: prompt eval time =   20027.78 ms /    92 tokens (  217.69
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    7776.86 ms /    29 runs   (  268.17
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   27925.06 ms /   121 tokens
No. of rows:  87%| | 1091/1258 [10:01:07<1:24:56, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.54 ms /    40 runs   (    0.26
ms per token,  3796.87 tokens per second)
llama_print_timings: prompt eval time =   19084.17 ms /    88 tokens (  216.87
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   10460.70 ms /    39 runs   (  268.22
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   29701.12 ms /   127 tokens
No. of rows:  87%| | 1092/1258 [10:01:37<1:23:45, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =       6.12 ms /    22 runs   (    0.28
ms per token,  3596.53 tokens per second)
llama_print_timings: prompt eval time =   18590.87 ms /    83 tokens (  223.99
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =    5620.99 ms /    21 runs   (  267.67
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   24297.95 ms /   104 tokens
No. of rows:  87%| | 1093/1258 [10:02:01<1:18:19, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.79 ms /    50 runs   (    0.26
ms per token,  3909.92 tokens per second)
llama_print_timings: prompt eval time =   21348.86 ms /   100 tokens (  213.49
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13173.14 ms /    49 runs   (  268.84
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   34716.02 ms /   149 tokens
No. of rows:  87%| | 1094/1258 [10:02:36<1:22:58, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.37 ms /    39 runs   (    0.27
ms per token,  3760.12 tokens per second)
llama_print_timings: prompt eval time =   21526.69 ms /   101 tokens (  213.14
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   10076.00 ms /    38 runs   (  265.16
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   31754.45 ms /   139 tokens
No. of rows:  87%| | 1095/1258 [10:03:08<1:23:36, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.10 ms /    50 runs   (    0.28
ms per token,  3546.35 tokens per second)
llama_print_timings: prompt eval time =   38232.25 ms /   179 tokens (  213.59
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13755.45 ms /    49 runs   (  280.72
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   52204.49 ms /   228 tokens
No. of rows:  87%| | 1096/1258 [10:04:00<1:40:27, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.39 ms /    47 runs   (    0.26
ms per token,  3792.46 tokens per second)
llama_print_timings: prompt eval time =   23860.58 ms /   110 tokens (  216.91
ms per token,     4.61 tokens per second)
```

```
llama_print_timings:        eval time =   12310.30 ms /    46 runs   (  267.62
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   36353.90 ms /   156 tokens
No. of rows:  87%| | 1097/1258 [10:04:37<1:39:09, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.11 ms /    44 runs   (    0.28
ms per token,  3634.26 tokens per second)
llama_print_timings: prompt eval time =   22815.79 ms /   104 tokens (  219.38
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   11399.01 ms /    43 runs   (  265.09
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   34387.84 ms /   147 tokens
No. of rows:  87%| | 1098/1258 [10:05:11<1:36:29, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.62 ms /    28 runs   (    0.27
ms per token,  3676.95 tokens per second)
llama_print_timings: prompt eval time =   19777.31 ms /    86 tokens (  229.97
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =    7191.38 ms /    27 runs   (  266.35
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   27081.64 ms /   113 tokens
No. of rows:  87%| | 1099/1258 [10:05:38<1:28:39, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.54 ms /    25 runs   (    0.26
ms per token,  3820.29 tokens per second)
llama_print_timings: prompt eval time =   17901.44 ms /    80 tokens (  223.77
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =    6427.45 ms /    24 runs   (  267.81
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   24425.52 ms /   104 tokens
No. of rows:  87%| | 1100/1258 [10:06:02<1:20:58, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.93 ms /    50 runs   (    0.28
ms per token,  3590.41 tokens per second)
llama_print_timings: prompt eval time =   22774.80 ms /   107 tokens (  212.85
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   15169.37 ms /    49 runs   (  309.58
ms per token,     3.23 tokens per second)
llama_print_timings:       total time =   38142.67 ms /   156 tokens
No. of rows:  88%| | 1101/1258 [10:06:41<1:26:16, Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.27 ms /    50 runs   (    0.29
ms per token,  3503.12 tokens per second)
llama_print_timings: prompt eval time =   24433.13 ms /   114 tokens (  214.33
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13339.52 ms /    49 runs   (  272.24
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   37974.91 ms /   163 tokens
No. of rows:  88%| | 1102/1258 [10:07:19<1:29:37, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.56 ms /    50 runs   (    0.27
ms per token,  3688.13 tokens per second)
llama_print_timings: prompt eval time =   22501.66 ms /   105 tokens (  214.30
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13730.57 ms /    49 runs   (  280.22
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   36439.42 ms /   154 tokens
No. of rows:  88%| | 1103/1258 [10:07:55<1:30:35, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.93 ms /    50 runs   (    0.28
ms per token,  3590.15 tokens per second)
llama_print_timings: prompt eval time =   25463.47 ms /   119 tokens (  213.98
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13225.03 ms /    49 runs   (  269.90
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   38887.02 ms /   168 tokens
No. of rows:  88%| | 1104/1258 [10:08:34<1:32:57, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.11 ms /    50 runs   (    0.26
ms per token,  3814.17 tokens per second)
llama_print_timings: prompt eval time =   28654.32 ms /   137 tokens (  209.16
ms per token,     4.78 tokens per second)
llama_print_timings:        eval time =   13214.09 ms /    49 runs   (  269.68
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   42062.93 ms /   186 tokens
No. of rows:  88%| | 1105/1258 [10:09:16<1:36:49, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.58 ms /    31 runs   (    0.28
```

```
ms per token,  3613.90 tokens per second)
llama_print_timings: prompt eval time =   19287.40 ms /    89 tokens (  216.71
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    8131.09 ms /    30 runs   (  271.04
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   27542.43 ms /   119 tokens
No. of rows:  88%| | 1106/1258 [10:09:44<1:28:16, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.76 ms /    34 runs   (    0.26
ms per token,  3879.51 tokens per second)
llama_print_timings: prompt eval time =   24042.76 ms /   111 tokens (  216.60
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    8955.23 ms /    33 runs   (  271.37
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   33132.59 ms /   144 tokens
No. of rows:  88%| | 1107/1258 [10:10:17<1:26:24, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.23 ms /    36 runs   (    0.28
ms per token,  3518.72 tokens per second)
llama_print_timings: prompt eval time =   26182.11 ms /   121 tokens (  216.38
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    9443.11 ms /    35 runs   (  269.80
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   35770.78 ms /   156 tokens
No. of rows:  88%| | 1108/1258 [10:10:52<1:26:55, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.60 ms /    34 runs   (    0.28
ms per token,  3541.67 tokens per second)
llama_print_timings: prompt eval time =   26390.97 ms /   123 tokens (  214.56
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    8805.79 ms /    33 runs   (  266.84
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   35332.66 ms /   156 tokens
No. of rows:  88%| | 1109/1258 [10:11:28<1:26:45, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.45 ms /    44 runs   (    0.28
ms per token,  3535.56 tokens per second)
llama_print_timings: prompt eval time =   21807.96 ms /   100 tokens (  218.08
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =   11793.13 ms /    43 runs   (  274.26
```

```
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   33782.73 ms /   143 tokens
No. of rows:  88%| | 1110/1258 [10:12:02<1:25:19, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.89 ms /    37 runs   (    0.27
ms per token,  3740.40 tokens per second)
llama_print_timings: prompt eval time =   20017.58 ms /    93 tokens (  215.24
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9613.81 ms /    36 runs   (  267.05
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   29778.58 ms /   129 tokens
No. of rows:  88%| | 1111/1258 [10:12:31<1:21:13, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.82 ms /    47 runs   (    0.27
ms per token,  3665.00 tokens per second)
llama_print_timings: prompt eval time =   24311.76 ms /   114 tokens (  213.26
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   12356.12 ms /    46 runs   (  268.61
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   36856.80 ms /   160 tokens
No. of rows:  88%| | 1112/1258 [10:13:08<1:23:22, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.07 ms /    50 runs   (    0.28
ms per token,  3553.41 tokens per second)
llama_print_timings: prompt eval time =   27998.21 ms /   129 tokens (  217.04
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   13437.41 ms /    49 runs   (  274.23
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   41638.08 ms /   178 tokens
No. of rows:  88%| | 1113/1258 [10:13:50<1:28:09, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.78 ms /    22 runs   (    0.26
ms per token,  3806.23 tokens per second)
llama_print_timings: prompt eval time =   18441.84 ms /    84 tokens (  219.55
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =    5525.13 ms /    21 runs   (  263.10
ms per token,     3.80 tokens per second)
llama_print_timings:       total time =   24052.33 ms /   105 tokens
No. of rows:  89%| | 1114/1258 [10:14:14<1:18:36, Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.36 ms /    28 runs   (    0.26
ms per token,  3804.35 tokens per second)
llama_print_timings: prompt eval time =   17158.99 ms /    78 tokens (  219.99
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =    7215.95 ms /    27 runs   (  267.26
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   24486.84 ms /   105 tokens
No. of rows:  89%| | 1115/1258 [10:14:38<1:12:09, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.27 ms /    48 runs   (    0.28
ms per token,  3616.64 tokens per second)
llama_print_timings: prompt eval time =   24353.66 ms /   115 tokens (  211.77
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   12681.84 ms /    47 runs   (  269.83
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   37229.41 ms /   162 tokens
No. of rows:  89%| | 1116/1258 [10:15:16<1:16:35, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.15 ms /    31 runs   (    0.26
ms per token,  3801.35 tokens per second)
llama_print_timings: prompt eval time =   20702.34 ms /    96 tokens (  215.65
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    8029.47 ms /    30 runs   (  267.65
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   28853.78 ms /   126 tokens
No. of rows:  89%| | 1117/1258 [10:15:45<1:13:34, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.23 ms /    50 runs   (    0.26
ms per token,  3779.00 tokens per second)
llama_print_timings: prompt eval time =   32755.94 ms /   151 tokens (  216.93
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =   13132.05 ms /    49 runs   (  268.00
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   46083.86 ms /   200 tokens
No. of rows:  89%| | 1118/1258 [10:16:31<1:23:24, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.52 ms /    29 runs   (    0.26
ms per token,  3854.85 tokens per second)
```

```
llama_print_timings: prompt eval time =   20015.71 ms /    92 tokens (  217.56
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    7477.21 ms /    28 runs   (  267.04
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   27606.82 ms /   120 tokens
No. of rows:  89%| | 1119/1258 [10:16:58<1:17:09, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.13 ms /    35 runs   (    0.26
ms per token,  3834.78 tokens per second)
llama_print_timings: prompt eval time =   21258.11 ms /    99 tokens (  214.73
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =    9142.74 ms /    34 runs   (  268.90
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   30536.35 ms /   133 tokens
No. of rows:  89%| | 1120/1258 [10:17:29<1:14:41, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.92 ms /    34 runs   (    0.26
ms per token,  3810.38 tokens per second)
llama_print_timings: prompt eval time =   20056.93 ms /    87 tokens (  230.54
ms per token,     4.34 tokens per second)
llama_print_timings:        eval time =    8966.20 ms /    33 runs   (  271.70
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   29160.40 ms /   120 tokens
No. of rows:  89%| | 1121/1258 [10:17:58<1:11:53, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.34 ms /    50 runs   (    0.27
ms per token,  3748.41 tokens per second)
llama_print_timings: prompt eval time =   23835.67 ms /   108 tokens (  220.70
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =   13185.45 ms /    49 runs   (  269.09
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   37216.95 ms /   157 tokens
No. of rows:  89%| | 1122/1258 [10:18:35<1:15:16, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.11 ms /    40 runs   (    0.28
ms per token,  3601.33 tokens per second)
llama_print_timings: prompt eval time =   17718.44 ms /    82 tokens (  216.08
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   10477.32 ms /    39 runs   (  268.65
ms per token,     3.72 tokens per second)
```

```
llama_print_timings:       total time =   28354.77 ms /   121 tokens
No. of rows:  89%| | 1123/1258 [10:19:04<1:11:26, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.59 ms /    31 runs   (    0.28
ms per token,  3609.69 tokens per second)
llama_print_timings: prompt eval time =   18909.48 ms /    88 tokens (  214.88
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9728.79 ms /    30 runs   (  324.29
ms per token,     3.08 tokens per second)
llama_print_timings:       total time =   28763.84 ms /   118 tokens
No. of rows:  89%| | 1124/1258 [10:19:32<1:08:55, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.15 ms /    49 runs   (    0.27
ms per token,  3725.95 tokens per second)
llama_print_timings: prompt eval time =   20396.98 ms /    93 tokens (  219.32
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   12913.15 ms /    48 runs   (  269.02
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   33505.08 ms /   141 tokens
No. of rows:  89%| | 1125/1258 [10:20:06<1:10:09, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.27 ms /    48 runs   (    0.28
ms per token,  3616.64 tokens per second)
llama_print_timings: prompt eval time =   21068.22 ms /    95 tokens (  221.77
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =   12839.48 ms /    47 runs   (  273.18
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   34099.92 ms /   142 tokens
No. of rows:  90%| | 1126/1258 [10:20:40<1:11:15, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.81 ms /    50 runs   (    0.26
ms per token,  3902.90 tokens per second)
llama_print_timings: prompt eval time =   24819.30 ms /   117 tokens (  212.13
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   13088.23 ms /    49 runs   (  267.11
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   38106.93 ms /   166 tokens
No. of rows:  90%| | 1127/1258 [10:21:18<1:14:28, Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.71 ms /    43 runs   (    0.27
ms per token,  3671.76 tokens per second)
llama_print_timings: prompt eval time =   21600.51 ms /   100 tokens (  216.01
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   11890.28 ms /    42 runs   (  283.10
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   33663.66 ms /   142 tokens
No. of rows:  90%| | 1128/1258 [10:21:52<1:13:37, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.03 ms /    34 runs   (    0.27
ms per token,  3764.81 tokens per second)
llama_print_timings: prompt eval time =   22358.73 ms /   102 tokens (  219.20
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    8859.49 ms /    33 runs   (  268.47
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   31352.19 ms /   135 tokens
No. of rows:  90%| | 1129/1258 [10:22:23<1:11:21, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.58 ms /    43 runs   (    0.27
ms per token,  3713.62 tokens per second)
llama_print_timings: prompt eval time =   23347.18 ms /   108 tokens (  216.18
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   11176.48 ms /    42 runs   (  266.11
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   34695.32 ms /   150 tokens
No. of rows:  90%| | 1130/1258 [10:22:58<1:11:46, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.23 ms /    50 runs   (    0.26
ms per token,  3780.43 tokens per second)
llama_print_timings: prompt eval time =   22645.15 ms /   106 tokens (  213.63
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13083.49 ms /    49 runs   (  267.01
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   35933.70 ms /   155 tokens
No. of rows:  90%| | 1131/1258 [10:23:34<1:12:40, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.85 ms /    29 runs   (    0.27
ms per token,  3693.33 tokens per second)
llama_print_timings: prompt eval time =   21026.92 ms /    96 tokens (  219.03
```

```
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    7619.64 ms /    28 runs   (  272.13
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   28759.13 ms /   124 tokens
No. of rows:  90%| | 1132/1258 [10:24:02<1:08:35, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.70 ms /    34 runs   (    0.29
ms per token,  3504.07 tokens per second)
llama_print_timings: prompt eval time =   18794.71 ms /    85 tokens (  221.11
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    8733.09 ms /    33 runs   (  264.64
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   27671.14 ms /   118 tokens
No. of rows:  90%| | 1133/1258 [10:24:30<1:04:55, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.54 ms /    39 runs   (    0.27
ms per token,  3699.84 tokens per second)
llama_print_timings: prompt eval time =   23321.12 ms /   109 tokens (  213.96
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   10016.21 ms /    38 runs   (  263.58
ms per token,     3.79 tokens per second)
llama_print_timings:       total time =   33492.59 ms /   147 tokens
No. of rows:  90%| | 1134/1258 [10:25:04<1:05:51, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.49 ms /    35 runs   (    0.27
ms per token,  3689.26 tokens per second)
llama_print_timings: prompt eval time =   20422.36 ms /    92 tokens (  221.98
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =    9210.75 ms /    34 runs   (  270.90
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   29770.86 ms /   126 tokens
No. of rows:  90%| | 1135/1258 [10:25:33<1:04:02, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.94 ms /    36 runs   (    0.28
ms per token,  3623.19 tokens per second)
llama_print_timings: prompt eval time =   19831.34 ms /    90 tokens (  220.35
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    9461.06 ms /    35 runs   (  270.32
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   29441.27 ms /   125 tokens
```

```
No. of rows:  90%| | 1136/1258 [10:26:03<1:02:25, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.47 ms /    43 runs   (    0.27
ms per token,  3750.22 tokens per second)
llama_print_timings: prompt eval time =   24858.67 ms /   110 tokens (  225.99
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   11320.08 ms /    42 runs   (  269.53
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   36347.99 ms /   152 tokens
No. of rows:  90%| | 1137/1258 [10:26:39<1:05:20, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.73 ms /    35 runs   (    0.28
ms per token,  3595.27 tokens per second)
llama_print_timings: prompt eval time =   21172.92 ms /    98 tokens (  216.05
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   10963.30 ms /    34 runs   (  322.45
ms per token,     3.10 tokens per second)
llama_print_timings:       total time =   32279.08 ms /   132 tokens
No. of rows:  90%| | 1138/1258 [10:27:12<1:04:44, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.04 ms /    41 runs   (    0.27
ms per token,  3714.44 tokens per second)
llama_print_timings: prompt eval time =   23117.03 ms /   108 tokens (  214.05
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   10701.47 ms /    40 runs   (  267.54
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   33980.67 ms /   148 tokens
No. of rows:  91%| | 1139/1258 [10:27:46<1:05:09, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.87 ms /    50 runs   (    0.28
ms per token,  3605.16 tokens per second)
llama_print_timings: prompt eval time =   37059.56 ms /   168 tokens (  220.59
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =   13146.81 ms /    49 runs   (  268.30
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   50411.56 ms /   217 tokens
No. of rows:  91%| | 1140/1258 [10:28:36<1:14:58, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      13.61 ms /    50 runs   (    0.27
ms per token,  3672.96 tokens per second)
llama_print_timings: prompt eval time =   25927.88 ms /   122 tokens (  212.52
ms per token,     4.71 tokens per second)
llama_print_timings:        eval time =   13112.96 ms /    49 runs   (  267.61
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   39240.53 ms /   171 tokens
No. of rows:  91%| | 1141/1258 [10:29:15<1:14:59, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.61 ms /    36 runs   (    0.27
ms per token,  3747.27 tokens per second)
llama_print_timings: prompt eval time =   22601.33 ms /   105 tokens (  215.25
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    9255.50 ms /    35 runs   (  264.44
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   31999.63 ms /   140 tokens
No. of rows:  91%| | 1142/1258 [10:29:47<1:10:36, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.44 ms /    38 runs   (    0.27
ms per token,  3639.50 tokens per second)
llama_print_timings: prompt eval time =   21966.48 ms /    96 tokens (  228.82
ms per token,     4.37 tokens per second)
llama_print_timings:        eval time =   10390.95 ms /    37 runs   (  280.84
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   32513.10 ms /   133 tokens
No. of rows:  91%| | 1143/1258 [10:30:20<1:07:41, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.06 ms /    29 runs   (    0.28
ms per token,  3597.12 tokens per second)
llama_print_timings: prompt eval time =   21063.96 ms /    98 tokens (  214.94
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    7556.57 ms /    28 runs   (  269.88
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   28738.68 ms /   126 tokens
No. of rows:  91%| | 1144/1258 [10:30:48<1:03:21, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.46 ms /    40 runs   (    0.26
ms per token,  3824.82 tokens per second)
llama_print_timings: prompt eval time =   18686.37 ms /    85 tokens (  219.84
ms per token,     4.55 tokens per second)
```

```
llama_print_timings:        eval time =   10671.59 ms /    39 runs   (  273.63
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   29513.32 ms /   124 tokens
No. of rows:  91%| | 1145/1258 [10:31:18<1:00:38, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       2.97 ms /    11 runs   (    0.27
ms per token,  3703.70 tokens per second)
llama_print_timings: prompt eval time =   14980.75 ms /    67 tokens (  223.59
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =    2598.72 ms /    10 runs   (  259.87
ms per token,     3.85 tokens per second)
llama_print_timings:       total time =   17622.65 ms /    77 tokens
No. of rows:  91%| | 1146/1258 [10:31:36<51:57, 27Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.48 ms /    50 runs   (    0.27
ms per token,  3710.02 tokens per second)
llama_print_timings: prompt eval time =   31557.74 ms /   142 tokens (  222.24
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =   13551.89 ms /    49 runs   (  276.57
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   45311.51 ms /   191 tokens
No. of rows:  91%| | 1147/1258 [10:32:21<1:01:11, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.15 ms /    34 runs   (    0.27
ms per token,  3715.85 tokens per second)
llama_print_timings: prompt eval time =   18804.51 ms /    87 tokens (  216.14
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    8727.36 ms /    33 runs   (  264.47
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   27669.28 ms /   120 tokens
No. of rows:  91%| | 1148/1258 [10:32:49<57:39, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.74 ms /    31 runs   (    0.28
ms per token,  3546.50 tokens per second)
llama_print_timings: prompt eval time =   19822.77 ms /    91 tokens (  217.83
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    7963.71 ms /    30 runs   (  265.46
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   27906.83 ms /   121 tokens
No. of rows:  91%| | 1149/1258 [10:33:17<55:12, 30Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.91 ms /    25 runs   (    0.28
ms per token,  3618.47 tokens per second)
llama_print_timings: prompt eval time =   18614.85 ms /    86 tokens (  216.45
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    6285.15 ms /    24 runs   (  261.88
ms per token,     3.82 tokens per second)
llama_print_timings:       total time =   24995.86 ms /   110 tokens
No. of rows:  91%| | 1150/1258 [10:33:42<51:47, 28Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.84 ms /    50 runs   (    0.28
ms per token,  3612.98 tokens per second)
llama_print_timings: prompt eval time =   28506.71 ms /   128 tokens (  222.71
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =   13718.48 ms /    49 runs   (  279.97
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   42428.55 ms /   177 tokens
No. of rows:  91%| | 1151/1258 [10:34:24<58:37, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.37 ms /    31 runs   (    0.27
ms per token,  3705.92 tokens per second)
llama_print_timings: prompt eval time =   22489.26 ms /   104 tokens (  216.24
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7976.14 ms /    30 runs   (  265.87
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   30588.88 ms /   134 tokens
No. of rows:  92%| | 1152/1258 [10:34:55<56:52, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.91 ms /    25 runs   (    0.28
ms per token,  3615.85 tokens per second)
llama_print_timings: prompt eval time =   17803.53 ms /    82 tokens (  217.12
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    6523.12 ms /    24 runs   (  271.80
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   24427.41 ms /   106 tokens
No. of rows:  92%| | 1153/1258 [10:35:19<52:15, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.79 ms /    50 runs   (    0.28
```

```
ms per token,  3626.87 tokens per second)
llama_print_timings: prompt eval time =   20803.51 ms /    97 tokens (  214.47
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13372.02 ms /    49 runs   (  272.90
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   34377.45 ms /   146 tokens
No. of rows:  92%| | 1154/1258 [10:35:53<54:07, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.57 ms /    50 runs   (    0.27
ms per token,  3683.78 tokens per second)
llama_print_timings: prompt eval time =   32434.56 ms /   148 tokens (  219.15
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   13235.41 ms /    49 runs   (  270.11
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   45866.74 ms /   197 tokens
No. of rows:  92%| | 1155/1258 [10:36:39<1:01:08, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.67 ms /    50 runs   (    0.27
ms per token,  3658.18 tokens per second)
llama_print_timings: prompt eval time =   30343.33 ms /   145 tokens (  209.26
ms per token,     4.78 tokens per second)
llama_print_timings:        eval time =   13181.63 ms /    49 runs   (  269.01
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   43727.53 ms /   194 tokens
No. of rows:  92%| | 1156/1258 [10:37:23<1:04:41, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.75 ms /    29 runs   (    0.27
ms per token,  3744.35 tokens per second)
llama_print_timings: prompt eval time =   18330.03 ms /    83 tokens (  220.84
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    7576.76 ms /    28 runs   (  270.60
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   26022.70 ms /   111 tokens
No. of rows:  92%| | 1157/1258 [10:37:49<57:59, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.84 ms /    50 runs   (    0.28
ms per token,  3613.50 tokens per second)
llama_print_timings: prompt eval time =   26573.69 ms /   117 tokens (  227.13
ms per token,     4.40 tokens per second)
llama_print_timings:        eval time =   13653.30 ms /    49 runs   (  278.64
```

```
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   40427.14 ms /   166 tokens
No. of rows:  92%| | 1158/1258 [10:38:29<1:00:24, Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.15 ms /    50 runs   (    0.26
ms per token,  3802.86 tokens per second)
llama_print_timings: prompt eval time =   23220.97 ms /   106 tokens (  219.07
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   12982.19 ms /    49 runs   (  264.94
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   36399.96 ms /   155 tokens
No. of rows:  92%| | 1159/1258 [10:39:06<59:52, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.38 ms /    50 runs   (    0.27
ms per token,  3736.08 tokens per second)
llama_print_timings: prompt eval time =   24261.85 ms /   115 tokens (  210.97
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =   13047.30 ms /    49 runs   (  266.27
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   37505.22 ms /   164 tokens
No. of rows:  92%| | 1160/1258 [10:39:43<59:52, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.42 ms /    28 runs   (    0.27
ms per token,  3772.57 tokens per second)
llama_print_timings: prompt eval time =   17954.37 ms /    80 tokens (  224.43
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =    7202.19 ms /    27 runs   (  266.75
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   25267.31 ms /   107 tokens
No. of rows:  92%| | 1161/1258 [10:40:09<53:44, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.82 ms /    30 runs   (    0.26
ms per token,  3834.85 tokens per second)
llama_print_timings: prompt eval time =   20131.29 ms /    90 tokens (  223.68
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =    7769.38 ms /    29 runs   (  267.91
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   28016.38 ms /   119 tokens
No. of rows:  92%| | 1162/1258 [10:40:37<50:41, 31Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.89 ms /    34 runs   (    0.26
ms per token,  3823.66 tokens per second)
llama_print_timings: prompt eval time =   22641.60 ms /   105 tokens (  215.63
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =    8994.30 ms /    33 runs   (  272.55
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   31772.36 ms /   138 tokens
No. of rows:  92%| | 1163/1258 [10:41:08<50:12, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.18 ms /    27 runs   (    0.27
ms per token,  3758.88 tokens per second)
llama_print_timings: prompt eval time =   21151.50 ms /    99 tokens (  213.65
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =    6904.02 ms /    26 runs   (  265.54
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   28159.94 ms /   125 tokens
No. of rows:  93%| | 1164/1258 [10:41:37<48:00, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.32 ms /    50 runs   (    0.25
ms per token,  4058.11 tokens per second)
llama_print_timings: prompt eval time =   23877.58 ms /   108 tokens (  221.09
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =   13070.83 ms /    49 runs   (  266.75
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   37141.95 ms /   157 tokens
No. of rows:  93%| | 1165/1258 [10:42:14<50:31, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.19 ms /    32 runs   (    0.26
ms per token,  3906.25 tokens per second)
llama_print_timings: prompt eval time =   23262.00 ms /   104 tokens (  223.67
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =    8391.74 ms /    31 runs   (  270.70
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   31779.53 ms /   135 tokens
No. of rows:  93%| | 1166/1258 [10:42:46<49:36, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.53 ms /    42 runs   (    0.27
ms per token,  3643.30 tokens per second)
```

```
llama_print_timings: prompt eval time =   23201.73 ms /   108 tokens (  214.83
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   10883.27 ms /    41 runs   (  265.45
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   34250.99 ms /   149 tokens
No. of rows:  93%| | 1167/1258 [10:43:20<49:56, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.95 ms /    30 runs   (    0.27
ms per token,  3772.64 tokens per second)
llama_print_timings: prompt eval time =   22168.64 ms /   102 tokens (  217.34
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    7834.49 ms /    29 runs   (  270.15
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   30121.56 ms /   131 tokens
No. of rows:  93%| | 1168/1258 [10:43:50<48:07, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.30 ms /    50 runs   (    0.29
ms per token,  3496.01 tokens per second)
llama_print_timings: prompt eval time =   30812.39 ms /   144 tokens (  213.97
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13687.47 ms /    49 runs   (  279.34
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   44708.24 ms /   193 tokens
No. of rows:  93%| | 1169/1258 [10:44:35<53:12, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.93 ms /    50 runs   (    0.28
ms per token,  3588.35 tokens per second)
llama_print_timings: prompt eval time =   25548.61 ms /   119 tokens (  214.69
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   13314.75 ms /    49 runs   (  271.73
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   39065.23 ms /   168 tokens
No. of rows:  93%| | 1170/1258 [10:45:14<54:01, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.73 ms /    39 runs   (    0.28
ms per token,  3632.98 tokens per second)
llama_print_timings: prompt eval time =   21687.93 ms /    99 tokens (  219.07
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   11100.42 ms /    38 runs   (  292.12
ms per token,     3.42 tokens per second)
```

```
llama_print_timings:       total time =   32950.66 ms /   137 tokens
No. of rows:  93%| | 1171/1258 [10:45:47<51:43, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.31 ms /    32 runs   (    0.26
ms per token,  3849.39 tokens per second)
llama_print_timings: prompt eval time =   20728.12 ms /    94 tokens (  220.51
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    8395.66 ms /    31 runs   (  270.83
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   29248.37 ms /   125 tokens
No. of rows:  93%| | 1172/1258 [10:46:16<48:22, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.68 ms /    32 runs   (    0.27
ms per token,  3686.21 tokens per second)
llama_print_timings: prompt eval time =   18574.99 ms /    84 tokens (  221.13
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    8321.26 ms /    31 runs   (  268.43
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   27021.43 ms /   115 tokens
No. of rows:  93%| | 1173/1258 [10:46:43<44:57, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.48 ms /    50 runs   (    0.27
ms per token,  3708.65 tokens per second)
llama_print_timings: prompt eval time =   25530.52 ms /   120 tokens (  212.75
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   13326.43 ms /    49 runs   (  271.97
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   39053.82 ms /   169 tokens
No. of rows:  93%| | 1174/1258 [10:47:22<47:30, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.27 ms /    45 runs   (    0.27
ms per token,  3668.08 tokens per second)
llama_print_timings: prompt eval time =   24208.35 ms /   113 tokens (  214.23
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   12194.78 ms /    44 runs   (  277.15
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   36588.80 ms /   157 tokens
No. of rows:  93%| | 1175/1258 [10:47:59<48:02, 34Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.31 ms /    31 runs   (    0.27
ms per token,  3730.00 tokens per second)
llama_print_timings: prompt eval time =   18603.57 ms /    86 tokens (  216.32
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    7969.57 ms /    30 runs   (  265.65
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   26697.91 ms /   116 tokens
No. of rows:  93%| | 1176/1258 [10:48:25<44:10, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.13 ms /    33 runs   (    0.28
ms per token,  3616.04 tokens per second)
llama_print_timings: prompt eval time =   18656.83 ms /    84 tokens (  222.11
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =    9065.90 ms /    32 runs   (  283.31
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   27858.14 ms /   116 tokens
No. of rows:  94%| | 1177/1258 [10:48:53<41:49, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.32 ms /    34 runs   (    0.27
ms per token,  3648.85 tokens per second)
llama_print_timings: prompt eval time =   20452.02 ms /    96 tokens (  213.04
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    8789.50 ms /    33 runs   (  266.35
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   29375.70 ms /   129 tokens
No. of rows:  94%| | 1178/1258 [10:49:23<40:40, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.03 ms /    33 runs   (    0.27
ms per token,  3656.10 tokens per second)
llama_print_timings: prompt eval time =   19360.57 ms /    89 tokens (  217.53
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =    8593.14 ms /    32 runs   (  268.54
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   28087.95 ms /   121 tokens
No. of rows:  94%| | 1179/1258 [10:49:51<39:12, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.28 ms /    33 runs   (    0.28
ms per token,  3556.03 tokens per second)
llama_print_timings: prompt eval time =   21364.68 ms /    98 tokens (  218.01
```

```
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    8642.15 ms /    32 runs   (  270.07
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   30146.90 ms /   130 tokens
No. of rows:  94%| | 1180/1258 [10:50:21<38:51, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.05 ms /    35 runs   (    0.26
ms per token,  3868.69 tokens per second)
llama_print_timings: prompt eval time =   20379.33 ms /    94 tokens (  216.80
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    9203.55 ms /    34 runs   (  270.69
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   29718.75 ms /   128 tokens
No. of rows:  94%| | 1181/1258 [10:50:51<38:17, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.48 ms /    50 runs   (    0.27
ms per token,  3709.75 tokens per second)
llama_print_timings: prompt eval time =   21216.42 ms /    99 tokens (  214.31
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13077.06 ms /    49 runs   (  266.88
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   34494.01 ms /   148 tokens
No. of rows:  94%| | 1182/1258 [10:51:25<39:34, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.26 ms /    50 runs   (    0.27
ms per token,  3770.74 tokens per second)
llama_print_timings: prompt eval time =   25560.69 ms /   120 tokens (  213.01
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   13565.83 ms /    49 runs   (  276.85
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   39330.78 ms /   169 tokens
No. of rows:  94%| | 1183/1258 [10:52:04<42:05, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       6.87 ms /    24 runs   (    0.29
ms per token,  3493.45 tokens per second)
llama_print_timings: prompt eval time =   19397.58 ms /    87 tokens (  222.96
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =    6060.03 ms /    23 runs   (  263.48
ms per token,     3.80 tokens per second)
llama_print_timings:       total time =   25560.36 ms /   110 tokens
```

```
No. of rows:  94%| | 1184/1258 [10:52:30<38:31, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.22 ms /    50 runs   (    0.26
ms per token,  3781.00 tokens per second)
llama_print_timings: prompt eval time =   21615.44 ms /    99 tokens (  218.34
ms per token,     4.58 tokens per second)
llama_print_timings:        eval time =   13057.43 ms /    49 runs   (  266.48
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   34872.13 ms /   148 tokens
No. of rows:  94%| | 1185/1258 [10:53:05<39:20, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.32 ms /    38 runs   (    0.27
ms per token,  3683.60 tokens per second)
llama_print_timings: prompt eval time =   24596.55 ms /   113 tokens (  217.67
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =    9952.01 ms /    37 runs   (  268.97
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   34701.53 ms /   150 tokens
No. of rows:  94%| | 1186/1258 [10:53:40<39:39, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.39 ms /    27 runs   (    0.27
ms per token,  3655.56 tokens per second)
llama_print_timings: prompt eval time =   22110.92 ms /   101 tokens (  218.92
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =    6861.39 ms /    26 runs   (  263.90
ms per token,     3.79 tokens per second)
llama_print_timings:       total time =   29078.02 ms /   127 tokens
No. of rows:  94%| | 1187/1258 [10:54:09<37:41, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.11 ms /    45 runs   (    0.27
ms per token,  3716.24 tokens per second)
llama_print_timings: prompt eval time =   25494.89 ms /   119 tokens (  214.24
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   12133.44 ms /    44 runs   (  275.76
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   37808.06 ms /   163 tokens
No. of rows:  94%| | 1188/1258 [10:54:46<39:15, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      14.35 ms /    50 runs   (    0.29
ms per token,  3484.81 tokens per second)
llama_print_timings: prompt eval time =   25306.94 ms /   119 tokens (  212.66
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   13228.68 ms /    49 runs   (  269.97
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   38739.91 ms /   168 tokens
No. of rows:  95%| | 1189/1258 [10:55:25<40:27, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.57 ms /    32 runs   (    0.27
ms per token,  3735.70 tokens per second)
llama_print_timings: prompt eval time =   21308.60 ms /   100 tokens (  213.09
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    8635.72 ms /    31 runs   (  278.57
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   30072.96 ms /   131 tokens
No. of rows:  95%| | 1190/1258 [10:55:55<38:07, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.28 ms /    37 runs   (    0.28
ms per token,  3599.92 tokens per second)
llama_print_timings: prompt eval time =   21269.09 ms /    98 tokens (  217.03
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    9612.76 ms /    36 runs   (  267.02
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   31030.14 ms /   134 tokens
No. of rows:  95%| | 1191/1258 [10:56:26<36:41, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.04 ms /    41 runs   (    0.27
ms per token,  3715.11 tokens per second)
llama_print_timings: prompt eval time =   24000.09 ms /   105 tokens (  228.57
ms per token,     4.37 tokens per second)
llama_print_timings:        eval time =   10712.73 ms /    40 runs   (  267.82
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   34875.90 ms /   145 tokens
No. of rows:  95%| | 1192/1258 [10:57:01<36:49, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.07 ms /    34 runs   (    0.27
ms per token,  3746.97 tokens per second)
llama_print_timings: prompt eval time =   20220.80 ms /    93 tokens (  217.43
ms per token,     4.60 tokens per second)
```

```
llama_print_timings:        eval time =    8840.62 ms /    33 runs   (  267.90
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   29194.75 ms /   126 tokens
No. of rows:  95%| | 1193/1258 [10:57:30<34:52, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.04 ms /    37 runs   (    0.27
ms per token,  3685.26 tokens per second)
llama_print_timings: prompt eval time =   19328.18 ms /    89 tokens (  217.17
ms per token,     4.60 tokens per second)
llama_print_timings:        eval time =   10067.36 ms /    36 runs   (  279.65
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   29542.10 ms /   125 tokens
No. of rows:  95%| | 1194/1258 [10:58:00<33:29, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.34 ms /    43 runs   (    0.26
ms per token,  3791.89 tokens per second)
llama_print_timings: prompt eval time =   20593.76 ms /    96 tokens (  214.52
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   11341.82 ms /    42 runs   (  270.04
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   32104.62 ms /   138 tokens
No. of rows:  95%| | 1195/1258 [10:58:32<33:11, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.81 ms /    36 runs   (    0.27
ms per token,  3670.85 tokens per second)
llama_print_timings: prompt eval time =   28157.52 ms /   132 tokens (  213.31
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =    9411.09 ms /    35 runs   (  268.89
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   37709.68 ms /   167 tokens
No. of rows:  95%| | 1196/1258 [10:59:10<34:33, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.88 ms /    33 runs   (    0.27
ms per token,  3716.63 tokens per second)
llama_print_timings: prompt eval time =   22609.63 ms /   103 tokens (  219.51
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    8534.12 ms /    32 runs   (  266.69
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   31272.85 ms /   135 tokens
No. of rows:  95%| | 1197/1258 [10:59:41<33:20, 32Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.25 ms /    38 runs   (    0.27
ms per token,  3706.96 tokens per second)
llama_print_timings: prompt eval time =   20824.76 ms /    95 tokens (  219.21
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    9910.53 ms /    37 runs   (  267.85
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   30885.22 ms /   132 tokens
No. of rows:  95%| | 1198/1258 [11:00:12<32:13, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.53 ms /    35 runs   (    0.27
ms per token,  3671.07 tokens per second)
llama_print_timings: prompt eval time =   22311.48 ms /   103 tokens (  216.62
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =    8995.36 ms /    34 runs   (  264.57
ms per token,     3.78 tokens per second)
llama_print_timings:       total time =   31443.27 ms /   137 tokens
No. of rows:  95%| | 1199/1258 [11:00:43<31:27, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.29 ms /    42 runs   (    0.29
ms per token,  3416.30 tokens per second)
llama_print_timings: prompt eval time =   22499.39 ms /    98 tokens (  229.59
ms per token,     4.36 tokens per second)
llama_print_timings:        eval time =   10919.46 ms /    41 runs   (  266.33
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   33585.00 ms /   139 tokens
No. of rows:  95%| | 1200/1258 [11:01:17<31:23, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.63 ms /    39 runs   (    0.27
ms per token,  3669.90 tokens per second)
llama_print_timings: prompt eval time =   20974.48 ms /    97 tokens (  216.23
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   10117.56 ms /    38 runs   (  266.25
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   31247.77 ms /   135 tokens
No. of rows:  95%| | 1201/1258 [11:01:48<30:30, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.37 ms /    38 runs   (    0.27
```

```
ms per token,  3665.83 tokens per second)
llama_print_timings: prompt eval time =   19352.81 ms /    88 tokens (  219.92
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =    9873.16 ms /    37 runs   (  266.84
ms per token,     3.75 tokens per second)
llama_print_timings:       total time =   29372.83 ms /   125 tokens
No. of rows:  96%| | 1202/1258 [11:02:18<29:12, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.33 ms /    50 runs   (    0.27
ms per token,  3751.78 tokens per second)
llama_print_timings: prompt eval time =   24959.39 ms /   117 tokens (  213.33
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   13554.55 ms /    49 runs   (  276.62
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   38720.32 ms /   166 tokens
No. of rows:  96%| | 1203/1258 [11:02:56<30:43, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.13 ms /    42 runs   (    0.27
ms per token,  3772.23 tokens per second)
llama_print_timings: prompt eval time =   24291.22 ms /   113 tokens (  214.97
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   11089.87 ms /    41 runs   (  270.48
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   35548.13 ms /   154 tokens
No. of rows:  96%| | 1204/1258 [11:03:32<30:42, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.89 ms /    34 runs   (    0.26
ms per token,  3825.38 tokens per second)
llama_print_timings: prompt eval time =   27793.76 ms /   128 tokens (  217.14
ms per token,     4.61 tokens per second)
llama_print_timings:        eval time =    9286.44 ms /    33 runs   (  281.41
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   37216.22 ms /   161 tokens
No. of rows:  96%| | 1205/1258 [11:04:09<30:58, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.45 ms /    50 runs   (    0.27
ms per token,  3717.47 tokens per second)
llama_print_timings: prompt eval time =   22709.11 ms /   105 tokens (  216.28
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   14894.45 ms /    49 runs   (  303.97
```

```
ms per token,     3.29 tokens per second)
llama_print_timings:       total time =   37801.96 ms /   154 tokens
No. of rows:  96%| | 1206/1258 [11:04:47<31:05, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.55 ms /    50 runs   (    0.27
ms per token,  3688.95 tokens per second)
llama_print_timings: prompt eval time =   26099.58 ms /   121 tokens (  215.70
ms per token,     4.64 tokens per second)
llama_print_timings:        eval time =   13223.12 ms /    49 runs   (  269.86
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   39522.40 ms /   170 tokens
No. of rows:  96%| | 1207/1258 [11:05:26<31:25, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.86 ms /    30 runs   (    0.26
ms per token,  3814.85 tokens per second)
llama_print_timings: prompt eval time =   18717.07 ms /    87 tokens (  215.14
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =    7693.58 ms /    29 runs   (  265.30
ms per token,     3.77 tokens per second)
llama_print_timings:       total time =   26527.15 ms /   116 tokens
No. of rows:  96%| | 1208/1258 [11:05:53<28:12, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.81 ms /    50 runs   (    0.26
ms per token,  3904.42 tokens per second)
llama_print_timings: prompt eval time =   24849.40 ms /   116 tokens (  214.22
ms per token,     4.67 tokens per second)
llama_print_timings:        eval time =   13266.94 ms /    49 runs   (  270.75
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   38311.56 ms /   165 tokens
No. of rows:  96%| | 1209/1258 [11:06:31<28:44, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.77 ms /    34 runs   (    0.29
ms per token,  3479.68 tokens per second)
llama_print_timings: prompt eval time =   23448.28 ms /   103 tokens (  227.65
ms per token,     4.39 tokens per second)
llama_print_timings:        eval time =    9044.20 ms /    33 runs   (  274.07
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   32628.24 ms /   136 tokens
No. of rows:  96%| | 1210/1258 [11:07:04<27:32, 34Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.15 ms /    25 runs   (    0.29
ms per token,  3498.46 tokens per second)
llama_print_timings: prompt eval time =   19240.69 ms /    89 tokens (  216.19
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =    6818.62 ms /    24 runs   (  284.11
ms per token,     3.52 tokens per second)
llama_print_timings:       total time =   26162.76 ms /   113 tokens
No. of rows:  96%| | 1211/1258 [11:07:30<25:01, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.61 ms /    42 runs   (    0.28
ms per token,  3617.88 tokens per second)
llama_print_timings: prompt eval time =   24897.44 ms /   118 tokens (  211.00
ms per token,     4.74 tokens per second)
llama_print_timings:        eval time =   11191.79 ms /    41 runs   (  272.97
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   36254.87 ms /   159 tokens
No. of rows:  96%| | 1212/1258 [11:08:06<25:29, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.59 ms /    50 runs   (    0.27
ms per token,  3678.63 tokens per second)
llama_print_timings: prompt eval time =   24779.54 ms /   110 tokens (  225.27
ms per token,     4.44 tokens per second)
llama_print_timings:        eval time =   13177.28 ms /    49 runs   (  268.92
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   38156.11 ms /   159 tokens
No. of rows:  96%| | 1213/1258 [11:08:45<26:02, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.34 ms /    45 runs   (    0.25
ms per token,  3968.25 tokens per second)
llama_print_timings: prompt eval time =   26447.50 ms /   116 tokens (  228.00
ms per token,     4.39 tokens per second)
llama_print_timings:        eval time =   11942.77 ms /    44 runs   (  271.43
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   38563.95 ms /   160 tokens
No. of rows:  97%| | 1214/1258 [11:09:23<26:18, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.71 ms /    22 runs   (    0.26
ms per token,  3853.56 tokens per second)
```

```
llama_print_timings: prompt eval time =   18903.73 ms /    85 tokens (  222.40
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =    5679.12 ms /    21 runs   (  270.43
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   24668.48 ms /   106 tokens
No. of rows:  97%| | 1215/1258 [11:09:48<23:18, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.13 ms /    50 runs   (    0.26
ms per token,  3808.65 tokens per second)
llama_print_timings: prompt eval time =   26800.86 ms /   121 tokens (  221.49
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =   13766.95 ms /    49 runs   (  280.96
ms per token,     3.56 tokens per second)
llama_print_timings:       total time =   40768.14 ms /   170 tokens
No. of rows:  97%| | 1216/1258 [11:10:29<24:29, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.54 ms /    32 runs   (    0.27
ms per token,  3748.39 tokens per second)
llama_print_timings: prompt eval time =   19498.66 ms /    88 tokens (  221.58
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    8442.24 ms /    31 runs   (  272.33
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   28065.08 ms /   119 tokens
No. of rows:  97%| | 1217/1258 [11:10:57<22:29, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.22 ms /    44 runs   (    0.26
ms per token,  3920.17 tokens per second)
llama_print_timings: prompt eval time =   23795.12 ms /    99 tokens (  240.35
ms per token,     4.16 tokens per second)
llama_print_timings:        eval time =   11606.29 ms /    43 runs   (  269.91
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   35573.61 ms /   142 tokens
No. of rows:  97%| | 1218/1258 [11:11:32<22:28, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.47 ms /    31 runs   (    0.27
ms per token,  3660.41 tokens per second)
llama_print_timings: prompt eval time =   20195.42 ms /    91 tokens (  221.93
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    8293.82 ms /    30 runs   (  276.46
ms per token,     3.62 tokens per second)
```

```
llama_print_timings:       total time =   28614.90 ms /   121 tokens
No. of rows:  97%| | 1219/1258 [11:12:01<20:55, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.23 ms /    31 runs   (    0.27
ms per token,  3768.08 tokens per second)
llama_print_timings: prompt eval time =   20104.25 ms /    90 tokens (  223.38
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =    8138.39 ms /    30 runs   (  271.28
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   28362.82 ms /   120 tokens
No. of rows:  97%| | 1220/1258 [11:12:29<19:39, 31Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.77 ms /    41 runs   (    0.26
ms per token,  3806.16 tokens per second)
llama_print_timings: prompt eval time =   27919.49 ms /   123 tokens (  226.99
ms per token,     4.41 tokens per second)
llama_print_timings:        eval time =   12699.35 ms /    40 runs   (  317.48
ms per token,     3.15 tokens per second)
llama_print_timings:       total time =   40782.37 ms /   163 tokens
No. of rows:  97%| | 1221/1258 [11:13:10<20:56, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.61 ms /    21 runs   (    0.27
ms per token,  3743.98 tokens per second)
llama_print_timings: prompt eval time =   16930.83 ms /    75 tokens (  225.74
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =    5476.61 ms /    20 runs   (  273.83
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   22492.24 ms /    95 tokens
No. of rows:  97%| | 1222/1258 [11:13:32<18:18, 30Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.67 ms /    36 runs   (    0.27
ms per token,  3724.39 tokens per second)
llama_print_timings: prompt eval time =   22115.70 ms /    99 tokens (  223.39
ms per token,     4.48 tokens per second)
llama_print_timings:        eval time =    9618.20 ms /    35 runs   (  274.81
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   31877.48 ms /   134 tokens
No. of rows:  97%| | 1223/1258 [11:14:04<18:02, 30Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.23 ms /    31 runs   (    0.27
ms per token,  3768.54 tokens per second)
llama_print_timings: prompt eval time =   18745.99 ms /    85 tokens (  220.54
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    8442.65 ms /    30 runs   (  281.42
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   27312.05 ms /   115 tokens
No. of rows:  97%| | 1224/1258 [11:14:32<16:54, 29Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.81 ms /    47 runs   (    0.27
ms per token,  3667.58 tokens per second)
llama_print_timings: prompt eval time =   28326.63 ms /   130 tokens (  217.90
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =   14489.41 ms /    46 runs   (  314.99
ms per token,     3.17 tokens per second)
llama_print_timings:       total time =   43005.86 ms /   176 tokens
No. of rows:  97%| | 1225/1258 [11:15:15<18:35, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.26 ms /    39 runs   (    0.26
ms per token,  3800.06 tokens per second)
llama_print_timings: prompt eval time =   21351.90 ms /    95 tokens (  224.76
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =   10151.15 ms /    38 runs   (  267.14
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   31653.97 ms /   133 tokens
No. of rows:  97%| | 1226/1258 [11:15:46<17:41, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.13 ms /    37 runs   (    0.27
ms per token,  3652.88 tokens per second)
llama_print_timings: prompt eval time =   21056.88 ms /    95 tokens (  221.65
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =    9677.11 ms /    36 runs   (  268.81
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   30880.46 ms /   131 tokens
No. of rows:  98%| | 1227/1258 [11:16:17<16:46, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.66 ms /    45 runs   (    0.26
ms per token,  3859.68 tokens per second)
llama_print_timings: prompt eval time =   23165.17 ms /   105 tokens (  220.62
```

```
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =   12043.22 ms /    44 runs   (  273.71
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   35385.85 ms /   149 tokens
No. of rows:  98%| | 1228/1258 [11:16:53<16:40, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.96 ms /    50 runs   (    0.26
ms per token,  3858.32 tokens per second)
llama_print_timings: prompt eval time =   44845.43 ms /   210 tokens (  213.55
ms per token,     4.68 tokens per second)
llama_print_timings:        eval time =   13403.47 ms /    49 runs   (  273.54
ms per token,     3.66 tokens per second)
llama_print_timings:       total time =   58487.36 ms /   259 tokens
No. of rows:  98%| | 1229/1258 [11:17:51<19:46, 40Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      14.06 ms /    50 runs   (    0.28
ms per token,  3556.95 tokens per second)
llama_print_timings: prompt eval time =   28210.08 ms /   129 tokens (  218.68
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =   13352.98 ms /    49 runs   (  272.51
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   41760.28 ms /   178 tokens
No. of rows:  98%| | 1230/1258 [11:18:33<19:12, 41Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.27 ms /    50 runs   (    0.27
ms per token,  3768.18 tokens per second)
llama_print_timings: prompt eval time =   25125.17 ms /   110 tokens (  228.41
ms per token,     4.38 tokens per second)
llama_print_timings:        eval time =   13446.88 ms /    49 runs   (  274.43
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   38769.72 ms /   159 tokens
No. of rows:  98%| | 1231/1258 [11:19:12<18:11, 40Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.72 ms /    21 runs   (    0.27
ms per token,  3668.76 tokens per second)
llama_print_timings: prompt eval time =   18503.15 ms /    83 tokens (  222.93
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =    5380.30 ms /    20 runs   (  269.01
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   23966.30 ms /   103 tokens
```

```
No. of rows:  98%| | 1232/1258 [11:19:36<15:23, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.69 ms /    50 runs   (    0.27
ms per token,  3652.03 tokens per second)
llama_print_timings: prompt eval time =   26581.37 ms /   121 tokens (  219.68
ms per token,     4.55 tokens per second)
llama_print_timings:        eval time =   13557.11 ms /    49 runs   (  276.68
ms per token,     3.61 tokens per second)
llama_print_timings:       total time =   40335.15 ms /   170 tokens
No. of rows:  98%| | 1233/1258 [11:20:16<15:23, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       9.64 ms /    36 runs   (    0.27
ms per token,  3735.21 tokens per second)
llama_print_timings: prompt eval time =   19826.81 ms /    89 tokens (  222.77
ms per token,     4.49 tokens per second)
llama_print_timings:        eval time =    9581.96 ms /    35 runs   (  273.77
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   29549.73 ms /   124 tokens
No. of rows:  98%| | 1234/1258 [11:20:46<13:53, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.46 ms /    44 runs   (    0.26
ms per token,  3840.78 tokens per second)
llama_print_timings: prompt eval time =   29290.09 ms /   134 tokens (  218.58
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =   11605.83 ms /    43 runs   (  269.90
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   41067.39 ms /   177 tokens
No. of rows:  98%| | 1235/1258 [11:21:27<14:02, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.24 ms /    50 runs   (    0.26
ms per token,  3776.44 tokens per second)
llama_print_timings: prompt eval time =   30271.75 ms /   139 tokens (  217.78
ms per token,     4.59 tokens per second)
llama_print_timings:        eval time =   13640.47 ms /    49 runs   (  278.38
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   44115.06 ms /   188 tokens
No. of rows:  98%| | 1236/1258 [11:22:11<14:15, 38Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
```

```
llama_print_timings:      sample time =      12.97 ms /    50 runs   (    0.26
ms per token,  3855.64 tokens per second)
llama_print_timings: prompt eval time =   31070.36 ms /   139 tokens (  223.53
ms per token,     4.47 tokens per second)
llama_print_timings:        eval time =   14913.60 ms /    49 runs   (  304.36
ms per token,     3.29 tokens per second)
llama_print_timings:       total time =   46178.62 ms /   188 tokens
No. of rows:  98%| | 1237/1258 [11:22:57<14:22, 41Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.18 ms /    31 runs   (    0.26
ms per token,  3789.27 tokens per second)
llama_print_timings: prompt eval time =   21547.58 ms /    97 tokens (  222.14
ms per token,     4.50 tokens per second)
llama_print_timings:        eval time =    8049.47 ms /    30 runs   (  268.32
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   29718.91 ms /   127 tokens
No. of rows:  98%| | 1238/1258 [11:23:27<12:33, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.33 ms /    47 runs   (    0.28
ms per token,  3525.09 tokens per second)
llama_print_timings: prompt eval time =   25404.44 ms /   116 tokens (  219.00
ms per token,     4.57 tokens per second)
llama_print_timings:        eval time =   13018.00 ms /    46 runs   (  283.00
ms per token,     3.53 tokens per second)
llama_print_timings:       total time =   38616.63 ms /   162 tokens
No. of rows:  98%| | 1239/1258 [11:24:05<12:01, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.67 ms /    29 runs   (    0.26
ms per token,  3778.99 tokens per second)
llama_print_timings: prompt eval time =   19240.65 ms /    87 tokens (  221.16
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =    7633.28 ms /    28 runs   (  272.62
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   26985.83 ms /   115 tokens
No. of rows:  99%| | 1240/1258 [11:24:32<10:24, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.09 ms /    27 runs   (    0.26
ms per token,  3808.72 tokens per second)
llama_print_timings: prompt eval time =   23116.73 ms /    99 tokens (  233.50
ms per token,     4.28 tokens per second)
```

```
llama_print_timings:        eval time =    6969.32 ms /    26 runs   (  268.05
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   30192.54 ms /   125 tokens
No. of rows:  99%| | 1241/1258 [11:25:02<09:26, 33Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.66 ms /    50 runs   (    0.27
ms per token,  3660.32 tokens per second)
llama_print_timings: prompt eval time =   27628.09 ms /   126 tokens (  219.27
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   13285.40 ms /    49 runs   (  271.13
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   41107.60 ms /   175 tokens
No. of rows:  99%| | 1242/1258 [11:25:44<09:30, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.84 ms /    45 runs   (    0.26
ms per token,  3800.03 tokens per second)
llama_print_timings: prompt eval time =   25275.41 ms /   114 tokens (  221.71
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =   11824.85 ms /    44 runs   (  268.75
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   37275.38 ms /   158 tokens
No. of rows:  99%| | 1243/1258 [11:26:21<09:02, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.67 ms /    42 runs   (    0.28
ms per token,  3598.36 tokens per second)
llama_print_timings: prompt eval time =   26228.64 ms /   114 tokens (  230.08
ms per token,     4.35 tokens per second)
llama_print_timings:        eval time =   10953.37 ms /    41 runs   (  267.16
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   37349.60 ms /   155 tokens
No. of rows:  99%| | 1244/1258 [11:26:58<08:31, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       7.79 ms /    29 runs   (    0.27
ms per token,  3721.29 tokens per second)
llama_print_timings: prompt eval time =   22295.72 ms /   101 tokens (  220.75
ms per token,     4.53 tokens per second)
llama_print_timings:        eval time =    7493.90 ms /    28 runs   (  267.64
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   29905.84 ms /   129 tokens
No. of rows:  99%| | 1245/1258 [11:27:28<07:28, 34Llama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.23 ms /    38 runs   (    0.27
ms per token,  3715.29 tokens per second)
llama_print_timings: prompt eval time =   22700.23 ms /   103 tokens (  220.39
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =    9958.15 ms /    37 runs   (  269.14
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   32805.38 ms /   140 tokens
No. of rows:  99%| | 1246/1258 [11:28:01<06:48, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.20 ms /    41 runs   (    0.27
ms per token,  3662.02 tokens per second)
llama_print_timings: prompt eval time =   26459.39 ms /   120 tokens (  220.49
ms per token,     4.54 tokens per second)
llama_print_timings:        eval time =   10816.23 ms /    40 runs   (  270.41
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   37437.43 ms /   160 tokens
No. of rows:  99%| | 1247/1258 [11:28:38<06:25, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.66 ms /    39 runs   (    0.27
ms per token,  3657.85 tokens per second)
llama_print_timings: prompt eval time =   23765.32 ms /   110 tokens (  216.05
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   10313.53 ms /    38 runs   (  271.41
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   34231.42 ms /   148 tokens
No. of rows:  99%| | 1248/1258 [11:29:13<05:48, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.08 ms /    38 runs   (    0.27
ms per token,  3769.09 tokens per second)
llama_print_timings: prompt eval time =   25661.27 ms /   117 tokens (  219.33
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    9992.34 ms /    37 runs   (  270.06
ms per token,     3.70 tokens per second)
llama_print_timings:       total time =   35799.97 ms /   154 tokens
No. of rows:  99%| | 1249/1258 [11:29:48<05:15, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      10.49 ms /    39 runs   (    0.27
```

```
ms per token,  3716.41 tokens per second)
llama_print_timings: prompt eval time =   24314.33 ms /   110 tokens (  221.04
ms per token,     4.52 tokens per second)
llama_print_timings:        eval time =   10544.64 ms /    38 runs   (  277.49
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   35017.65 ms /   148 tokens
No. of rows:  99%| | 1250/1258 [11:30:23<04:40, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.27 ms /    30 runs   (    0.28
ms per token,  3628.89 tokens per second)
llama_print_timings: prompt eval time =   19281.83 ms /    88 tokens (  219.11
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =    7758.55 ms /    29 runs   (  267.54
ms per token,     3.74 tokens per second)
llama_print_timings:       total time =   27156.90 ms /   117 tokens
No. of rows:  99%| | 1251/1258 [11:30:51<03:48, 32Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.25 ms /    50 runs   (    0.27
ms per token,  3773.30 tokens per second)
llama_print_timings: prompt eval time =   26294.00 ms /   120 tokens (  219.12
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   13683.28 ms /    49 runs   (  279.25
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =   40179.18 ms /   169 tokens
No. of rows: 100%| | 1252/1258 [11:31:31<03:29, 34Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      13.02 ms /    45 runs   (    0.29
ms per token,  3457.28 tokens per second)
llama_print_timings: prompt eval time =   27233.41 ms /   126 tokens (  216.14
ms per token,     4.63 tokens per second)
llama_print_timings:        eval time =   12408.40 ms /    44 runs   (  282.01
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   39824.78 ms /   170 tokens
No. of rows: 100%| | 1253/1258 [11:32:11<03:02, 36Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.84 ms /    44 runs   (    0.27
ms per token,  3716.53 tokens per second)
llama_print_timings: prompt eval time =   29627.82 ms /   137 tokens (  216.26
ms per token,     4.62 tokens per second)
llama_print_timings:        eval time =   11700.71 ms /    43 runs   (  272.11
```

```
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   41500.16 ms /   180 tokens
No. of rows: 100%| | 1254/1258 [11:32:52<02:31, 37Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      12.88 ms /    47 runs   (    0.27
ms per token,  3649.35 tokens per second)
llama_print_timings: prompt eval time =   34727.95 ms /   154 tokens (  225.51
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   12508.47 ms /    46 runs   (  271.92
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   47423.53 ms /   200 tokens
No. of rows: 100%| | 1255/1258 [11:33:40<02:02, 40Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =      11.70 ms /    44 runs   (    0.27
ms per token,  3759.72 tokens per second)
llama_print_timings: prompt eval time =   24201.69 ms /   108 tokens (  224.09
ms per token,     4.46 tokens per second)
llama_print_timings:        eval time =   11770.57 ms /    43 runs   (  273.73
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   36153.58 ms /   151 tokens
No. of rows: 100%| | 1256/1258 [11:34:16<01:18, 39Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       5.72 ms /    21 runs   (    0.27
ms per token,  3673.26 tokens per second)
llama_print_timings: prompt eval time =   21592.59 ms /    96 tokens (  224.92
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =    5363.90 ms /    20 runs   (  268.20
ms per token,     3.73 tokens per second)
llama_print_timings:       total time =   27036.60 ms /   116 tokens
No. of rows: 100%| | 1257/1258 [11:34:43<00:35, 35Llama.generate: prefix-match
hit

llama_print_timings:        load time =  138797.67 ms
llama_print_timings:      sample time =       8.48 ms /    32 runs   (    0.27
ms per token,  3773.14 tokens per second)
llama_print_timings: prompt eval time =   19116.99 ms /    85 tokens (  224.91
ms per token,     4.45 tokens per second)
llama_print_timings:        eval time =    8366.81 ms /    31 runs   (  269.90
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   27606.89 ms /   116 tokens
No. of rows: 100%| | 1258/1258 [11:35:10<00:00, 33
```

[120]:
```
hard_a = df_a[df_a['difficulty']=='hard']
```

[121]:
```
hard_b = df_b[df_b['difficulty']=='hard']
```

[125]:
```
hard_questions = pd.concat([hard_a, hard_b])
```

[128]:
```
from datetime import datetime

def predict(df, llm, out):
    for point in tqdm(df.iloc, desc="No. of rows", total=df.shape[0]):
        start = datetime.now()
        prompt = generate_test_prompt(point)
        out['prompt'].append(prompt)
        out['actu'].append(point['answer'])
        result = llm(prompt=prompt,
            max_tokens = 150,
            temperature = 0.2,
            stop = ['```'])
        answer = result
        end = datetime.now()
        out['inf_time'].append((end - start).total_seconds())
        out['pred'].append(answer['choices'][0]['text'].strip())
        out['temperature'].append(0.2)
        out['difficulty'].append(point['difficulty'])
        out['token_in'].append(result['usage']['prompt_tokens'])
        out['token_out'].append(result['usage']['completion_tokens']+1)
        out['tokens_per_sec'].append(result['usage']['completion_tokens']/((end␣
 ↪- start).total_seconds()))
    return out
```

[129]:
```
from llama_cpp import Llama

phi2 = Llama(model_path="../phi2_sqlcoder_f16.gguf")
out = {"prompt": [], "pred": [], "actu": [], "inf_time": [], "temperature": [],␣
 ↪"difficulty": [], "token_in": [], "token_out": [], "tokens_per_sec": []}

out = predict(hard_questions, phi2, out)
json.dump(out, open("phi2_eval_hard.json", "w"))
```

```
llama_model_loader: loaded meta data with 19 key-value pairs and 453 tensors
from ../phi2_sqlcoder_f16.gguf (version GGUF V3 (latest))
llama_model_loader: Dumping metadata keys/values. Note: KV overrides do not
apply in this output.
llama_model_loader: - kv   0:                       general.architecture str
= phi2
llama_model_loader: - kv   1:                               general.name str
= Phi2
llama_model_loader: - kv   2:                      phi2.context_length u32
```

```
= 2048
llama_model_loader: - kv   3:                      phi2.embedding_length u32
= 2560
llama_model_loader: - kv   4:                   phi2.feed_forward_length u32
= 10240
llama_model_loader: - kv   5:                           phi2.block_count u32
= 32
llama_model_loader: - kv   6:                  phi2.attention.head_count u32
= 32
llama_model_loader: - kv   7:               phi2.attention.head_count_kv u32
= 32
llama_model_loader: - kv   8:          phi2.attention.layer_norm_epsilon f32
= 0.000010
llama_model_loader: - kv   9:                  phi2.rope.dimension_count u32
= 32
llama_model_loader: - kv  10:                          general.file_type u32
= 1
llama_model_loader: - kv  11:               tokenizer.ggml.add_bos_token bool
= false
llama_model_loader: - kv  12:                       tokenizer.ggml.model str
= gpt2
llama_model_loader: - kv  13:                      tokenizer.ggml.tokens
arr[str,51200]   = ["!", "\"", "#", "$", "%", "&", "'", …
llama_model_loader: - kv  14:                  tokenizer.ggml.token_type
arr[i32,51200]   = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
llama_model_loader: - kv  15:                      tokenizer.ggml.merges
arr[str,50000]   = ["Ġ t", "Ġ a", "h e", "i n", "r e",…
llama_model_loader: - kv  16:                tokenizer.ggml.bos_token_id u32
= 50256
llama_model_loader: - kv  17:                tokenizer.ggml.eos_token_id u32
= 50256
llama_model_loader: - kv  18:            tokenizer.ggml.unknown_token_id u32
= 50256
llama_model_loader: - type  f32:  259 tensors
llama_model_loader: - type  f16:  194 tensors
llm_load_vocab: mismatch in special tokens definition ( 910/51200 vs 944/51200
).
llm_load_print_meta: format           = GGUF V3 (latest)
llm_load_print_meta: arch             = phi2
llm_load_print_meta: vocab type       = BPE
llm_load_print_meta: n_vocab          = 51200
llm_load_print_meta: n_merges         = 50000
llm_load_print_meta: n_ctx_train      = 2048
llm_load_print_meta: n_embd           = 2560
llm_load_print_meta: n_head           = 32
llm_load_print_meta: n_head_kv        = 32
llm_load_print_meta: n_layer          = 32
llm_load_print_meta: n_rot            = 32
```

```
llm_load_print_meta: n_embd_head_k    = 80
llm_load_print_meta: n_embd_head_v    = 80
llm_load_print_meta: n_gqa            = 1
llm_load_print_meta: n_embd_k_gqa     = 2560
llm_load_print_meta: n_embd_v_gqa     = 2560
llm_load_print_meta: f_norm_eps       = 1.0e-05
llm_load_print_meta: f_norm_rms_eps   = 0.0e+00
llm_load_print_meta: f_clamp_kqv      = 0.0e+00
llm_load_print_meta: f_max_alibi_bias = 0.0e+00
llm_load_print_meta: n_ff             = 10240
llm_load_print_meta: n_expert         = 0
llm_load_print_meta: n_expert_used    = 0
llm_load_print_meta: rope scaling     = linear
llm_load_print_meta: freq_base_train  = 10000.0
llm_load_print_meta: freq_scale_train = 1
llm_load_print_meta: n_yarn_orig_ctx  = 2048
llm_load_print_meta: rope_finetuned   = unknown
llm_load_print_meta: model type       = 3B
llm_load_print_meta: model ftype      = F16
llm_load_print_meta: model params     = 2.78 B
llm_load_print_meta: model size       = 5.18 GiB (16.01 BPW)
llm_load_print_meta: general.name     = Phi2
llm_load_print_meta: BOS token        = 50256 '<|endoftext|>'
llm_load_print_meta: EOS token        = 50256 '<|endoftext|>'
llm_load_print_meta: UNK token        = 50256 '<|endoftext|>'
llm_load_print_meta: LF token         = 30 '?'
llm_load_tensors: ggml ctx size =    0.17 MiB
llm_load_tensors:        CPU buffer size =  5303.65 MiB
…
…
llama_new_context_with_model: n_ctx      = 512
llama_new_context_with_model: freq_base  = 10000.0
llama_new_context_with_model: freq_scale = 1
llama_kv_cache_init:        CPU KV buffer size =   160.00 MiB
llama_new_context_with_model: KV self size  =  160.00 MiB, K (f16):   80.00 MiB,
V (f16):   80.00 MiB
llama_new_context_with_model:        CPU input buffer size   =     7.01 MiB
llama_new_context_with_model:        CPU compute buffer size =   105.00 MiB
llama_new_context_with_model: graph splits (measure): 1
AVX = 1 | AVX_VNNI = 0 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI =
0 | FMA = 1 | NEON = 0 | ARM_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 |
BLAS = 0 | SSE3 = 1 | SSSE3 = 0 | VSX = 0 | MATMUL_INT8 = 0 |
Model metadata: {'general.architecture': 'phi2', 'phi2.context_length': '2048',
'general.name': 'Phi2', 'phi2.attention.head_count_kv': '32',
'phi2.embedding_length': '2560', 'tokenizer.ggml.add_bos_token': 'false',
'phi2.feed_forward_length': '10240', 'tokenizer.ggml.bos_token_id': '50256',
'phi2.block_count': '32', 'phi2.attention.head_count': '32',
'phi2.attention.layer_norm_epsilon': '0.000010', 'phi2.rope.dimension_count':
```

```
'32', 'tokenizer.ggml.eos_token_id': '50256', 'general.file_type': '1',
'tokenizer.ggml.model': 'gpt2', 'tokenizer.ggml.unknown_token_id': '50256'}
No. of rows:   0%|          | 0/33 [00:00<?, ?it/s]
llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      31.54 ms /    90 runs   (    0.35
ms per token,  2853.25 tokens per second)
llama_print_timings: prompt eval time =  169140.77 ms /   155 tokens ( 1091.23
ms per token,     0.92 tokens per second)
llama_print_timings:        eval time =   24856.70 ms /    89 runs   (  279.29
ms per token,     3.58 tokens per second)
llama_print_timings:       total time =  194532.74 ms /   244 tokens
No. of rows:   3%| | 1/33 [03:14<1:43:45, 194.54s/Llama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      34.47 ms /    89 runs   (    0.39
ms per token,  2581.96 tokens per second)
llama_print_timings: prompt eval time =   10845.78 ms /   111 tokens (   97.71
ms per token,    10.23 tokens per second)
llama_print_timings:        eval time =   23993.07 ms /    88 runs   (  272.65
ms per token,     3.67 tokens per second)
llama_print_timings:       total time =   35406.45 ms /   199 tokens
No. of rows:   6%| | 2/33 [03:49<52:09, 100.94s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      29.35 ms /    76 runs   (    0.39
ms per token,  2589.44 tokens per second)
llama_print_timings: prompt eval time =   13463.19 ms /   139 tokens (   96.86
ms per token,    10.32 tokens per second)
llama_print_timings:        eval time =   17961.08 ms /    75 runs   (  239.48
ms per token,     4.18 tokens per second)
llama_print_timings:       total time =   31892.35 ms /   214 tokens
No. of rows:   9%| | 3/33 [04:21<34:42, 69.41s/it]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      34.50 ms /    85 runs   (    0.41
ms per token,  2463.77 tokens per second)
llama_print_timings: prompt eval time =   14304.65 ms /   143 tokens (  100.03
ms per token,    10.00 tokens per second)
llama_print_timings:        eval time =   21410.87 ms /    84 runs   (  254.89
ms per token,     3.92 tokens per second)
llama_print_timings:       total time =   36260.21 ms /   227 tokens
No. of rows:  12%| | 4/33 [04:58<27:13, 56.33s/it]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
```

```
llama_print_timings:      sample time =      30.94 ms /    77 runs   (    0.40
ms per token,  2488.45 tokens per second)
llama_print_timings: prompt eval time =   12879.24 ms /   130 tokens (   99.07
ms per token,    10.09 tokens per second)
llama_print_timings:        eval time =   19891.50 ms /    76 runs   (  261.73
ms per token,     3.82 tokens per second)
llama_print_timings:       total time =   33267.99 ms /   206 tokens
No. of rows:  15%| | 5/33 [05:31<22:24, 48.02s/it]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      34.68 ms /    79 runs   (    0.44
ms per token,  2278.10 tokens per second)
llama_print_timings: prompt eval time =   14451.28 ms /   137 tokens (  105.48
ms per token,     9.48 tokens per second)
llama_print_timings:        eval time =   18343.13 ms /    78 runs   (  235.17
ms per token,     4.25 tokens per second)
llama_print_timings:       total time =   33322.26 ms /   215 tokens
No. of rows:  18%| | 6/33 [06:04<19:21, 43.03s/it]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      30.02 ms /    80 runs   (    0.38
ms per token,  2664.54 tokens per second)
llama_print_timings: prompt eval time =   15275.78 ms /   151 tokens (  101.16
ms per token,     9.88 tokens per second)
llama_print_timings:        eval time =   18389.70 ms /    79 runs   (  232.78
ms per token,     4.30 tokens per second)
llama_print_timings:       total time =   34143.67 ms /   230 tokens
No. of rows:  21%| | 7/33 [06:38<17:23, 40.12s/it]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      27.24 ms /    74 runs   (    0.37
ms per token,  2716.29 tokens per second)
llama_print_timings: prompt eval time =   13190.73 ms /   136 tokens (   96.99
ms per token,    10.31 tokens per second)
llama_print_timings:        eval time =   17155.57 ms /    73 runs   (  235.01
ms per token,     4.26 tokens per second)
llama_print_timings:       total time =   30783.23 ms /   209 tokens
No. of rows:  24%| | 8/33 [07:09<15:28, 37.15s/it]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      16.26 ms /    42 runs   (    0.39
ms per token,  2583.18 tokens per second)
llama_print_timings: prompt eval time =   13167.98 ms /   138 tokens (   95.42
ms per token,    10.48 tokens per second)
```

```
llama_print_timings:        eval time =   10630.61 ms /    41 runs   (  259.28
ms per token,     3.86 tokens per second)
llama_print_timings:       total time =   24052.68 ms /   179 tokens
No. of rows:  27%| | 9/33 [07:33<13:13, 33.06s/it]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      28.58 ms /    77 runs   (    0.37
ms per token,  2694.10 tokens per second)
llama_print_timings: prompt eval time =   12304.42 ms /   124 tokens (   99.23
ms per token,    10.08 tokens per second)
llama_print_timings:        eval time =   16370.76 ms /    76 runs   (  215.40
ms per token,     4.64 tokens per second)
llama_print_timings:       total time =   29150.01 ms /   200 tokens
No. of rows:  30%| | 10/33 [08:02<12:12, 31.85s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      27.07 ms /    73 runs   (    0.37
ms per token,  2696.81 tokens per second)
llama_print_timings: prompt eval time =   13800.20 ms /   136 tokens (  101.47
ms per token,     9.85 tokens per second)
llama_print_timings:        eval time =   17283.54 ms /    72 runs   (  240.05
ms per token,     4.17 tokens per second)
llama_print_timings:       total time =   31529.61 ms /   208 tokens
No. of rows:  33%| | 11/33 [08:34<11:38, 31.76s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      49.81 ms /   123 runs   (    0.40
ms per token,  2469.14 tokens per second)
llama_print_timings: prompt eval time =   13044.39 ms /   133 tokens (   98.08
ms per token,    10.20 tokens per second)
llama_print_timings:        eval time =   29124.69 ms /   122 runs   (  238.73
ms per token,     4.19 tokens per second)
llama_print_timings:       total time =   42948.15 ms /   255 tokens
No. of rows:  36%| | 12/33 [09:17<12:18, 35.16s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      27.44 ms /    63 runs   (    0.44
ms per token,  2296.25 tokens per second)
llama_print_timings: prompt eval time =   11692.41 ms /   122 tokens (   95.84
ms per token,    10.43 tokens per second)
llama_print_timings:        eval time =   13989.78 ms /    62 runs   (  225.64
ms per token,     4.43 tokens per second)
llama_print_timings:       total time =   26092.43 ms /   184 tokens
No. of rows:  39%| | 13/33 [09:43<10:48, 32.41s/itLlama.generate: prefix-match
```

```
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      26.07 ms /    68 runs   (    0.38
ms per token,  2608.56 tokens per second)
llama_print_timings: prompt eval time =   15271.74 ms /   144 tokens (  106.05
ms per token,     9.43 tokens per second)
llama_print_timings:        eval time =   15228.60 ms /    67 runs   (  227.29
ms per token,     4.40 tokens per second)
llama_print_timings:       total time =   30923.15 ms /   211 tokens
No. of rows:  42%| | 14/33 [10:14<10:07, 31.97s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      28.95 ms /    74 runs   (    0.39
ms per token,  2556.04 tokens per second)
llama_print_timings: prompt eval time =   14320.49 ms /   137 tokens (  104.53
ms per token,     9.57 tokens per second)
llama_print_timings:        eval time =   16135.07 ms /    73 runs   (  221.03
ms per token,     4.52 tokens per second)
llama_print_timings:       total time =   30916.19 ms /   210 tokens
No. of rows:  45%| | 15/33 [10:45<09:29, 31.66s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      29.08 ms /    68 runs   (    0.43
ms per token,  2338.46 tokens per second)
llama_print_timings: prompt eval time =   16128.31 ms /   137 tokens (  117.72
ms per token,     8.49 tokens per second)
llama_print_timings:        eval time =   16832.34 ms /    67 runs   (  251.23
ms per token,     3.98 tokens per second)
llama_print_timings:       total time =   33390.19 ms /   204 tokens
No. of rows:  48%| | 16/33 [11:18<09:07, 32.18s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      25.67 ms /    65 runs   (    0.39
ms per token,  2531.94 tokens per second)
llama_print_timings: prompt eval time =   14652.41 ms /   139 tokens (  105.41
ms per token,     9.49 tokens per second)
llama_print_timings:        eval time =   14395.52 ms /    64 runs   (  224.93
ms per token,     4.45 tokens per second)
llama_print_timings:       total time =   29453.17 ms /   203 tokens
No. of rows:  52%| | 17/33 [11:48<08:21, 31.36s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      37.70 ms /    94 runs   (    0.40
```

```
ms per token,  2493.30 tokens per second)
llama_print_timings: prompt eval time =   17434.54 ms /   165 tokens (  105.66
ms per token,     9.46 tokens per second)
llama_print_timings:        eval time =   20786.34 ms /    93 runs   (  223.51
ms per token,     4.47 tokens per second)
llama_print_timings:       total time =   38823.35 ms /   258 tokens
No. of rows:  55%| | 18/33 [12:27<08:24, 33.60s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      29.19 ms /    78 runs   (    0.37
ms per token,  2672.24 tokens per second)
llama_print_timings: prompt eval time =   15100.43 ms /   128 tokens (  117.97
ms per token,     8.48 tokens per second)
llama_print_timings:        eval time =   18452.31 ms /    77 runs   (  239.64
ms per token,     4.17 tokens per second)
llama_print_timings:       total time =   34037.98 ms /   205 tokens
No. of rows:  58%| | 19/33 [13:01<07:52, 33.74s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      59.25 ms /   150 runs   (    0.40
ms per token,  2531.60 tokens per second)
llama_print_timings: prompt eval time =   17127.14 ms /   146 tokens (  117.31
ms per token,     8.52 tokens per second)
llama_print_timings:        eval time =   35632.43 ms /   149 runs   (  239.14
ms per token,     4.18 tokens per second)
llama_print_timings:       total time =   53720.51 ms /   295 tokens
No. of rows:  61%| | 20/33 [13:54<08:36, 39.74s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      48.94 ms /   125 runs   (    0.39
ms per token,  2554.20 tokens per second)
llama_print_timings: prompt eval time =   14220.96 ms /   128 tokens (  111.10
ms per token,     9.00 tokens per second)
llama_print_timings:        eval time =   30606.01 ms /   124 runs   (  246.82
ms per token,     4.05 tokens per second)
llama_print_timings:       total time =   45631.66 ms /   252 tokens
No. of rows:  64%| | 21/33 [14:40<08:18, 41.51s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      31.66 ms /    82 runs   (    0.39
ms per token,  2590.35 tokens per second)
llama_print_timings: prompt eval time =   12931.47 ms /   120 tokens (  107.76
ms per token,     9.28 tokens per second)
llama_print_timings:        eval time =   18729.68 ms /    81 runs   (  231.23
```

```
ms per token,     4.32 tokens per second)
llama_print_timings:       total time =   32171.18 ms /   201 tokens
No. of rows:  67%| | 22/33 [15:12<07:05, 38.71s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      51.34 ms /   124 runs   (    0.41
ms per token,  2415.41 tokens per second)
llama_print_timings: prompt eval time =   13637.01 ms /   127 tokens (  107.38
ms per token,     9.31 tokens per second)
llama_print_timings:        eval time =   28500.90 ms /   123 runs   (  231.71
ms per token,     4.32 tokens per second)
llama_print_timings:       total time =   42955.29 ms /   250 tokens
No. of rows:  70%| | 23/33 [15:55<06:39, 39.99s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      28.39 ms /    63 runs   (    0.45
ms per token,  2219.33 tokens per second)
llama_print_timings: prompt eval time =   16822.57 ms /   151 tokens (  111.41
ms per token,     8.98 tokens per second)
llama_print_timings:        eval time =   17085.87 ms /    62 runs   (  275.58
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   34353.04 ms /   213 tokens
No. of rows:  73%| | 24/33 [16:29<05:44, 38.30s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      44.36 ms /   109 runs   (    0.41
ms per token,  2457.45 tokens per second)
llama_print_timings: prompt eval time =   17581.26 ms /   152 tokens (  115.67
ms per token,     8.65 tokens per second)
llama_print_timings:        eval time =   26295.28 ms /   108 runs   (  243.47
ms per token,     4.11 tokens per second)
llama_print_timings:       total time =   44605.76 ms /   260 tokens
No. of rows:  76%| | 25/33 [17:14<05:21, 40.19s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      37.28 ms /    93 runs   (    0.40
ms per token,  2494.84 tokens per second)
llama_print_timings: prompt eval time =   17719.12 ms /   153 tokens (  115.81
ms per token,     8.63 tokens per second)
llama_print_timings:        eval time =   21758.52 ms /    92 runs   (  236.51
ms per token,     4.23 tokens per second)
llama_print_timings:       total time =   40095.20 ms /   245 tokens
No. of rows:  79%| | 26/33 [17:54<04:41, 40.17s/itLlama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      36.78 ms /    87 runs   (    0.42
ms per token,  2365.54 tokens per second)
llama_print_timings: prompt eval time =   14747.40 ms /   133 tokens (  110.88
ms per token,     9.02 tokens per second)
llama_print_timings:        eval time =   21555.21 ms /    86 runs   (  250.64
ms per token,     3.99 tokens per second)
llama_print_timings:       total time =   36874.88 ms /   219 tokens
No. of rows:  82%| | 27/33 [18:31<03:55, 39.18s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      37.33 ms /    92 runs   (    0.41
ms per token,  2464.51 tokens per second)
llama_print_timings: prompt eval time =   17125.31 ms /   150 tokens (  114.17
ms per token,     8.76 tokens per second)
llama_print_timings:        eval time =   20921.08 ms /    91 runs   (  229.90
ms per token,     4.35 tokens per second)
llama_print_timings:       total time =   38636.71 ms /   241 tokens
No. of rows:  85%| | 28/33 [19:10<03:15, 39.02s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      64.85 ms /   150 runs   (    0.43
ms per token,  2313.07 tokens per second)
llama_print_timings: prompt eval time =   18801.26 ms /   153 tokens (  122.88
ms per token,     8.14 tokens per second)
llama_print_timings:        eval time =   35184.34 ms /   149 runs   (  236.14
ms per token,     4.23 tokens per second)
llama_print_timings:       total time =   55034.18 ms /   302 tokens
No. of rows:  88%| | 29/33 [20:05<02:55, 43.82s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      23.43 ms /    61 runs   (    0.38
ms per token,  2603.83 tokens per second)
llama_print_timings: prompt eval time =   14200.33 ms /   120 tokens (  118.34
ms per token,     8.45 tokens per second)
llama_print_timings:        eval time =   15583.38 ms /    60 runs   (  259.72
ms per token,     3.85 tokens per second)
llama_print_timings:       total time =   30185.27 ms /   180 tokens
No. of rows:  91%| | 30/33 [20:35<01:59, 39.74s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      33.89 ms /    83 runs   (    0.41
ms per token,  2449.39 tokens per second)
```

```
llama_print_timings: prompt eval time =   18306.15 ms /   154 tokens (  118.87
ms per token,     8.41 tokens per second)
llama_print_timings:        eval time =   20821.08 ms /    82 runs   (  253.92
ms per token,     3.94 tokens per second)
llama_print_timings:       total time =   39681.37 ms /   236 tokens
No. of rows:  94%|▉| 31/33 [21:15<01:19, 39.73s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      30.60 ms /    76 runs   (    0.40
ms per token,  2483.58 tokens per second)
llama_print_timings: prompt eval time =   15826.00 ms /   141 tokens (  112.24
ms per token,     8.91 tokens per second)
llama_print_timings:        eval time =   17346.39 ms /    75 runs   (  231.29
ms per token,     4.32 tokens per second)
llama_print_timings:       total time =   33667.04 ms /   216 tokens
No. of rows:  97%|▉| 32/33 [21:48<00:37, 37.91s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  169141.40 ms
llama_print_timings:      sample time =      33.79 ms /    77 runs   (    0.44
ms per token,  2279.05 tokens per second)
llama_print_timings: prompt eval time =   15279.51 ms /   130 tokens (  117.53
ms per token,     8.51 tokens per second)
llama_print_timings:        eval time =   18312.65 ms /    76 runs   (  240.96
ms per token,     4.15 tokens per second)
llama_print_timings:       total time =   34132.57 ms /   206 tokens
No. of rows: 100%|▉| 33/33 [22:22<00:00, 40.70s/it
```

[130]:
```
sqlc = Llama(model_path="../sqlcoder-7b-q5_k_m.gguf")
out = {"prompt": [], "pred": [], "actu": [], "inf_time": [], "temperature": [],␣
 ↪"difficulty": [], "token_in": [], "token_out": [], "tokens_per_sec": []}

out = predict(hard_questions, sqlc, out)
json.dump(out, open("sqlc_eval_hard.json", "w"))
```

```
llama_model_loader: loaded meta data with 22 key-value pairs and 291 tensors
from ../sqlcoder-7b-q5_k_m.gguf (version GGUF V3 (latest))
llama_model_loader: Dumping metadata keys/values. Note: KV overrides do not
apply in this output.
llama_model_loader: - kv   0:                       general.architecture str
= llama
llama_model_loader: - kv   1:                               general.name str
= .
llama_model_loader: - kv   2:                      llama.context_length u32
= 16384
llama_model_loader: - kv   3:                    llama.embedding_length u32
= 4096
```

```
llama_model_loader: - kv   4:                          llama.block_count u32
= 32
llama_model_loader: - kv   5:                  llama.feed_forward_length u32
= 11008
llama_model_loader: - kv   6:                 llama.rope.dimension_count u32
= 128
llama_model_loader: - kv   7:                 llama.attention.head_count u32
= 32
llama_model_loader: - kv   8:              llama.attention.head_count_kv u32
= 32
llama_model_loader: - kv   9:     llama.attention.layer_norm_rms_epsilon f32
= 0.000010
llama_model_loader: - kv  10:                       llama.rope.freq_base f32
= 1000000.000000
llama_model_loader: - kv  11:                          general.file_type u32
= 17
llama_model_loader: - kv  12:                       tokenizer.ggml.model str
= llama
llama_model_loader: - kv  13:                      tokenizer.ggml.tokens
arr[str,32016]   = ["<unk>", "<s>", "</s>", "<0x00>", "<…
llama_model_loader: - kv  14:                      tokenizer.ggml.scores
arr[f32,32016]   = [0.000000, 0.000000, 0.000000, 0.0000…
llama_model_loader: - kv  15:                  tokenizer.ggml.token_type
arr[i32,32016]   = [2, 3, 3, 6, 6, 6, 6, 6, 6, 6, 6, 6, …
llama_model_loader: - kv  16:                tokenizer.ggml.bos_token_id u32
= 1
llama_model_loader: - kv  17:                tokenizer.ggml.eos_token_id u32
= 2
llama_model_loader: - kv  18:            tokenizer.ggml.unknown_token_id u32
= 0
llama_model_loader: - kv  19:               tokenizer.ggml.add_bos_token bool
= true
llama_model_loader: - kv  20:               tokenizer.ggml.add_eos_token bool
= false
llama_model_loader: - kv  21:               general.quantization_version u32
= 2
llama_model_loader: - type  f32:   65 tensors
llama_model_loader: - type q5_K:  193 tensors
llama_model_loader: - type q6_K:   33 tensors
llm_load_vocab: mismatch in special tokens definition ( 264/32016 vs 259/32016
).
llm_load_print_meta: format           = GGUF V3 (latest)
llm_load_print_meta: arch             = llama
llm_load_print_meta: vocab type       = SPM
llm_load_print_meta: n_vocab          = 32016
llm_load_print_meta: n_merges         = 0
llm_load_print_meta: n_ctx_train      = 16384
llm_load_print_meta: n_embd           = 4096
```

```
llm_load_print_meta: n_head           = 32
llm_load_print_meta: n_head_kv        = 32
llm_load_print_meta: n_layer          = 32
llm_load_print_meta: n_rot            = 128
llm_load_print_meta: n_embd_head_k    = 128
llm_load_print_meta: n_embd_head_v    = 128
llm_load_print_meta: n_gqa            = 1
llm_load_print_meta: n_embd_k_gqa     = 4096
llm_load_print_meta: n_embd_v_gqa     = 4096
llm_load_print_meta: f_norm_eps       = 0.0e+00
llm_load_print_meta: f_norm_rms_eps   = 1.0e-05
llm_load_print_meta: f_clamp_kqv      = 0.0e+00
llm_load_print_meta: f_max_alibi_bias = 0.0e+00
llm_load_print_meta: n_ff             = 11008
llm_load_print_meta: n_expert         = 0
llm_load_print_meta: n_expert_used    = 0
llm_load_print_meta: rope scaling     = linear
llm_load_print_meta: freq_base_train  = 1000000.0
llm_load_print_meta: freq_scale_train = 1
llm_load_print_meta: n_yarn_orig_ctx  = 16384
llm_load_print_meta: rope_finetuned   = unknown
llm_load_print_meta: model type       = 7B
llm_load_print_meta: model ftype      = Q5_K - Medium
llm_load_print_meta: model params     = 6.74 B
llm_load_print_meta: model size       = 4.45 GiB (5.68 BPW)
llm_load_print_meta: general.name     = .
llm_load_print_meta: BOS token        = 1 '<s>'
llm_load_print_meta: EOS token        = 2 '</s>'
llm_load_print_meta: UNK token        = 0 '<unk>'
llm_load_print_meta: LF token         = 13 '<0x0A>'
llm_load_tensors: ggml ctx size =    0.11 MiB
llm_load_tensors:        CPU buffer size =  4560.96 MiB
…
…
llama_new_context_with_model: n_ctx      = 512
llama_new_context_with_model: freq_base  = 1000000.0
llama_new_context_with_model: freq_scale = 1
llama_kv_cache_init:        CPU KV buffer size =   256.00 MiB
llama_new_context_with_model: KV self size  =  256.00 MiB, K (f16):  128.00 MiB,
V (f16):  128.00 MiB
llama_new_context_with_model:        CPU input buffer size   =    10.01 MiB
llama_new_context_with_model:        CPU compute buffer size =    70.53 MiB
llama_new_context_with_model: graph splits (measure): 1
AVX = 1 | AVX_VNNI = 0 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI =
0 | FMA = 1 | NEON = 0 | ARM_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 |
BLAS = 0 | SSE3 = 1 | SSSE3 = 0 | VSX = 0 | MATMUL_INT8 = 0 |
Model metadata: {'general.name': '.', 'general.architecture': 'llama',
'llama.context_length': '16384', 'llama.rope.dimension_count': '128',
```

```
'llama.embedding_length': '4096', 'llama.block_count': '32',
'llama.feed_forward_length': '11008', 'llama.attention.head_count': '32',
'tokenizer.ggml.eos_token_id': '2', 'general.file_type': '17',
'llama.attention.head_count_kv': '32', 'llama.attention.layer_norm_rms_epsilon':
'0.000010', 'llama.rope.freq_base': '1000000.000000', 'tokenizer.ggml.model':
'llama', 'general.quantization_version': '2', 'tokenizer.ggml.bos_token_id':
'1', 'tokenizer.ggml.unknown_token_id': '0', 'tokenizer.ggml.add_bos_token':
'true', 'tokenizer.ggml.add_eos_token': 'false'}
No. of rows:   0%|          | 0/33 [00:00<?, ?it/s]
llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      15.25 ms /    64 runs   (    0.24
ms per token,  4198.10 tokens per second)
llama_print_timings: prompt eval time =  441235.51 ms /   145 tokens ( 3043.00
ms per token,     0.33 tokens per second)
llama_print_timings:        eval time =   26715.97 ms /    63 runs   (  424.06
ms per token,     2.36 tokens per second)
llama_print_timings:       total time =  468266.67 ms /   208 tokens
No. of rows:   3%| | 1/33 [07:48<4:09:45, 468.30s/Llama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      21.51 ms /    89 runs   (    0.24
ms per token,  4137.80 tokens per second)
llama_print_timings: prompt eval time =   19309.73 ms /   102 tokens (  189.31
ms per token,     5.28 tokens per second)
llama_print_timings:        eval time =   23741.95 ms /    88 runs   (  269.79
ms per token,     3.71 tokens per second)
llama_print_timings:       total time =   43371.86 ms /   190 tokens
No. of rows:   6%| | 2/33 [08:31<1:52:48, 218.35s/Llama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      17.64 ms /    73 runs   (    0.24
ms per token,  4138.79 tokens per second)
llama_print_timings: prompt eval time =   26139.63 ms /   136 tokens (  192.20
ms per token,     5.20 tokens per second)
llama_print_timings:        eval time =   20075.60 ms /    72 runs   (  278.83
ms per token,     3.59 tokens per second)
llama_print_timings:       total time =   46483.98 ms /   208 tokens
No. of rows:   9%| | 3/33 [09:18<1:09:56, 139.87s/Llama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      16.60 ms /    65 runs   (    0.26
ms per token,  3915.66 tokens per second)
llama_print_timings: prompt eval time =   25537.64 ms /   125 tokens (  204.30
ms per token,     4.89 tokens per second)
llama_print_timings:        eval time =   17027.09 ms /    64 runs   (  266.05
```

```
ms per token,     3.76 tokens per second)
llama_print_timings:       total time =   42812.80 ms /   189 tokens
No. of rows:  12%| | 4/33 [10:00<49:05, 101.56s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =       6.74 ms /    25 runs   (    0.27
ms per token,  3707.00 tokens per second)
llama_print_timings: prompt eval time =   24548.37 ms /   120 tokens (  204.57
ms per token,     4.89 tokens per second)
llama_print_timings:        eval time =    8629.67 ms /    24 runs   (  359.57
ms per token,     2.78 tokens per second)
llama_print_timings:       total time =   33273.58 ms /   144 tokens
No. of rows:  15%| | 5/33 [10:34<35:54, 76.94s/it]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      19.77 ms /    77 runs   (    0.26
ms per token,  3894.20 tokens per second)
llama_print_timings: prompt eval time =   27719.23 ms /   132 tokens (  209.99
ms per token,     4.76 tokens per second)
llama_print_timings:        eval time =   22421.98 ms /    76 runs   (  295.03
ms per token,     3.39 tokens per second)
llama_print_timings:       total time =   50440.30 ms /   208 tokens
No. of rows:  18%| | 6/33 [11:24<30:34, 67.93s/it]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      16.77 ms /    64 runs   (    0.26
ms per token,  3817.02 tokens per second)
llama_print_timings: prompt eval time =   29011.87 ms /   139 tokens (  208.72
ms per token,     4.79 tokens per second)
llama_print_timings:        eval time =   17086.46 ms /    63 runs   (  271.21
ms per token,     3.69 tokens per second)
llama_print_timings:       total time =   46347.39 ms /   202 tokens
No. of rows:  21%| | 7/33 [12:11<26:22, 60.88s/it]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      19.02 ms /    69 runs   (    0.28
ms per token,  3628.52 tokens per second)
llama_print_timings: prompt eval time =   25907.95 ms /   124 tokens (  208.94
ms per token,     4.79 tokens per second)
llama_print_timings:        eval time =   19839.88 ms /    68 runs   (  291.76
ms per token,     3.43 tokens per second)
llama_print_timings:       total time =   46023.67 ms /   192 tokens
No. of rows:  24%| | 8/33 [12:57<23:23, 56.15s/it]Llama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      15.36 ms /    58 runs   (    0.26
ms per token,  3775.30 tokens per second)
llama_print_timings: prompt eval time =   25706.73 ms /   123 tokens (  209.00
ms per token,     4.78 tokens per second)
llama_print_timings:        eval time =   17197.86 ms /    57 runs   (  301.72
ms per token,     3.31 tokens per second)
llama_print_timings:       total time =   43131.83 ms /   180 tokens
No. of rows:  27%| | 9/33 [13:40<20:50, 52.09s/it]Llama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      20.71 ms /    78 runs   (    0.27
ms per token,  3766.48 tokens per second)
llama_print_timings: prompt eval time =   26554.54 ms /   121 tokens (  219.46
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   22516.80 ms /    77 runs   (  292.43
ms per token,     3.42 tokens per second)
llama_print_timings:       total time =   49382.98 ms /   198 tokens
No. of rows:  30%| | 10/33 [14:29<19:38, 51.25s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      15.64 ms /    60 runs   (    0.26
ms per token,  3836.32 tokens per second)
llama_print_timings: prompt eval time =   26464.08 ms /   125 tokens (  211.71
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   17473.03 ms /    59 runs   (  296.15
ms per token,     3.38 tokens per second)
llama_print_timings:       total time =   44168.61 ms /   184 tokens
No. of rows:  33%| | 11/33 [15:13<17:59, 49.09s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      27.33 ms /    92 runs   (    0.30
ms per token,  3366.63 tokens per second)
llama_print_timings: prompt eval time =   31335.71 ms /   146 tokens (  214.63
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   25077.05 ms /    91 runs   (  275.57
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   56789.91 ms /   237 tokens
No. of rows:  36%| | 12/33 [16:10<18:00, 51.43s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      16.05 ms /    58 runs   (    0.28
ms per token,  3613.03 tokens per second)
```

```
llama_print_timings: prompt eval time =   23247.53 ms /   111 tokens (  209.44
ms per token,     4.77 tokens per second)
llama_print_timings:        eval time =   17610.55 ms /    57 runs   (  308.96
ms per token,     3.24 tokens per second)
llama_print_timings:       total time =   41086.66 ms /   168 tokens
No. of rows:  39%| | 13/33 [16:51<16:05, 48.30s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      13.80 ms /    53 runs   (    0.26
ms per token,  3841.69 tokens per second)
llama_print_timings: prompt eval time =   29381.21 ms /   137 tokens (  214.46
ms per token,     4.66 tokens per second)
llama_print_timings:        eval time =   14239.58 ms /    52 runs   (  273.84
ms per token,     3.65 tokens per second)
llama_print_timings:       total time =   43826.23 ms /   189 tokens
No. of rows:  42%| | 14/33 [17:35<14:52, 46.95s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      16.41 ms /    62 runs   (    0.26
ms per token,  3777.03 tokens per second)
llama_print_timings: prompt eval time =   25824.10 ms /   122 tokens (  211.67
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   16389.87 ms /    61 runs   (  268.69
ms per token,     3.72 tokens per second)
llama_print_timings:       total time =   42454.62 ms /   183 tokens
No. of rows:  45%| | 15/33 [18:18<13:40, 45.60s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      13.72 ms /    52 runs   (    0.26
ms per token,  3790.36 tokens per second)
llama_print_timings: prompt eval time =   26830.46 ms /   130 tokens (  206.39
ms per token,     4.85 tokens per second)
llama_print_timings:        eval time =   15225.16 ms /    51 runs   (  298.53
ms per token,     3.35 tokens per second)
llama_print_timings:       total time =   42257.20 ms /   181 tokens
No. of rows:  48%| | 16/33 [19:00<12:38, 44.59s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      12.69 ms /    49 runs   (    0.26
ms per token,  3861.61 tokens per second)
llama_print_timings: prompt eval time =   26391.74 ms /   126 tokens (  209.46
ms per token,     4.77 tokens per second)
llama_print_timings:        eval time =   14606.12 ms /    48 runs   (  304.29
ms per token,     3.29 tokens per second)
```

```
llama_print_timings:       total time =   41186.30 ms /   174 tokens
No. of rows:  52%| | 17/33 [19:41<11:37, 43.57s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      33.34 ms /   112 runs   (    0.30
ms per token,  3358.82 tokens per second)
llama_print_timings: prompt eval time =   33481.98 ms /   157 tokens (  213.26
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   32972.16 ms /   111 runs   (  297.05
ms per token,     3.37 tokens per second)
llama_print_timings:       total time =   66939.71 ms /   268 tokens
No. of rows:  55%| | 18/33 [20:48<12:38, 50.60s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      14.03 ms /    53 runs   (    0.26
ms per token,  3777.89 tokens per second)
llama_print_timings: prompt eval time =   25366.89 ms /   119 tokens (  213.17
ms per token,     4.69 tokens per second)
llama_print_timings:        eval time =   15824.43 ms /    52 runs   (  304.32
ms per token,     3.29 tokens per second)
llama_print_timings:       total time =   41400.61 ms /   171 tokens
No. of rows:  58%| | 19/33 [21:29<11:09, 47.84s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      22.73 ms /    82 runs   (    0.28
ms per token,  3606.77 tokens per second)
llama_print_timings: prompt eval time =   30677.56 ms /   136 tokens (  225.57
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   22694.53 ms /    81 runs   (  280.18
ms per token,     3.57 tokens per second)
llama_print_timings:       total time =   53697.30 ms /   217 tokens
No. of rows:  61%| | 20/33 [22:23<10:44, 49.60s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      23.80 ms /    87 runs   (    0.27
ms per token,  3655.16 tokens per second)
llama_print_timings: prompt eval time =   27789.05 ms /   123 tokens (  225.93
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =   26500.69 ms /    86 runs   (  308.15
ms per token,     3.25 tokens per second)
llama_print_timings:       total time =   54636.20 ms /   209 tokens
No. of rows:  64%| | 21/33 [23:18<10:13, 51.11s/itLlama.generate: prefix-match
hit
```

```
llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      21.28 ms /    67 runs   (    0.32
ms per token,  3148.35 tokens per second)
llama_print_timings: prompt eval time =   22778.02 ms /   106 tokens (  214.89
ms per token,     4.65 tokens per second)
llama_print_timings:        eval time =   18325.64 ms /    66 runs   (  277.66
ms per token,     3.60 tokens per second)
llama_print_timings:       total time =   41378.68 ms /   172 tokens
No. of rows:  67%| | 22/33 [23:59<08:50, 48.19s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      18.49 ms /    70 runs   (    0.26
ms per token,  3785.63 tokens per second)
llama_print_timings: prompt eval time =   26522.23 ms /   121 tokens (  219.19
ms per token,     4.56 tokens per second)
llama_print_timings:        eval time =   20749.64 ms /    69 runs   (  300.72
ms per token,     3.33 tokens per second)
llama_print_timings:       total time =   47547.88 ms /   190 tokens
No. of rows:  70%| | 23/33 [24:47<07:59, 48.00s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      16.80 ms /    62 runs   (    0.27
ms per token,  3689.38 tokens per second)
llama_print_timings: prompt eval time =   31069.57 ms /   151 tokens (  205.76
ms per token,     4.86 tokens per second)
llama_print_timings:        eval time =   16581.51 ms /    61 runs   (  271.83
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   47897.72 ms /   212 tokens
No. of rows:  73%| | 24/33 [25:35<07:11, 47.97s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      20.27 ms /    72 runs   (    0.28
ms per token,  3552.92 tokens per second)
llama_print_timings: prompt eval time =   28503.85 ms /   134 tokens (  212.72
ms per token,     4.70 tokens per second)
llama_print_timings:        eval time =   20021.58 ms /    71 runs   (  281.99
ms per token,     3.55 tokens per second)
llama_print_timings:       total time =   48818.20 ms /   205 tokens
No. of rows:  76%| | 25/33 [26:23<06:25, 48.23s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      25.13 ms /    89 runs   (    0.28
ms per token,  3541.30 tokens per second)
llama_print_timings: prompt eval time =   32218.91 ms /   153 tokens (  210.58
```

```
ms per token,     4.75 tokens per second)
llama_print_timings:        eval time =   25842.47 ms /    88 runs   (  293.66
ms per token,     3.41 tokens per second)
llama_print_timings:       total time =   58419.05 ms /   241 tokens
No. of rows:  79%| | 26/33 [27:22<05:59, 51.29s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      16.31 ms /    46 runs   (    0.35
ms per token,  2821.05 tokens per second)
llama_print_timings: prompt eval time =   25440.25 ms /   120 tokens (  212.00
ms per token,     4.72 tokens per second)
llama_print_timings:        eval time =   13141.32 ms /    45 runs   (  292.03
ms per token,     3.42 tokens per second)
llama_print_timings:       total time =   38767.06 ms /   165 tokens
No. of rows:  82%| | 27/33 [28:01<04:45, 47.53s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      29.44 ms /   108 runs   (    0.27
ms per token,  3668.85 tokens per second)
llama_print_timings: prompt eval time =   29191.15 ms /   140 tokens (  208.51
ms per token,     4.80 tokens per second)
llama_print_timings:        eval time =   29560.84 ms /   107 runs   (  276.27
ms per token,     3.62 tokens per second)
llama_print_timings:       total time =   59186.20 ms /   247 tokens
No. of rows:  85%| | 28/33 [29:00<04:15, 51.04s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      30.89 ms /   107 runs   (    0.29
ms per token,  3463.79 tokens per second)
llama_print_timings: prompt eval time =   33673.73 ms /   152 tokens (  221.54
ms per token,     4.51 tokens per second)
llama_print_timings:        eval time =   29953.26 ms /   106 runs   (  282.58
ms per token,     3.54 tokens per second)
llama_print_timings:       total time =   64067.47 ms /   258 tokens
No. of rows:  88%| | 29/33 [30:04<03:39, 54.95s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      13.51 ms /    47 runs   (    0.29
ms per token,  3480.19 tokens per second)
llama_print_timings: prompt eval time =   22935.21 ms /   110 tokens (  208.50
ms per token,     4.80 tokens per second)
llama_print_timings:        eval time =   14408.78 ms /    46 runs   (  313.23
ms per token,     3.19 tokens per second)
llama_print_timings:       total time =   37530.91 ms /   156 tokens
```

```
No. of rows:  91%| | 30/33 [30:41<02:29, 49.72s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      21.38 ms /    76 runs   (    0.28
ms per token,  3554.72 tokens per second)
llama_print_timings: prompt eval time =   31738.39 ms /   152 tokens (  208.81
ms per token,     4.79 tokens per second)
llama_print_timings:        eval time =   20663.12 ms /    75 runs   (  275.51
ms per token,     3.63 tokens per second)
llama_print_timings:       total time =   52709.07 ms /   227 tokens
No. of rows:  94%| | 31/33 [31:34<01:41, 50.62s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      10.33 ms /    37 runs   (    0.28
ms per token,  3582.49 tokens per second)
llama_print_timings: prompt eval time =   28876.61 ms /   128 tokens (  225.60
ms per token,     4.43 tokens per second)
llama_print_timings:        eval time =    9901.76 ms /    36 runs   (  275.05
ms per token,     3.64 tokens per second)
llama_print_timings:       total time =   38926.05 ms /   164 tokens
No. of rows:  97%| | 32/33 [32:13<00:47, 47.12s/itLlama.generate: prefix-match
hit

llama_print_timings:        load time =  441235.68 ms
llama_print_timings:      sample time =      17.12 ms /    64 runs   (    0.27
ms per token,  3739.41 tokens per second)
llama_print_timings: prompt eval time =   24635.34 ms /   119 tokens (  207.02
ms per token,     4.83 tokens per second)
llama_print_timings:        eval time =   17114.55 ms /    63 runs   (  271.66
ms per token,     3.68 tokens per second)
llama_print_timings:       total time =   42004.85 ms /   182 tokens
No. of rows: 100%| | 33/33 [32:55<00:00, 59.86s/it
```