final-project-technical-report-1

November 29, 2023

1 Enhancing Structured Narrative Generation in Language Models: A Fine-Tuning Approach Utilizing Classic Short Stories

1.1 Abstract

Storytelling is a fundamental human activity instrumental in communication and culture. Recent advancements in large language models (LLMs) have opened new possibilities in automated story generation. This project explores the fine-tuning of LLMs for dynamic and personalized story generation, capable of integrating user preferences into a coherent narrative structure inspired by classic short stories.

1.2 Introduction

Incorporating the intricacies of human storytelling into machine learning models presents a complex challenge—a challenge that, if addressed, can transform how we interact with and consume stories. By tailoring narratives to individual user preferences, we aim to create a new dimension of engagement. Leveraging LLama 2 as the base model, this project aims to fine-tune this model by using differentiating techniques that augment their generative storytelling capabilities.

1.3 Methodology

We grounded our approach in parameter-efficient fine-tuning techniques, primarily focusing on quantized Learning Rate Annealing (qLoRA). A data-driven curriculum was developed to sequentially introduce the model to various facets of storytelling through a large dataset. User preferences are encoded using meta-data tags and injected into the model as conditional elements guiding the generation process.

1.4 Experiments

We conducted a series of experiments aimed at evaluating model performance of Llama 2 versus Mistral, and in comparing these two, we found that Mistral had an unsupportable compute power and thus we decided on using Llama 2 as our base model.

[ ]:
```
!pip install -q accelerate==0.21.0 peft==0.4.0 bitsandbytes==0.40.2␣
↪transformers==4.31.0 trl==0.4.7 datasets
```

[ ]:
```
import os
import torch
from datasets import load_dataset
```

```
from transformers import (
    AutoModelForCausalLM,
    AutoTokenizer,
    BitsAndBytesConfig,
    HfArgumentParser,
    TrainingArguments,
    pipeline,
    logging,
)
from peft import LoraConfig, PeftModel
from trl import SFTTrainer
```

2 Training Configuration

Below is the training configuration for our Llama 2 model.

The training leverages QLoRA, an efficient fine-tuning method that significantly reduces memory usage to enable training large models on resource-constrained environments. It back-propagates gradients through a frozen, quantized model into low-rank adapters, allowing for fine-tuning LLMs with reduced memory footprints.

Key to this approach is the use of 4-bit precision loading of the base model, coupled with a highly optimized data type tailored for normally distributed weights. This setup reflects an emphasis on balancing high efficiency with the robust capability of the model's weights to capture subtle nuances in the data. Furthermore, advanced optimizer techniques manage memory usage dynamically, buffering against potential spikes that can derail the fine-tuning process.

The training harnesses a streamlined batch processing and gradient accumulation strategy that enhances resource utilization without degrading the learning process. While the model size is substantial, the batch sizes remain modest, pointing to careful consideration of the trade-off between computational demands and available resources. Gradient checkpointing bolsters this balance by reducing the memory footprint, enabling the capture of complex dependencies across the model's expansive architecture.

The fine-tuning process employs the AdamW optimizer, an adaptation of the traditional Adam optimizer which incorporates decoupled weight decay regularization. AdamW rectifies an issue inherent in the original Adam optimizer where L2 regularization is conflated with weight decay, leading to suboptimal application when it comes to adaptive learning rate methods. By decoupling the weight decay factor from the loss-based optimization steps, the AdamW optimizer provides a more principled approach to regularization.

[ ]:
```
# The model that you want to train from the Hugging Face hub
model_name = "NousResearch/Llama-2-7b-chat-hf"

# The instruction dataset to use
dataset_name = "siddrao11/test"

# Fine-tuned model name
```

```
new_model = "llama-2-7b-storytelling"

################################################################################
# QLoRA parameters
################################################################################

# LoRA attention dimension
lora_r = 64

# Alpha parameter for LoRA scaling
lora_alpha = 16

# Dropout probability for LoRA layers
lora_dropout = 0.1

################################################################################
# bitsandbytes parameters
################################################################################

# Activate 4-bit precision base model loading
use_4bit = True

# Compute dtype for 4-bit base models
bnb_4bit_compute_dtype = "float16"

# Quantization type (fp4 or nf4)
bnb_4bit_quant_type = "nf4"

# Activate nested quantization for 4-bit base models (double quantization)
use_nested_quant = False

################################################################################
# TrainingArguments parameters
################################################################################

# Output directory where the model predictions and checkpoints will be stored
output_dir = "./drive/MyDrive/cs180"

# Number of training epochs
num_train_epochs = 1

# Enable fp16/bf16 training (set bf16 to True with an A100)
fp16 = False
bf16 = False

# Batch size per GPU for training
per_device_train_batch_size = 4
```

```
# Batch size per GPU for evaluation
per_device_eval_batch_size = 4

# Number of update steps to accumulate the gradients for
gradient_accumulation_steps = 1

# Enable gradient checkpointing
gradient_checkpointing = True

# Maximum gradient normal (gradient clipping)
max_grad_norm = 0.3

# Initial learning rate (AdamW optimizer)
learning_rate = 2e-4

# Weight decay to apply to all layers except bias/LayerNorm weights
weight_decay = 0.001

# Optimizer to use
optim = "paged_adamw_32bit"

# Learning rate schedule
lr_scheduler_type = "cosine"

# Number of training steps (overrides num_train_epochs)
max_steps = -1

# Ratio of steps for a linear warmup (from 0 to learning rate)
warmup_ratio = 0.03

# Group sequences into batches with same length
# Saves memory and speeds up training considerably
group_by_length = True

# Save checkpoint every X updates steps
save_steps = 250

# Log every X updates steps
logging_steps = 25

################################################################################
# SFT parameters
################################################################################

# Maximum sequence length to use
max_seq_length = None
```

```
# Pack multiple short examples in the same input sequence to increase efficiency
packing = False

# Load the entire model on the GPU 0
device_map = {"": 0}
```

2.1 Data

This work collects a large dataset of 300,000 human-written stories paired with writing prompts from an online forum that enables hierarchical story generation, specifically found in the Hierarchical Neural Story Generation github. Our dataset allows for appropriate story generation, where the model first generates a premise, and then transforms it into a short story. The processed dataset is available here.

2.2 Task

The primary task was to generate structured, coherent, and personalized short stories using a fine-tuned model, challenging it to maintain narrative integrity while adapting to diverse user-defined elements.

[ ]:
```
# Load dataset (you can process it here)
dataset = load_dataset(dataset_name, split="train")

# Load tokenizer and model with QLoRA configuration
compute_dtype = getattr(torch, bnb_4bit_compute_dtype)

bnb_config = BitsAndBytesConfig(
    load_in_4bit=use_4bit,
    bnb_4bit_quant_type=bnb_4bit_quant_type,
    bnb_4bit_compute_dtype=compute_dtype,
    bnb_4bit_use_double_quant=use_nested_quant,
)

# Check GPU compatibility with bfloat16
if compute_dtype == torch.float16 and use_4bit:
    major, _ = torch.cuda.get_device_capability()
    if major >= 8:
        print("=" * 80)
        print("Your GPU supports bfloat16: accelerate training with bf16=True")
        print("=" * 80)

# Load base model
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    quantization_config=bnb_config,
    device_map=device_map
)
```

```
model.config.use_cache = False
model.config.pretraining_tp = 1

# Load LLaMA tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)
tokenizer.pad_token = tokenizer.eos_token
tokenizer.padding_side = "right" # Fix weird overflow issue with fp16 training

# Load LoRA configuration
peft_config = LoraConfig(
    lora_alpha=lora_alpha,
    lora_dropout=lora_dropout,
    r=lora_r,
    bias="none",
    task_type="CAUSAL_LM",
)

# Set training parameters
training_arguments = TrainingArguments(
    output_dir=output_dir,
    num_train_epochs=num_train_epochs,
    per_device_train_batch_size=per_device_train_batch_size,
    gradient_accumulation_steps=gradient_accumulation_steps,
    optim=optim,
    save_steps=save_steps,
    logging_steps=logging_steps,
    learning_rate=learning_rate,
    weight_decay=weight_decay,
    fp16=fp16,
    bf16=bf16,
    max_grad_norm=max_grad_norm,
    max_steps=max_steps,
    warmup_ratio=warmup_ratio,
    group_by_length=group_by_length,
    lr_scheduler_type=lr_scheduler_type,
    report_to="tensorboard"
)

# Set supervised fine-tuning parameters
# trainer = SFTTrainer(
#     model=model,
#     train_dataset=dataset,
#     peft_config=peft_config,
#     dataset_text_field="formatted_text",
#     max_seq_length=max_seq_length,
#     tokenizer=tokenizer,
#     args=training_arguments,
```

```
#     packing=packing,
# )

checkpoint_path = os.path.join(output_dir, 'checkpoint-5000')
# Train model
# trainer.train(checkpoint_path)

# # Save trained model
# trainer.model.save_pretrained(new_model)
```

```
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
<ipython-input-6-84c7309441c4> in <cell line: 2>()
      1 # Load dataset (you can process it here)
----> 2 dataset = load_dataset(dataset_name, split="train")
      3
      4 # Load tokenizer and model with QLoRA configuration
      5 compute_dtype = getattr(torch, bnb_4bit_compute_dtype)

/usr/local/lib/python3.10/dist-packages/datasets/load.py in load_dataset(path,␣
 ↪name, data_dir, data_files, split, cache_dir, features, download_config,␣
 ↪download_mode, verification_mode, ignore_verifications, keep_in_memory,␣
 ↪save_infos, revision, token, use_auth_token, task, streaming, num_proc,␣
 ↪storage_options, **config_kwargs)
   2126
   2127     # Create a dataset builder
-> 2128     builder_instance = load_dataset_builder(

   2129         path=path,
   2130         name=name,

/usr/local/lib/python3.10/dist-packages/datasets/load.py in␣
 ↪load_dataset_builder(path, name, data_dir, data_files, cache_dir, features,␣
 ↪download_config, download_mode, revision, token, use_auth_token,␣
 ↪storage_options, **config_kwargs)
   1812         download_config = download_config.copy() if download_config els␣
 ↪DownloadConfig()
   1813         download_config.storage_options.update(storage_options)
-> 1814     dataset_module = dataset_module_factory(

   1815         path,
   1816         revision=revision,

/usr/local/lib/python3.10/dist-packages/datasets/load.py in␣
 ↪dataset_module_factory(path, revision, download_config, download_mode,␣
 ↪dynamic_modules_path, data_dir, data_files, **download_kwargs)
   1493                     download_config=download_config,
   1494                     download_mode=download_mode,
```

```
-> 1495                 ).get_module()
   1496         except (
   1497             Exception

/usr/local/lib/python3.10/dist-packages/datasets/load.py in get_module(self)
   1013
   1014     def get_module(self) -> DatasetModule:
-> 1015         hfh_dataset_info = HfApi(config.HF_ENDPOINT).dataset_info(
   1016             self.name,
   1017             revision=self.revision,

/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_validators.py in
 ↪_inner_fn(*args, **kwargs)
    116             kwargs = smoothly_deprecate_use_auth_token(fn_name=fn.
 ↪__name__, has_token=has_token, kwargs=kwargs)
    117
--> 118         return fn(*args, **kwargs)
    119
    120     return _inner_fn  # type: ignore

/usr/local/lib/python3.10/dist-packages/huggingface_hub/hf_api.py in␣
 ↪dataset_info(self, repo_id, revision, timeout, files_metadata, token)
   1982             params["blobs"] = True
   1983
-> 1984         r = get_session().get(path, headers=headers, timeout=timeout,␣
 ↪params=params)
   1985         hf_raise_for_status(r)
   1986         data = r.json()

/usr/local/lib/python3.10/dist-packages/requests/sessions.py in get(self, url,␣
 ↪**kwargs)
    600
    601         kwargs.setdefault("allow_redirects", True)
--> 602         return self.request("GET", url, **kwargs)
    603
    604     def options(self, url, **kwargs):

/usr/local/lib/python3.10/dist-packages/requests/sessions.py in request(self,␣
 ↪method, url, params, data, headers, cookies, files, auth, timeout,␣
 ↪allow_redirects, proxies, hooks, stream, verify, cert, json)
    587         }
    588         send_kwargs.update(settings)
--> 589         resp = self.send(prep, **send_kwargs)
    590
    591         return resp
```

```
/usr/local/lib/python3.10/dist-packages/requests/sessions.py in send(self,␣
↪request, **kwargs)
    701
    702         # Send the request
--> 703         r = adapter.send(request, **kwargs)
    704
    705         # Total elapsed time of the request (approximately)

/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_http.py in␣
↪send(self, request, *args, **kwargs)
     61         """Catch any RequestException to append request id to the error␣
↪message for debugging."""
     62         try:
---> 63             return super().send(request, *args, **kwargs)
     64         except requests.RequestException as e:
     65             request_id = request.headers.get(X_AMZN_TRACE_ID)

/usr/local/lib/python3.10/dist-packages/requests/adapters.py in send(self,␣
↪request, stream, timeout, verify, cert, proxies)
    484
    485         try:
--> 486             resp = conn.urlopen(
    487                 method=request.method,
    488                 url=url,

/usr/local/lib/python3.10/dist-packages/urllib3/connectionpool.py in␣
↪urlopen(self, method, url, body, headers, retries, redirect, assert_same_host␣
↪timeout, pool_timeout, release_conn, chunked, body_pos, preload_content,␣
↪decode_content, **response_kw)
    789
    790             # Make the request on the HTTPConnection object
--> 791             response = self._make_request(
    792                 conn,
    793                 method,

/usr/local/lib/python3.10/dist-packages/urllib3/connectionpool.py in␣
↪_make_request(self, conn, method, url, body, headers, retries, timeout,␣
↪chunked, response_conn, preload_content, decode_content,␣
↪enforce_content_length)
    535         # Receive the response from the server
    536         try:
--> 537             response = conn.getresponse()
    538         except (BaseSSLError, OSError) as e:
    539             self._raise_timeout(err=e, url=url,␣
↪timeout_value=read_timeout)

/usr/local/lib/python3.10/dist-packages/urllib3/connection.py in␣
↪getresponse(self)
```

```
    459
    460         # Get the response from http.client.HTTPConnection
--> 461         httplib_response = super().getresponse()
    462
    463         try:

/usr/lib/python3.10/http/client.py in getresponse(self)
   1373         try:
   1374             try:
-> 1375                 response.begin()
   1376             except ConnectionError:
   1377                 self.close()

/usr/lib/python3.10/http/client.py in begin(self)
    316         # read until we get a non-100 response
    317         while True:
--> 318             version, status, reason = self._read_status()
    319             if status != CONTINUE:
    320                 break

/usr/lib/python3.10/http/client.py in _read_status(self)
    277
    278     def _read_status(self):
--> 279         line = str(self.fp.readline(_MAXLINE + 1), "iso-8859-1")
    280         if len(line) > _MAXLINE:
    281             raise LineTooLong("status line")

/usr/lib/python3.10/socket.py in readinto(self, b)
    703         while True:
    704             try:
--> 705                 return self._sock.recv_into(b)
    706             except timeout:
    707                 self._timeout_occurred = True

/usr/lib/python3.10/ssl.py in recv_into(self, buffer, nbytes, flags)
   1272                   "non-zero flags not allowed in calls to recv_into() o
 ↪%s" %
   1273                   self.__class__)
-> 1274             return self.read(nbytes, buffer)
   1275         else:
   1276             return super().recv_into(buffer, nbytes, flags)

/usr/lib/python3.10/ssl.py in read(self, len, buffer)
   1128         try:
   1129             if buffer is not None:
-> 1130                 return self._sslobj.read(len, buffer)
   1131             else:
   1132                 return self._sslobj.read(len)
```

```
KeyboardInterrupt:
```

2.3 Evaluation Protocol

Performance is difficult to assess here because we are analyzing a generative model and we can't necessarily directly compare to a solution, but what we can do is analyze a quantitative metric, such as the perplexity score, where the lower the perplexity score indicates a better response. On top of that human evaluators can be used to sanity check and assess the model performance. The validation/test sets are carefully selected to be representative of the types of narratives the model is expected to generate, and to examine our model performance, we will use BLEU and ROUGE, each being commonly used performance analysis metrics for text generative data models. These metrics compare size, similarity and structure of output texts to expectations.

[ ]:
```
from google.colab import drive
drive.mount('/content/drive')
```

[ ]:
```
# %load_ext tensorboard
# %tensorboard --logdir results/runs
```

[ ]:
```
# Ignore warnings
logging.set_verbosity(logging.CRITICAL)

# Run text generation pipeline with our next model
# prompt = "What is a large language model?"
# pipe = pipeline(task="text-generation", model=model, tokenizer=tokenizer,␣
 ↪max_length=200)
# result = pipe(f"<s>[INST] {prompt} [/INST]")
# print(result[0]['generated_text'])
```

[ ]:
```
# Empty VRAM
# del model
# del pipe
# del trainer
import gc
gc.collect()
gc.collect()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-7-35c102f0e46b> in <cell line: 3>()
      1 # Empty VRAM
      2 del model
----> 3 del pipe
      4 del trainer
      5 import gc
```

```
NameError: name 'pipe' is not defined
```

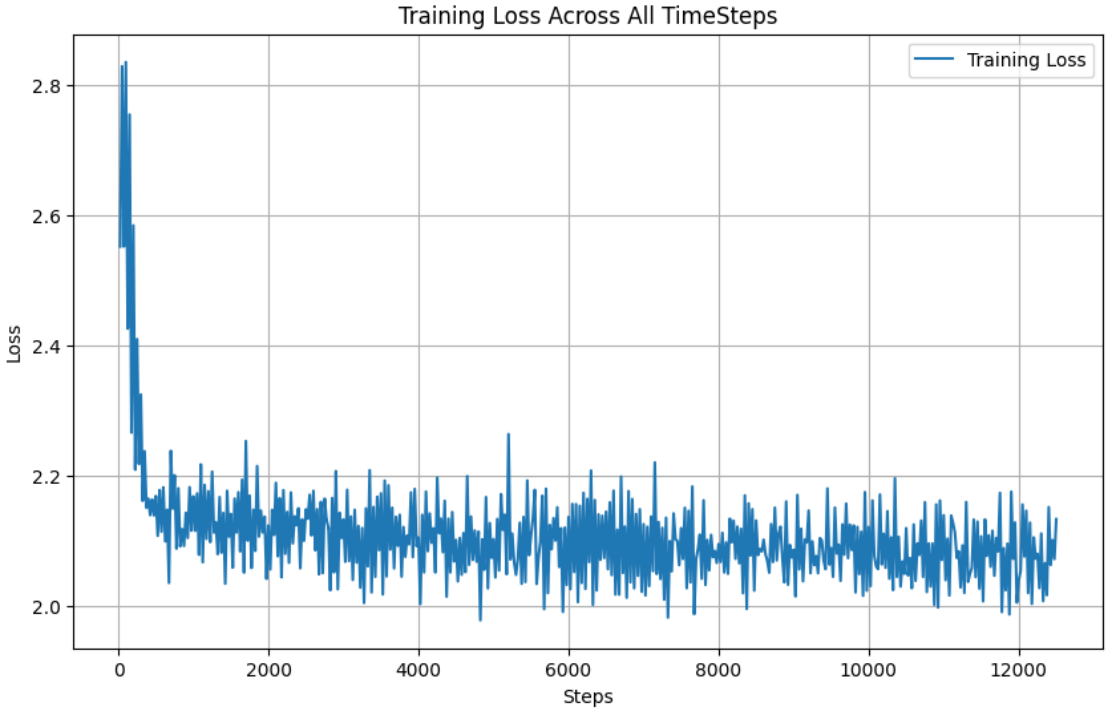
2.4 Results

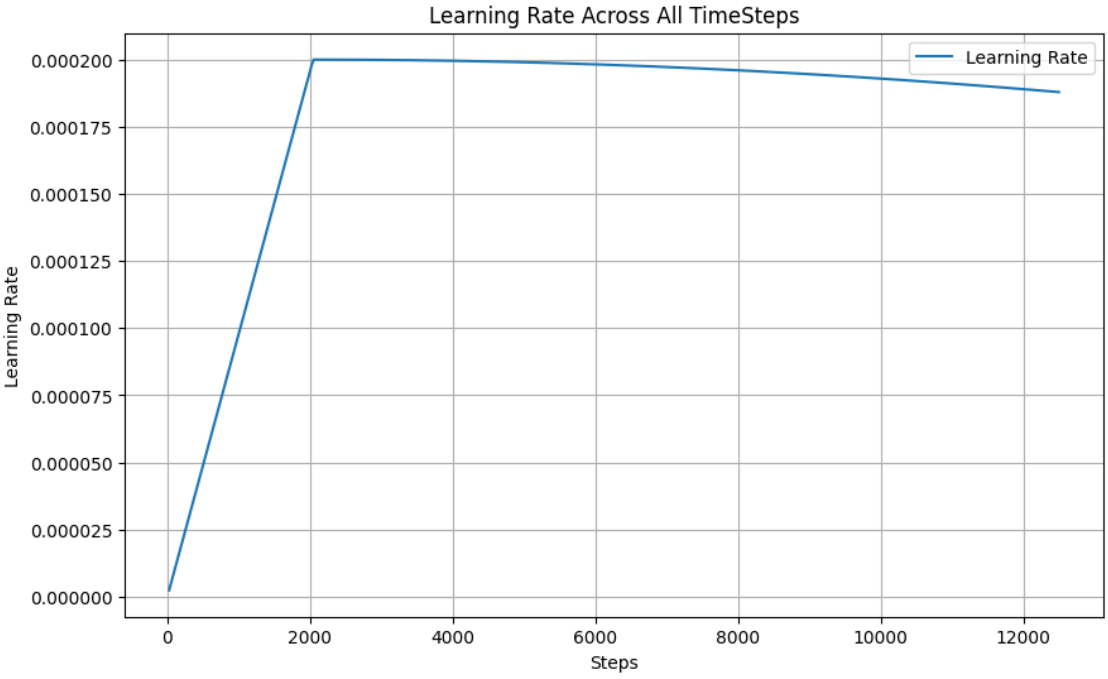
The fine-tuned Llama 2 model showed improved narrative structuring when evaluated with standard literary analysis criteria. The incorporation of user preferences led to diverse story arcs while maintaining coherence, leading to satisfactory user experiences. Quantitative analysis of the Llama 2 model is still ongoing.

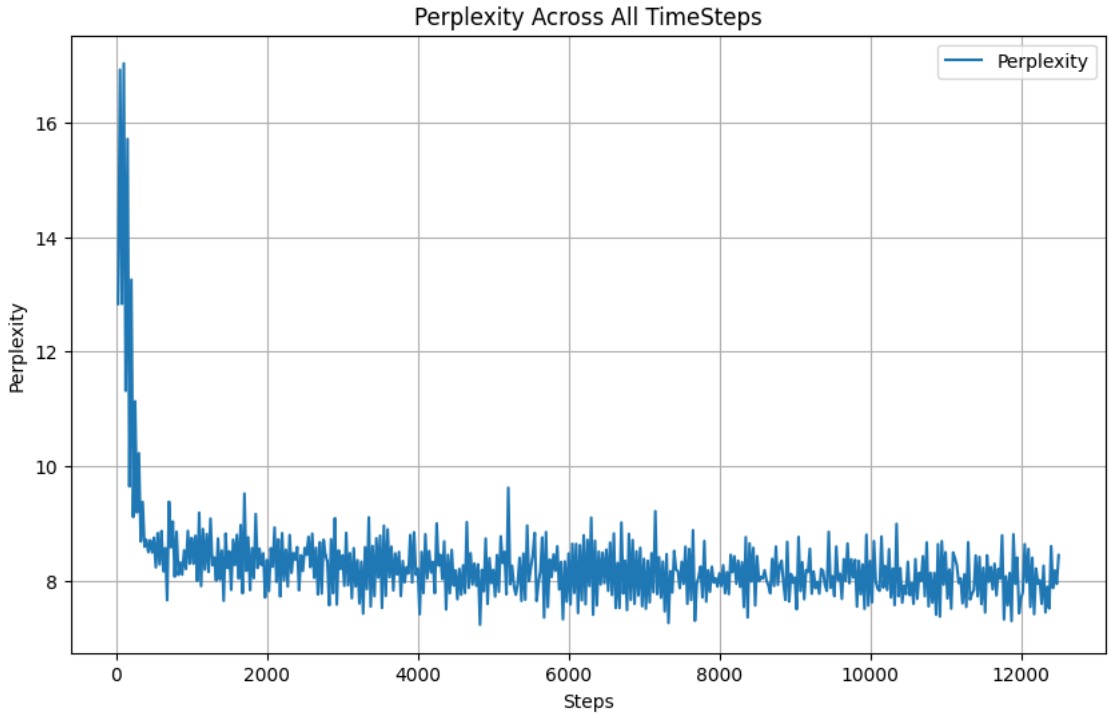
2.4.1 Training Evals (Loss, Perplexity)

[ ]:
```python
import json
import os
import matplotlib.pyplot as plt
import numpy as np

# Initialize lists to store steps and losses from all checkpoints
all_steps = []
all_losses = []
all_lr = []
all_perplexities = []

# List all checkpoint subdirectories in output_dir (assumes naming convention␣
 ↪starts with "checkpoint-")
checkpoint_dirs = [d for d in os.listdir(output_dir) if d.
 ↪startswith('checkpoint-') and os.path.isdir(os.path.join(output_dir, d))]

# Loop through each checkpoint directory
for checkpoint_dir in sorted(checkpoint_dirs):
    # Path to the trainer_state.json in the current checkpoint directory
    trainer_state_path = os.path.join(output_dir, checkpoint_dir,␣
 ↪'trainer_state.json')

    # Load the trainer_state.json file
    with open(trainer_state_path, 'r') as f:
        trainer_state = json.load(f)

    # Extract the log_history field
    log_history = trainer_state.get('log_history', [])

    # Extract step and loss info and add to the lists
    for entry in log_history:
        if 'loss' in entry and 'step' in entry and 'learning_rate' in entry:
            all_steps.append(entry['step'])
            all_losses.append(entry['loss'])
            all_lr.append(entry['learning_rate'])
```

```
            perplexity = np.exp(entry['loss'])
            all_perplexities.append(perplexity)

# Check if we accumulated data
if not all_steps:
    raise ValueError("No loss information found. Please check if the␣
 ↪checkpoints contain 'trainer_state.json' and 'log_history'.")

# Sorting the all_steps and all_losses based on step values
sorted_indices = sorted(range(len(all_steps)), key=lambda k: all_steps[k])
all_steps = [all_steps[i] for i in sorted_indices]
all_losses = [all_losses[i] for i in sorted_indices]
all_lr = [all_lr[i] for i in sorted_indices]
all_perplexities = [all_perplexities[i] for i in sorted_indices]

# Plot loss against steps from all checkpoints combined
plt.figure(figsize=(10, 6))
plt.plot(all_steps, all_losses, label='Training Loss')
plt.xlabel('Steps')
plt.ylabel('Loss')
plt.title('Training Loss Across All TimeSteps')
plt.legend()
plt.grid(True)
plt.show()

# Plot lr
plt.figure(figsize=(10, 6))
plt.plot(all_steps, all_lr, label='Learning Rate')
plt.xlabel('Steps')
plt.ylabel('Learning Rate')
plt.title('Learning Rate Across All TimeSteps')
plt.legend()
plt.grid(True)
plt.show()

# Plot perplexity
plt.figure(figsize=(10, 6))
plt.plot(all_steps, all_perplexities, label='Perplexity')
plt.xlabel('Steps')
plt.ylabel('Perplexity')
plt.title('Perplexity Across All TimeSteps')
plt.legend()
plt.grid(True)
plt.show()
```


Training Loss Across All TimeSteps
Training Loss
Loss
Steps
2.8
2.6
2.4
2.2
2.0
0
2000
4000
6000
8000
10000
12000


Learning Rate Across All TimeSteps
Learning Rate
Learning Rate
Steps
0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
0
2000
4000
6000
8000
10000
12000



Analyzing our plots, we see that perplexity seems to decrease as we increase the number of steps and begins to level off around 8. This makes sense becasue the lower the perplexity the better our text generative model is at generating choesize text. Furthermore for the learning rate we have a warm-up phase at the start, peaking and then as we increase the number of steps the learning rate gradually begins to decrease as we expect.

2.4.2 Test Evals (BLEU, ROUGE)

BLEU: Baseline

[ ]:
```
from datasets import load_dataset
from transformers import (
    AutoModelForCausalLM,
    AutoTokenizer,
    pipeline,
)
import nltk
from nltk.translate.bleu_score import sentence_bleu, SmoothingFunction
from tqdm import tqdm
import torch
import re

# Install NLTK if not already installed and download BLEU's tokenizer model
nltk.download('punkt')
```

```
# Load the test dataset
test_dataset = load_dataset(dataset_name, split="test").select(range(1000))

# Load base model
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    low_cpu_mem_usage=True,
    torch_dtype=torch.float16,
    # Adjust device_map as below if you have a device map, else use .to("cuda")
    device_map=device_map,
)

# Load the tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)
tokenizer.pad_token = tokenizer.eos_token
tokenizer.padding_side = "right"

# Create generation pipeline
translation_pipeline = pipeline("text-generation", model=model,␣
 ↪tokenizer=tokenizer, device=0) # Adjust device as needed

# Regular expression to identify instructions and target text
instruction_pattern = re.compile(r'<s>\[INST\] (.+?) \[/INST\](.+)<\/s>')

# Calculate BLEU scores
bleu_scores = []

# Define smoothing function for nltk BLEU calculation to handle potential zero␣
 ↪n-gram counts
chencherry = SmoothingFunction()

for instance in tqdm(test_dataset):
    formatted_text = instance['formatted_text']

    # Match the instruction pattern to separate the source and target
    match = instruction_pattern.match(formatted_text)
    if match:
        source_text, target_text = match.groups()
        target_text = [nltk.word_tokenize(target_text.strip())] # Tokenize␣
 ↪reference text

        # Generate the translation
        translated = translation_pipeline(source_text)[0]['generated_text']

        # Tokenize the predicted text
        predicted_tokens = nltk.word_tokenize(translated)
```

```
        # Calculate BLEU score for this instance, with smoothing
        bleu_score = sentence_bleu(target_text, predicted_tokens,␣
 ↪smoothing_function=chencherry.method1)
        bleu_scores.append(bleu_score)

# Calculate the average BLEU score over all instances
average_bleu = sum(bleu_scores) / len(bleu_scores)
print("Average BLEU score on the test set:", average_bleu)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
```

```
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
```

```
Xformers is not installed correctly. If you want to use
memory_efficient_attention to accelerate training use the following command to
install Xformers
pip install xformers.
  0%|          | 0/1000 [00:00<?, ?it/s]/usr/local/lib/python3.10/dist-
packages/transformers/generation/utils.py:1270: UserWarning: You have modified
the pretrained model configuration to control generation. This is a deprecated
strategy to control generation and will be removed soon, in a future version.
Please use a generation configuration file (see
https://huggingface.co/docs/transformers/main_classes/text_generation )
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/transformers/generation/utils.py:1369:
UserWarning: Using `max_length`'s default (20) to control the generation length.
This behaviour is deprecated and will be removed from the config in v5 of
Transformers -- we recommend using `max_new_tokens` to control the maximum
length of the generation.
  warnings.warn(
Input length of input_ids is 50, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  0%|          | 2/1000 [00:01<11:08,  1.49it/s]Input length of input_ids is 41,
but `max_length` is set to 20. This can lead to unexpected behavior. You should
consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  0%|          | 4/1000 [00:01<05:08,  3.23it/s]Input length of input_ids is 42,
but `max_length` is set to 20. This can lead to unexpected behavior. You should
consider increasing `max_new_tokens`.
  1%|          | 9/1000 [00:03<07:58,  2.07it/s]Input length of input_ids is 25,
but `max_length` is set to 20. This can lead to unexpected behavior. You should
consider increasing `max_new_tokens`.
/usr/local/lib/python3.10/dist-packages/transformers/pipelines/base.py:1083:
UserWarning: You seem to be using the pipelines sequentially on GPU. In order to
maximize efficiency please use a dataset
  warnings.warn(
```

```
  1%|          | 11/1000 [00:04<04:57,  3.32it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  1%|          | 13/1000 [00:04<03:37,  4.53it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 51, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  2%|          | 16/1000 [00:04<03:09,  5.19it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  2%|          | 18/1000 [00:05<03:27,  4.74it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  2%|          | 19/1000 [00:05<02:58,  5.49it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  2%|          | 21/1000 [00:05<02:20,  6.97it/s]Input length of input_ids is
50, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  2%|          | 22/1000 [00:05<02:10,  7.47it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  2%|          | 23/1000 [00:05<02:02,  7.95it/s]Input length of input_ids is
70, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  2%|          | 24/1000 [00:05<01:59,  8.16it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  3%|          | 26/1000 [00:06<03:49,  4.25it/s]Input length of input_ids is
74, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  3%|          | 27/1000 [00:06<03:17,  4.92it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 47, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  3%|          | 29/1000 [00:06<02:31,  6.42it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  3%|          | 32/1000 [00:08<05:18,  3.03it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  3%|          | 34/1000 [00:08<05:55,  2.72it/s]Input length of input_ids is
```

```
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 37/1000 [00:09<05:33,  2.89it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 38/1000 [00:09<04:40,  3.43it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 39/1000 [00:09<03:55,  4.08it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 41/1000 [00:10<05:05,  3.14it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 42/1000 [00:10<04:18,  3.71it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 44/1000 [00:11<05:32,  2.88it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 45/1000 [00:11<04:42,  3.38it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  5%|         | 46/1000 [00:12<03:57,  4.01it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 49, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  5%|         | 48/1000 [00:12<02:59,  5.30it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  5%|         | 50/1000 [00:12<02:25,  6.53it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  5%|         | 52/1000 [00:12<02:06,  7.50it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  5%|         | 54/1000 [00:12<01:53,  8.35it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 60, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
```

```
  6%|         | 56/1000 [00:13<01:46,  8.88it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  6%|         | 58/1000 [00:13<01:40,  9.36it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  6%|         | 63/1000 [00:15<05:16,  2.96it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  6%|         | 64/1000 [00:15<04:31,  3.45it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  7%|         | 66/1000 [00:15<03:52,  4.02it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 60, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  7%|         | 68/1000 [00:16<03:01,  5.13it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  7%|         | 69/1000 [00:16<02:44,  5.64it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 43, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  7%|         | 71/1000 [00:16<02:18,  6.72it/s]Input length of input_ids is
64, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  7%|         | 73/1000 [00:16<02:22,  6.53it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  7%|         | 74/1000 [00:16<02:13,  6.95it/s]Input length of input_ids is
70, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  8%|         | 77/1000 [00:18<07:53,  1.95it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  8%|         | 78/1000 [00:19<06:12,  2.48it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  8%|         | 80/1000 [00:19<04:07,  3.71it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 65, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  8%|         | 82/1000 [00:19<03:10,  4.81it/s]Input length of input_ids is
59, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  8%|         | 85/1000 [00:19<02:36,  5.86it/s]Input length of input_ids is
30, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  9%|         | 86/1000 [00:19<02:23,  6.35it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  9%|         | 88/1000 [00:20<02:00,  7.54it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  9%|         | 91/1000 [00:21<03:43,  4.07it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  9%|         | 93/1000 [00:21<02:53,  5.24it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 10%|         | 96/1000 [00:21<02:50,  5.30it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 10%|         | 98/1000 [00:21<02:20,  6.42it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 10%|         | 101/1000 [00:23<05:54,  2.54it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 10%|         | 102/1000 [00:23<04:56,  3.03it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 10%|         | 104/1000 [00:23<03:34,  4.18it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 11%|         | 108/1000 [00:26<06:39,  2.23it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
 11%|          | 109/1000 [00:26<05:24,  2.75it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 11%|          | 110/1000 [00:26<04:25,  3.36it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 11%|          | 112/1000 [00:27<06:40,  2.22it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 11%|          | 113/1000 [00:27<05:27,  2.71it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 12%|          | 117/1000 [00:29<05:30,  2.67it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 71, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 12%|          | 119/1000 [00:29<03:53,  3.77it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 12%|          | 122/1000 [00:30<04:04,  3.60it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 12%|          | 123/1000 [00:30<03:28,  4.20it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 12%|          | 125/1000 [00:30<04:01,  3.62it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 13%|          | 126/1000 [00:30<03:21,  4.33it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 44, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 13%|          | 130/1000 [00:31<03:07,  4.63it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 13%|          | 132/1000 [00:31<02:23,  6.05it/s]Input length of input_ids is
72, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 13%|          | 134/1000 [00:32<03:29,  4.14it/s]Input length of input_ids is
60, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
 14%|         | 136/1000 [00:32<03:01,  4.75it/s]Input length of input_ids is
72, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 14%|         | 138/1000 [00:33<03:29,  4.11it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 14%|         | 140/1000 [00:33<02:33,  5.61it/s]Input length of input_ids is
59, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 14%|         | 142/1000 [00:34<03:09,  4.52it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 14%|         | 143/1000 [00:34<02:42,  5.29it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 14%|         | 145/1000 [00:34<02:57,  4.82it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 56, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 15%|         | 147/1000 [00:34<02:14,  6.35it/s]Input length of input_ids is
30, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 15%|         | 149/1000 [00:34<01:51,  7.65it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 43, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 15%|         | 152/1000 [00:35<03:27,  4.09it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 15%|         | 153/1000 [00:35<02:59,  4.71it/s]Input length of input_ids is
59, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 16%|         | 155/1000 [00:36<02:20,  6.00it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 58, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 16%|         | 157/1000 [00:36<02:01,  6.97it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 16%|         | 159/1000 [00:36<02:26,  5.74it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 16%|         | 161/1000 [00:37<02:04,  6.73it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 57, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 17%|         | 166/1000 [00:38<03:50,  3.62it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 17%|         | 167/1000 [00:38<03:20,  4.16it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 17%|         | 169/1000 [00:39<04:35,  3.01it/s]Input length of input_ids is
56, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 17%|         | 170/1000 [00:39<03:55,  3.52it/s]Input length of input_ids is
67, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 17%|         | 171/1000 [00:39<03:21,  4.12it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 50, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 17%|         | 173/1000 [00:40<02:34,  5.34it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 18%|         | 175/1000 [00:41<04:00,  3.43it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 18%|         | 177/1000 [00:41<03:26,  3.99it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 18%|         | 179/1000 [00:41<02:43,  5.02it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 18%|         | 182/1000 [00:43<05:37,  2.42it/s]Input length of input_ids is
57, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 18%|         | 183/1000 [00:43<04:44,  2.87it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 34, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 18%|         | 185/1000 [00:43<03:25,  3.96it/s]Input length of input_ids is
```

```
72, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 19%|         | 188/1000 [00:44<04:54,  2.76it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 19%|         | 189/1000 [00:45<04:07,  3.28it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 25, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 19%|         | 193/1000 [00:45<03:28,  3.86it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 20%|         | 195/1000 [00:46<03:02,  4.40it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 20%|         | 196/1000 [00:46<02:41,  4.99it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 20%|         | 198/1000 [00:46<02:41,  4.95it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 20%|         | 200/1000 [00:46<02:11,  6.08it/s]Input length of input_ids is
58, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 20%|         | 202/1000 [00:47<01:52,  7.09it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 20%|         | 204/1000 [00:47<01:39,  8.02it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 21%|         | 206/1000 [00:47<01:31,  8.67it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 21%|         | 207/1000 [00:47<01:29,  8.83it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 21%|         | 208/1000 [00:47<01:27,  9.04it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
```

```
 21%|         | 213/1000 [00:48<02:41,  4.89it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 21%|         | 214/1000 [00:48<02:26,  5.35it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 22%|         | 217/1000 [00:50<04:28,  2.92it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 22%|         | 218/1000 [00:50<03:44,  3.48it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 50, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 22%|         | 222/1000 [00:51<05:16,  2.46it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 22%|         | 224/1000 [00:52<05:24,  2.39it/s]Input length of input_ids is
75, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 23%|         | 226/1000 [00:53<05:31,  2.33it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 23%|         | 227/1000 [00:53<04:26,  2.90it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 23%|         | 230/1000 [00:54<04:37,  2.77it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 23%|         | 232/1000 [00:55<04:27,  2.87it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 23%|         | 234/1000 [00:55<03:43,  3.43it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 24%|         | 235/1000 [00:55<03:10,  4.02it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 44, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 24%|         | 237/1000 [00:55<02:21,  5.38it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 54, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 24%|         | 239/1000 [00:56<01:54,  6.62it/s]Input length of input_ids is
```

```
50, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 42, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 24%|         | 241/1000 [00:56<01:39,  7.61it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 49, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 24%|         | 243/1000 [00:56<01:29,  8.46it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 24%|         | 245/1000 [00:56<01:22,  9.11it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 25%|         | 247/1000 [00:56<01:18,  9.65it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 25%|         | 250/1000 [00:57<02:02,  6.12it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 25%|         | 253/1000 [00:58<02:57,  4.20it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 26%|         | 256/1000 [00:59<04:06,  3.01it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 26%|         | 258/1000 [00:59<03:23,  3.64it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 26%|         | 261/1000 [01:01<05:04,  2.43it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 26%|         | 263/1000 [01:02<04:34,  2.68it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 26%|         | 265/1000 [01:02<03:45,  3.26it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 27%|         | 268/1000 [01:03<03:49,  3.19it/s]Input length of input_ids is
63, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 27%|         | 269/1000 [01:03<03:16,  3.72it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 27%|         | 271/1000 [01:03<02:28,  4.92it/s]Input length of input_ids is
58, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 27%|         | 273/1000 [01:04<02:50,  4.26it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 27%|         | 274/1000 [01:04<02:30,  4.81it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 28%|         | 276/1000 [01:04<02:00,  6.03it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 28%|         | 278/1000 [01:04<01:56,  6.19it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 33, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 28%|         | 280/1000 [01:05<01:39,  7.22it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 28%|         | 282/1000 [01:05<01:28,  8.15it/s]Input length of input_ids is
63, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 28%|         | 284/1000 [01:05<01:29,  7.99it/s]Input length of input_ids is
50, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 29%|         | 287/1000 [01:06<02:31,  4.70it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 29%|         | 289/1000 [01:06<01:59,  5.96it/s]Input length of input_ids is
67, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 29%|         | 290/1000 [01:06<01:50,  6.44it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 29%|         | 292/1000 [01:06<01:32,  7.66it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

should consider increasing `max_new_tokens`.
 29%| | 293/1000 [01:07<01:27, 8.06it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 30%| | 295/1000 [01:07<01:18, 9.02it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 30%| | 298/1000 [01:08<03:02, 3.84it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 30%| | 302/1000 [01:09<03:34, 3.26it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 30%| | 303/1000 [01:09<02:59, 3.87it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 30%| | 305/1000 [01:09<02:37, 4.42it/s]Input length of input_ids is
63, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 31%| | 307/1000 [01:10<02:49, 4.08it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 31%| | 308/1000 [01:10<02:26, 4.71it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 76, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 31%| | 311/1000 [01:10<02:02, 5.65it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 31%| | 313/1000 [01:11<01:39, 6.91it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 31%| | 314/1000 [01:11<01:32, 7.38it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 32%| | 316/1000 [01:12<02:41, 4.25it/s]Input length of input_ids is
75, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 32%| | 317/1000 [01:12<02:24, 4.73it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.

```
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 32%|        | 320/1000 [01:12<02:23,  4.75it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 48, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 32%|        | 322/1000 [01:12<01:52,  6.03it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 32%|        | 324/1000 [01:13<01:34,  7.15it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 33%|        | 328/1000 [01:14<03:23,  3.30it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 33%|        | 330/1000 [01:15<03:54,  2.85it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 33%|        | 332/1000 [01:15<02:53,  3.85it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 55, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 34%|        | 335/1000 [01:16<03:02,  3.65it/s]Input length of input_ids is
64, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 34%|        | 336/1000 [01:16<02:38,  4.20it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 34%|        | 339/1000 [01:17<04:05,  2.69it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 31, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 34%|        | 341/1000 [01:18<02:55,  3.76it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 34, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 34%|        | 345/1000 [01:19<04:25,  2.47it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 35%|        | 348/1000 [01:20<03:04,  3.54it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 35%|        | 350/1000 [01:20<02:16,  4.75it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 35%|        | 353/1000 [01:21<03:28,  3.10it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 36%|        | 355/1000 [01:21<02:26,  4.40it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 36%|        | 356/1000 [01:21<02:08,  5.03it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 74, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 36%|        | 358/1000 [01:21<01:42,  6.27it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 36%|        | 360/1000 [01:22<02:47,  3.81it/s]Input length of input_ids is
75, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 36%|        | 364/1000 [01:24<03:44,  2.83it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 37%|        | 366/1000 [01:24<03:19,  3.17it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 37%|        | 367/1000 [01:24<02:46,  3.79it/s]Input length of input_ids is
60, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 37%|        | 369/1000 [01:25<02:40,  3.94it/s]Input length of input_ids is
68, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 37%|        | 371/1000 [01:25<02:44,  3.81it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 37%|        | 373/1000 [01:25<02:01,  5.18it/s]Input length of input_ids is
```

```
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 38%|        | 375/1000 [01:26<01:37,  6.44it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 43, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 38%|        | 377/1000 [01:26<01:24,  7.38it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 38%|        | 382/1000 [01:27<02:00,  5.14it/s]Input length of input_ids is
59, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 38%|        | 384/1000 [01:27<01:34,  6.49it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 39%|        | 386/1000 [01:27<01:20,  7.61it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 46, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 39%|        | 388/1000 [01:27<01:13,  8.33it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 39%|        | 391/1000 [01:28<01:24,  7.20it/s]Input length of input_ids is
69, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 39%|        | 392/1000 [01:28<01:21,  7.44it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 39%|        | 394/1000 [01:28<01:12,  8.37it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 72, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 40%|        | 396/1000 [01:28<01:08,  8.86it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
Input length of input_ids is 31, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 40%|        | 398/1000 [01:29<01:03,  9.44it/s]Input length of input_ids is
61, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 40%|        | 400/1000 [01:29<01:02,  9.68it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 40%|        | 402/1000 [01:29<00:59, 10.02it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 72, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 40%|        | 404/1000 [01:29<00:59, 10.02it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 41%|        | 410/1000 [01:30<01:20,  7.32it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 41%|        | 413/1000 [01:31<01:34,  6.24it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 41%|        | 414/1000 [01:31<01:26,  6.74it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 72, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 42%|        | 417/1000 [01:32<02:17,  4.23it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 42%|        | 419/1000 [01:32<01:47,  5.42it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 42%|        | 421/1000 [01:32<01:28,  6.54it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 42%|        | 423/1000 [01:32<01:49,  5.28it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
 42%|      | 424/1000 [01:33<01:39,  5.80it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 34, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 43%|      | 426/1000 [01:33<01:22,  6.93it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 43%|      | 428/1000 [01:33<01:12,  7.84it/s]Input length of input_ids is
30, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 43%|      | 431/1000 [01:34<02:01,  4.68it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 43%|      | 433/1000 [01:35<03:10,  2.97it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 44%|      | 435/1000 [01:35<02:29,  3.78it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 44%|      | 437/1000 [01:36<02:29,  3.78it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 44%|      | 440/1000 [01:36<02:35,  3.61it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 67, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 44%|      | 444/1000 [01:37<01:31,  6.09it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 45%|      | 446/1000 [01:37<01:17,  7.16it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 45%|      | 448/1000 [01:38<02:35,  3.54it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 45%|      | 451/1000 [01:39<02:46,  3.30it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 45%|      | 452/1000 [01:39<02:22,  3.84it/s]Input length of input_ids is
```

```
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 45%|        | 453/1000 [01:39<02:02,  4.46it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 48, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 46%|        | 455/1000 [01:39<01:33,  5.80it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 46%|        | 458/1000 [01:40<02:32,  3.56it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 46%|        | 461/1000 [01:41<02:39,  3.38it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 67, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 46%|        | 463/1000 [01:41<02:00,  4.45it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 46%|        | 465/1000 [01:42<02:09,  4.14it/s]Input length of input_ids is
72, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 47%|        | 467/1000 [01:43<02:43,  3.26it/s]Input length of input_ids is
62, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 47%|        | 468/1000 [01:43<02:14,  3.97it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 47%|        | 469/1000 [01:43<01:51,  4.77it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 65, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 47%|        | 471/1000 [01:43<01:25,  6.19it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 47%|        | 474/1000 [01:44<02:59,  2.93it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 48%|        | 476/1000 [01:45<02:02,  4.29it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 48%|        | 478/1000 [01:45<01:33,  5.55it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 48%|        | 480/1000 [01:45<01:17,  6.68it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 48%|        | 482/1000 [01:45<01:07,  7.67it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 48%|        | 484/1000 [01:45<01:00,  8.48it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 49%|        | 486/1000 [01:45<00:56,  9.17it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 44, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 49%|        | 488/1000 [01:46<00:53,  9.57it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 49%|        | 492/1000 [01:48<02:54,  2.91it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 49%|        | 494/1000 [01:48<03:07,  2.69it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 50%|        | 496/1000 [01:50<03:55,  2.14it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 32, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 50%|        | 498/1000 [01:50<02:52,  2.91it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 58, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 50%|        | 500/1000 [01:50<02:11,  3.79it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 50%|        | 502/1000 [01:50<01:44,  4.77it/s]Input length of input_ids is
```

```
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 50%|        | 505/1000 [01:51<02:32,  3.24it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 51%|        | 506/1000 [01:51<02:07,  3.88it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 51%|        | 508/1000 [01:52<01:33,  5.24it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 51%|        | 509/1000 [01:52<01:23,  5.85it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 51%|        | 512/1000 [01:52<01:51,  4.38it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 51%|        | 514/1000 [01:53<01:26,  5.64it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 52%|        | 518/1000 [01:55<03:51,  2.08it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 52%|        | 519/1000 [01:55<03:07,  2.56it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 52%|        | 520/1000 [01:55<02:32,  3.14it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 52%|        | 522/1000 [01:55<01:58,  4.05it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 53%|        | 527/1000 [01:58<03:07,  2.52it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 53%|        | 529/1000 [01:58<02:07,  3.70it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
```

```
 53%|       | 531/1000 [01:58<01:35,  4.89it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 53%|       | 533/1000 [01:58<01:27,  5.35it/s]Input length of input_ids is
64, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 53%|       | 534/1000 [01:58<01:19,  5.89it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 54%|       | 536/1000 [01:59<01:06,  6.96it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 32, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 54%|       | 538/1000 [01:59<00:58,  7.86it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 54%|       | 540/1000 [01:59<00:53,  8.61it/s]Input length of input_ids is
67, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 54%|       | 541/1000 [01:59<00:52,  8.73it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 54%|       | 543/1000 [02:00<01:21,  5.62it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 55%|       | 545/1000 [02:01<02:17,  3.30it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 55%|       | 547/1000 [02:01<01:40,  4.49it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 55%|       | 549/1000 [02:02<02:19,  3.23it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 55%|       | 550/1000 [02:02<02:00,  3.74it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 55%|       | 551/1000 [02:02<01:43,  4.33it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
```

```
unexpected behavior. You should consider increasing `max_new_tokens`.
 55%|       | 553/1000 [02:02<01:19,  5.63it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 56%|       | 555/1000 [02:02<01:05,  6.79it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 56%|       | 559/1000 [02:05<03:57,  1.86it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 56%|       | 561/1000 [02:05<02:42,  2.70it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 56%|       | 564/1000 [02:07<03:21,  2.17it/s]Input length of input_ids is
66, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 57%|       | 566/1000 [02:07<02:23,  3.03it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 57%|       | 568/1000 [02:08<02:09,  3.32it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 47, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 57%|       | 571/1000 [02:08<01:30,  4.72it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 57%|       | 572/1000 [02:08<01:19,  5.40it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 58, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 57%|       | 575/1000 [02:10<02:33,  2.77it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 58%|       | 577/1000 [02:10<01:47,  3.92it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 58%|       | 578/1000 [02:10<01:33,  4.49it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 58%|       | 580/1000 [02:11<01:41,  4.14it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 58%|       | 582/1000 [02:11<01:18,  5.30it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 58%|       | 584/1000 [02:11<01:05,  6.34it/s]Input length of input_ids is
65, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 58%|       | 585/1000 [02:11<01:01,  6.74it/s]Input length of input_ids is
65, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 59%|       | 586/1000 [02:11<00:58,  7.09it/s]Input length of input_ids is
56, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 59%|       | 589/1000 [02:13<02:41,  2.55it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 59%|       | 591/1000 [02:13<01:47,  3.80it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 59%|       | 593/1000 [02:13<01:19,  5.09it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 60%|       | 595/1000 [02:13<01:04,  6.33it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 68, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 60%|       | 599/1000 [02:15<02:02,  3.26it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 60%|       | 602/1000 [02:15<01:59,  3.34it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 60%|       | 604/1000 [02:16<01:36,  4.11it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 61%|       | 606/1000 [02:17<02:19,  2.82it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 53, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
```

```
 61%|       | 608/1000 [02:17<01:34,  4.14it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 54, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 61%|       | 611/1000 [02:17<01:18,  4.98it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 61%|       | 613/1000 [02:17<01:01,  6.28it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 62%|       | 615/1000 [02:18<01:06,  5.82it/s]Input length of input_ids is
52, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 62%|       | 616/1000 [02:18<01:00,  6.31it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 62%|       | 618/1000 [02:19<02:03,  3.10it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 62%|       | 620/1000 [02:19<01:32,  4.11it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 62%|       | 623/1000 [02:20<01:58,  3.19it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 63%|       | 626/1000 [02:21<01:45,  3.56it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 63%|       | 628/1000 [02:21<01:15,  4.94it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 63%|       | 632/1000 [02:23<02:17,  2.68it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 64%|       | 635/1000 [02:23<01:45,  3.47it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 64%|       | 637/1000 [02:24<01:16,  4.73it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 64%|       | 639/1000 [02:24<01:27,  4.11it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 64%|       | 642/1000 [02:25<01:22,  4.35it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 64%|       | 644/1000 [02:25<01:35,  3.71it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 64%|       | 645/1000 [02:26<01:23,  4.24it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 65%|       | 647/1000 [02:26<01:04,  5.45it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 65%|       | 649/1000 [02:26<01:23,  4.22it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 65%|       | 650/1000 [02:26<01:14,  4.72it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 51, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 65%|       | 652/1000 [02:27<00:58,  5.90it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 65%|       | 654/1000 [02:27<01:24,  4.09it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 66%|       | 656/1000 [02:28<01:07,  5.09it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 66%|       | 658/1000 [02:28<00:55,  6.12it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 66%|       | 660/1000 [02:28<01:05,  5.16it/s]Input length of input_ids is
67, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
 66%|       | 661/1000 [02:28<01:00,  5.56it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 66%|       | 662/1000 [02:29<00:55,  6.11it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 66%|       | 664/1000 [02:29<00:46,  7.29it/s]Input length of input_ids is
66, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 67%|       | 666/1000 [02:30<01:31,  3.65it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 67%|       | 667/1000 [02:30<01:17,  4.32it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 51, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 67%|       | 669/1000 [02:30<00:57,  5.76it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 67%|       | 671/1000 [02:30<01:00,  5.42it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 67%|       | 674/1000 [02:31<00:58,  5.55it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 68%|       | 675/1000 [02:31<00:53,  6.13it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 55, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 68%|       | 677/1000 [02:31<00:44,  7.32it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 68%|       | 679/1000 [02:31<00:38,  8.33it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 68%|       | 681/1000 [02:31<00:35,  8.99it/s]Input length of input_ids is
50, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 68%|       | 684/1000 [02:32<01:13,  4.32it/s]Input length of input_ids is
```

```
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 69%|       | 686/1000 [02:33<01:21,  3.87it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 50, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 69%|       | 688/1000 [02:33<01:00,  5.15it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 69%|       | 690/1000 [02:33<00:48,  6.40it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 69%|       | 692/1000 [02:34<00:41,  7.43it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 70%|       | 695/1000 [02:34<00:53,  5.74it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 70%|       | 696/1000 [02:34<00:48,  6.28it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 33, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 70%|       | 699/1000 [02:35<01:07,  4.45it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 70%|       | 701/1000 [02:36<01:17,  3.84it/s]Input length of input_ids is
57, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 70%|       | 703/1000 [02:36<01:06,  4.50it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 70%|       | 704/1000 [02:36<00:58,  5.05it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 71%|       | 707/1000 [02:37<01:22,  3.57it/s]Input length of input_ids is
62, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 71%|       | 709/1000 [02:37<01:01,  4.73it/s]Input length of input_ids is
64, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
 71%|       | 711/1000 [02:38<01:03,  4.54it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 71%|       | 713/1000 [02:38<01:12,  3.93it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 72%|       | 715/1000 [02:39<00:56,  5.07it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 58, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 72%|       | 717/1000 [02:39<00:45,  6.17it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 25, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 72%|       | 720/1000 [02:39<00:45,  6.18it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 72%|       | 722/1000 [02:39<00:38,  7.28it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 72%|       | 725/1000 [02:41<01:32,  2.98it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 73%|       | 727/1000 [02:41<01:07,  4.03it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 73%|       | 729/1000 [02:41<00:52,  5.14it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 25, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 73%|       | 732/1000 [02:42<00:45,  5.85it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 73%|       | 734/1000 [02:42<00:38,  6.95it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 68, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 74%|      | 736/1000 [02:42<00:34,  7.73it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 74%|      | 737/1000 [02:42<00:33,  7.87it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 74%|      | 740/1000 [02:43<00:43,  5.96it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 74%|      | 742/1000 [02:43<00:36,  7.10it/s]Input length of input_ids is
56, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 74%|      | 743/1000 [02:43<00:34,  7.55it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 75%|      | 746/1000 [02:45<01:27,  2.90it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 75%|      | 747/1000 [02:45<01:13,  3.46it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 75%|      | 749/1000 [02:45<00:52,  4.77it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 34, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 75%|      | 752/1000 [02:46<01:19,  3.14it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 75%|      | 754/1000 [02:46<01:10,  3.47it/s]Input length of input_ids is
70, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 76%|      | 755/1000 [02:47<01:01,  3.98it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
```

```
 76%|      | 757/1000 [02:47<00:46,  5.18it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 76%|      | 758/1000 [02:47<00:42,  5.72it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 76%|      | 760/1000 [02:47<00:34,  6.94it/s]Input length of input_ids is
66, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 76%|      | 761/1000 [02:47<00:32,  7.38it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 33, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 76%|      | 763/1000 [02:47<00:28,  8.46it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 77%|      | 769/1000 [02:50<01:36,  2.40it/s]Input length of input_ids is
73, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 77%|      | 770/1000 [02:50<01:20,  2.87it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 68, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 77%|      | 773/1000 [02:51<00:57,  3.98it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 48, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 78%|      | 775/1000 [02:51<00:42,  5.27it/s]Input length of input_ids is
50, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 78%|      | 779/1000 [02:51<00:39,  5.59it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 78%|      | 781/1000 [02:52<00:31,  6.88it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
```

```
unexpected behavior. You should consider increasing `max_new_tokens`.
 78%|      | 784/1000 [02:53<01:06,  3.26it/s]Input length of input_ids is
50, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 46, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 79%|      | 786/1000 [02:53<00:49,  4.36it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 79%|      | 791/1000 [02:55<01:12,  2.86it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 79%|      | 793/1000 [02:55<00:50,  4.07it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 80%|      | 796/1000 [02:57<01:42,  2.00it/s]Input length of input_ids is
84, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 80%|      | 798/1000 [02:57<01:18,  2.57it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 80%|      | 802/1000 [02:58<01:10,  2.81it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 80%|      | 804/1000 [02:59<00:48,  4.04it/s]Input length of input_ids is
69, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 81%|      | 806/1000 [03:00<01:19,  2.43it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 81%|      | 807/1000 [03:00<01:03,  3.03it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 81%|      | 808/1000 [03:00<00:51,  3.72it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 44, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 81%|      | 810/1000 [03:00<00:36,  5.14it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 43, but `max_length` is set to 20. This can lead to
```

unexpected behavior. You should consider increasing `max_new_tokens`.
 81%| | 812/1000 [03:00<00:29, 6.39it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 81%| | 813/1000 [03:00<00:27, 6.90it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 33, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 82%| | 815/1000 [03:01<00:23, 7.88it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 82%| | 817/1000 [03:01<00:38, 4.75it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 82%| | 818/1000 [03:01<00:34, 5.29it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 32, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 82%| | 820/1000 [03:02<00:27, 6.52it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 33, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 82%| | 823/1000 [03:02<00:26, 6.79it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 82%| | 825/1000 [03:02<00:22, 7.82it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 83%| | 827/1000 [03:03<00:49, 3.51it/s]Input length of input_ids is
69, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 83%| | 829/1000 [03:04<00:55, 3.08it/s]Input length of input_ids is
66, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 83%| | 830/1000 [03:04<00:46, 3.63it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 83%| | 832/1000 [03:04<00:34, 4.93it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 83%| | 834/1000 [03:05<00:39, 4.20it/s]Input length of input_ids is

```
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 84%|      | 835/1000 [03:05<00:34,  4.77it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 84%|      | 836/1000 [03:05<00:30,  5.38it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 53, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 84%|      | 838/1000 [03:05<00:24,  6.71it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 84%|      | 840/1000 [03:06<00:54,  2.93it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 84%|      | 841/1000 [03:06<00:44,  3.57it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 84%|      | 844/1000 [03:07<00:35,  4.39it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 85%|      | 848/1000 [03:10<01:19,  1.91it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 85%|      | 850/1000 [03:10<00:52,  2.85it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 86%|      | 855/1000 [03:12<00:55,  2.59it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 86%|      | 856/1000 [03:12<00:45,  3.20it/s]Input length of input_ids is
66, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 86%|      | 857/1000 [03:12<00:36,  3.89it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 86%|      | 860/1000 [03:13<00:41,  3.40it/s]Input length of input_ids is
65, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 86%|      | 861/1000 [03:13<00:34,  4.00it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 86%|      | 863/1000 [03:13<00:25,  5.41it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 86%|      | 865/1000 [03:13<00:20,  6.56it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 87%|      | 870/1000 [03:14<00:26,  5.00it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 67, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 87%|      | 873/1000 [03:16<00:45,  2.80it/s]Input length of input_ids is
56, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 88%|      | 878/1000 [03:18<01:07,  1.80it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 88%|      | 880/1000 [03:18<00:48,  2.48it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 88%|      | 882/1000 [03:18<00:33,  3.57it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 89%|      | 886/1000 [03:20<00:45,  2.48it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 89%|      | 887/1000 [03:20<00:37,  3.03it/s]Input length of input_ids is
52, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 89%|      | 889/1000 [03:21<00:31,  3.58it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 89%|      | 892/1000 [03:22<00:36,  2.99it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 89%|      | 894/1000 [03:23<00:46,  2.28it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 90%|      | 899/1000 [03:25<00:42,  2.35it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 90%|      | 902/1000 [03:26<00:35,  2.79it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 90%|      | 904/1000 [03:26<00:24,  3.93it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 91%|      | 906/1000 [03:26<00:18,  5.09it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 91%|      | 908/1000 [03:27<00:27,  3.40it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 91%|      | 911/1000 [03:28<00:29,  3.05it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 91%|      | 914/1000 [03:29<00:36,  2.33it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 92%|      | 916/1000 [03:30<00:30,  2.79it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 92%|      | 919/1000 [03:31<00:26,  3.02it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 92%|      | 921/1000 [03:31<00:19,  4.07it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 41, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 92%|      | 924/1000 [03:31<00:17,  4.33it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 93%|      | 926/1000 [03:32<00:22,  3.30it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 93%|      | 928/1000 [03:33<00:16,  4.31it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 93%|      | 930/1000 [03:33<00:13,  5.37it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 52, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 93%|      | 932/1000 [03:33<00:10,  6.37it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 93%|      | 934/1000 [03:34<00:17,  3.77it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 94%|      | 936/1000 [03:34<00:13,  4.71it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 47, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 94%|      | 939/1000 [03:35<00:13,  4.69it/s]Input length of input_ids is
57, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 94%|      | 941/1000 [03:35<00:12,  4.74it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 94%|      | 943/1000 [03:35<00:10,  5.70it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 95%|      | 946/1000 [03:37<00:19,  2.78it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 65, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 95%|      | 948/1000 [03:37<00:13,  3.89it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 95%|      | 950/1000 [03:37<00:11,  4.39it/s]Input length of input_ids is
30, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
 95%|      | 951/1000 [03:37<00:09,  4.92it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 95%|      | 953/1000 [03:37<00:07,  6.21it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 25, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 96%|      | 955/1000 [03:38<00:06,  7.32it/s]Input length of input_ids is
68, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 96%|      | 956/1000 [03:38<00:05,  7.62it/s]Input length of input_ids is
52, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 96%|      | 958/1000 [03:38<00:08,  4.85it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 96%|      | 959/1000 [03:38<00:07,  5.42it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 96%|      | 961/1000 [03:39<00:05,  6.69it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 96%|      | 963/1000 [03:39<00:04,  7.74it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 97%|      | 966/1000 [03:39<00:05,  5.85it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 97%|      | 968/1000 [03:40<00:04,  6.51it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 41, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 97%|      | 970/1000 [03:40<00:04,  7.44it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 97%|      | 972/1000 [03:41<00:07,  3.56it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
```

```
 98%|     | 976/1000 [03:43<00:09,  2.59it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 98%|     | 977/1000 [03:43<00:07,  3.10it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 98%|     | 980/1000 [03:44<00:06,  2.86it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 98%|     | 982/1000 [03:45<00:09,  1.89it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 98%|     | 984/1000 [03:46<00:07,  2.13it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 98%|     | 985/1000 [03:46<00:05,  2.77it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 99%|     | 988/1000 [03:47<00:05,  2.30it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 99%|     | 990/1000 [03:47<00:02,  3.40it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 99%|     | 992/1000 [03:47<00:01,  4.57it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 99%|     | 993/1000 [03:48<00:01,  5.09it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
100%|     | 995/1000 [03:48<00:00,  6.40it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
100%|     | 997/1000 [03:48<00:00,  7.51it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
100%|     | 1000/1000 [03:49<00:00,  4.36it/s]

Average BLEU score on the test set: 2.1448724179416255e-05
```

BLEU: Checkpoint

[ ]:
```
from datasets import load_dataset
from transformers import (
    AutoModelForCausalLM,
    AutoTokenizer,
    pipeline,
)
import nltk
from nltk.translate.bleu_score import sentence_bleu, SmoothingFunction
from tqdm import tqdm
import torch
import re

# Install NLTK if not already installed and download BLEU's tokenizer model
nltk.download('punkt')

# Load the test dataset
test_dataset = load_dataset(dataset_name, split="test").select(range(1000))

# Load your trained model
base_model = AutoModelForCausalLM.from_pretrained(
    model_name,
    low_cpu_mem_usage=True,
    return_dict=True,
    torch_dtype=torch.float16,
    device_map=device_map,
)
# Merge fine-tuned model
model = PeftModel.from_pretrained(base_model, os.path.join(output_dir,␣
 ↪'checkpoint-11000'))
model = model.merge_and_unload()

# Load the tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)
tokenizer.pad_token = tokenizer.eos_token
tokenizer.padding_side = "right"

# Create generation pipeline
translation_pipeline = pipeline("text-generation", model=model,␣
 ↪tokenizer=tokenizer, device=0) # Adjust device as needed

# Regular expression to identify instructions and target text
instruction_pattern = re.compile(r'<s>\[INST\] (.+?) \[/INST\](.+)<\/s>')

# Calculate BLEU scores
bleu_scores = []
```

```
# Define smoothing function for nltk BLEU calculation to handle potential zero␣
 ↪n-gram counts
chencherry = SmoothingFunction()

for instance in tqdm(test_dataset):
    formatted_text = instance['formatted_text']

    # Match the instruction pattern to separate the source and target
    match = instruction_pattern.match(formatted_text)
    if match:
        source_text, target_text = match.groups()
        target_text = [nltk.word_tokenize(target_text.strip())] # Tokenize␣
 ↪reference text

        # Generate the translation
        translated = translation_pipeline(source_text)[0]['generated_text']

        # Tokenize the predicted text
        predicted_tokens = nltk.word_tokenize(translated)

        # Calculate BLEU score for this instance, with smoothing
        bleu_score = sentence_bleu(target_text, predicted_tokens,␣
 ↪smoothing_function=chencherry.method1)
        bleu_scores.append(bleu_score)

# Calculate the average BLEU score over all instances
average_bleu = sum(bleu_scores) / len(bleu_scores)
print("Average BLEU score on the test set:", average_bleu)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
```

```
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
```

```
Xformers is not installed correctly. If you want to use
memory_efficient_attention to accelerate training use the following command to
install Xformers
pip install xformers.
  0%|          | 0/1000 [00:00<?, ?it/s]/usr/local/lib/python3.10/dist-
packages/transformers/generation/utils.py:1270: UserWarning: You have modified
the pretrained model configuration to control generation. This is a deprecated
strategy to control generation and will be removed soon, in a future version.
Please use a generation configuration file (see
https://huggingface.co/docs/transformers/main_classes/text_generation )
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/transformers/generation/utils.py:1369:
UserWarning: Using `max_length`'s default (20) to control the generation length.
This behaviour is deprecated and will be removed from the config in v5 of
```

```
Transformers -- we recommend using `max_new_tokens` to control the maximum
length of the generation.
  warnings.warn(
Input length of input_ids is 50, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  0%|          | 2/1000 [00:00<07:03,  2.36it/s]Input length of input_ids is 41,
but `max_length` is set to 20. This can lead to unexpected behavior. You should
consider increasing `max_new_tokens`.
  0%|          | 3/1000 [00:00<04:36,  3.61it/s]Input length of input_ids is 23,
but `max_length` is set to 20. This can lead to unexpected behavior. You should
consider increasing `max_new_tokens`.
Input length of input_ids is 42, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  1%|          | 9/1000 [00:03<07:34,  2.18it/s]Input length of input_ids is 25,
but `max_length` is set to 20. This can lead to unexpected behavior. You should
consider increasing `max_new_tokens`.
/usr/local/lib/python3.10/dist-packages/transformers/pipelines/base.py:1083:
UserWarning: You seem to be using the pipelines sequentially on GPU. In order to
maximize efficiency please use a dataset
  warnings.warn(
  1%|          | 11/1000 [00:03<04:45,  3.46it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  1%|          | 13/1000 [00:03<03:29,  4.71it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 51, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  2%|          | 16/1000 [00:04<03:04,  5.34it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  2%|          | 18/1000 [00:04<03:11,  5.13it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  2%|          | 20/1000 [00:04<02:36,  6.26it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 50, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  2%|          | 22/1000 [00:04<02:13,  7.31it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 70, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
```

```
  2%|         | 24/1000 [00:05<02:01,  8.04it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  3%|         | 26/1000 [00:05<03:30,  4.64it/s]Input length of input_ids is
74, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  3%|         | 27/1000 [00:05<03:07,  5.19it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 47, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  3%|         | 29/1000 [00:06<02:28,  6.53it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  3%|         | 32/1000 [00:07<05:07,  3.15it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  3%|         | 34/1000 [00:08<05:44,  2.80it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 37/1000 [00:09<05:33,  2.89it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 38/1000 [00:09<04:41,  3.42it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 32, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  4%|         | 41/1000 [00:10<05:17,  3.02it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 42/1000 [00:10<04:25,  3.61it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 44/1000 [00:11<05:37,  2.83it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  4%|         | 45/1000 [00:11<04:45,  3.34it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 31, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  5%|         | 47/1000 [00:11<03:30,  4.54it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  5%|         | 48/1000 [00:11<03:05,  5.14it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  5%|          | 50/1000 [00:11<02:27,  6.43it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  5%|          | 52/1000 [00:12<02:07,  7.46it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  5%|          | 54/1000 [00:12<01:53,  8.35it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 60, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  6%|          | 56/1000 [00:12<01:46,  8.90it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  6%|          | 58/1000 [00:12<01:39,  9.44it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  6%|          | 63/1000 [00:14<05:12,  2.99it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  6%|          | 64/1000 [00:14<04:28,  3.49it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  7%|          | 66/1000 [00:15<03:49,  4.06it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 60, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  7%|          | 68/1000 [00:15<03:00,  5.17it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  7%|          | 69/1000 [00:15<02:44,  5.67it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 43, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  7%|          | 71/1000 [00:15<02:17,  6.74it/s]Input length of input_ids is
64, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
  7%|         | 73/1000 [00:15<02:23,  6.44it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 70, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  8%|         | 77/1000 [00:18<07:23,  2.08it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  8%|         | 78/1000 [00:18<05:54,  2.60it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  8%|         | 80/1000 [00:18<04:01,  3.82it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 65, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  8%|         | 82/1000 [00:18<03:07,  4.90it/s]Input length of input_ids is
59, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  8%|         | 85/1000 [00:19<02:39,  5.75it/s]Input length of input_ids is
30, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  9%|         | 86/1000 [00:19<02:25,  6.30it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  9%|         | 88/1000 [00:19<02:00,  7.56it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  9%|         | 91/1000 [00:20<03:32,  4.27it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
  9%|         | 92/1000 [00:20<03:06,  4.87it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
  9%|         | 94/1000 [00:20<02:25,  6.24it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 10%|         | 96/1000 [00:20<02:38,  5.69it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 10%|         | 97/1000 [00:21<02:24,  6.23it/s]Input length of input_ids is
```

```
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 51, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 10%|          | 99/1000 [00:21<02:03,  7.31it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 10%|          | 101/1000 [00:22<05:16,  2.84it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 54, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 10%|          | 103/1000 [00:22<04:00,  3.73it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 11%|          | 108/1000 [00:25<06:37,  2.24it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 11%|          | 110/1000 [00:25<04:34,  3.24it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 11%|          | 112/1000 [00:26<06:26,  2.30it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 11%|          | 114/1000 [00:26<04:43,  3.13it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 12%|          | 117/1000 [00:28<05:18,  2.77it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 71, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 12%|          | 119/1000 [00:28<03:58,  3.70it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 12%|          | 122/1000 [00:29<04:06,  3.55it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 12%|          | 123/1000 [00:29<03:33,  4.11it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 12%|          | 125/1000 [00:29<04:17,  3.40it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
 13%|         | 126/1000 [00:30<03:32,  4.11it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 13%|         | 127/1000 [00:30<02:58,  4.90it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 13%|         | 130/1000 [00:30<03:16,  4.43it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 13%|         | 132/1000 [00:31<02:24,  5.99it/s]Input length of input_ids is
72, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 13%|         | 134/1000 [00:31<03:38,  3.97it/s]Input length of input_ids is
60, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 14%|         | 136/1000 [00:32<03:05,  4.67it/s]Input length of input_ids is
72, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 14%|         | 138/1000 [00:32<03:35,  3.99it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 14%|         | 140/1000 [00:32<02:36,  5.51it/s]Input length of input_ids is
59, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 14%|         | 142/1000 [00:33<03:00,  4.75it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 14%|         | 143/1000 [00:33<02:40,  5.34it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 14%|         | 145/1000 [00:33<02:55,  4.86it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 56, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 15%|         | 147/1000 [00:33<02:15,  6.28it/s]Input length of input_ids is
30, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 15%|         | 149/1000 [00:34<01:53,  7.52it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 43, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 15%|         | 152/1000 [00:35<03:25,  4.13it/s]Input length of input_ids is
```

```
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 59, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 15%|         | 154/1000 [00:35<02:40,  5.26it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 43, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 16%|         | 156/1000 [00:35<02:12,  6.35it/s]Input length of input_ids is
58, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 16%|         | 157/1000 [00:35<02:04,  6.80it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 16%|         | 159/1000 [00:36<02:41,  5.19it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 16%|         | 161/1000 [00:36<02:13,  6.28it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 57, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 17%|         | 166/1000 [00:37<03:55,  3.54it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 17%|         | 167/1000 [00:37<03:23,  4.10it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 17%|         | 169/1000 [00:38<04:36,  3.00it/s]Input length of input_ids is
56, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 17%|         | 170/1000 [00:39<03:55,  3.52it/s]Input length of input_ids is
67, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 17%|         | 171/1000 [00:39<03:21,  4.11it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 50, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 17%|         | 173/1000 [00:39<02:35,  5.33it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 18%|         | 175/1000 [00:40<04:04,  3.38it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
 18%|         | 177/1000 [00:40<03:30,  3.91it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 18%|         | 179/1000 [00:40<02:46,  4.93it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 18%|         | 182/1000 [00:42<05:43,  2.38it/s]Input length of input_ids is
57, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 18%|         | 183/1000 [00:42<04:48,  2.83it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 34, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 18%|         | 185/1000 [00:43<03:27,  3.92it/s]Input length of input_ids is
72, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 19%|         | 188/1000 [00:44<04:58,  2.72it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 19%|         | 189/1000 [00:44<04:11,  3.23it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 25, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 19%|         | 193/1000 [00:45<03:30,  3.83it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 20%|         | 195/1000 [00:45<03:05,  4.33it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 20%|         | 196/1000 [00:45<02:43,  4.92it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 20%|         | 198/1000 [00:46<02:43,  4.90it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 20%|         | 200/1000 [00:46<02:12,  6.02it/s]Input length of input_ids is
58, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 20%|         | 202/1000 [00:46<01:53,  7.05it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
 20%|         | 204/1000 [00:46<01:39,  8.01it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 21%|         | 206/1000 [00:46<01:31,  8.67it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 21%|         | 207/1000 [00:46<01:29,  8.85it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 21%|         | 209/1000 [00:47<01:23,  9.43it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 21%|         | 213/1000 [00:48<02:35,  5.05it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 22%|         | 215/1000 [00:48<02:04,  6.29it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 22%|         | 217/1000 [00:49<03:56,  3.31it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 22%|         | 218/1000 [00:49<03:25,  3.80it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 50, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 22%|         | 222/1000 [00:51<05:12,  2.49it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 22%|         | 224/1000 [00:52<05:48,  2.22it/s]Input length of input_ids is
75, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 23%|         | 226/1000 [00:53<05:50,  2.21it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 23%|         | 227/1000 [00:53<04:32,  2.84it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 23%|         | 230/1000 [00:54<04:45,  2.70it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
 23%|        | 232/1000 [00:54<04:30,  2.84it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 23%|        | 234/1000 [00:55<03:54,  3.27it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 47, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 24%|        | 236/1000 [00:55<02:52,  4.44it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 24%|        | 238/1000 [00:55<02:15,  5.62it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 50, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 24%|        | 240/1000 [00:55<01:54,  6.66it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 34, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 24%|        | 242/1000 [00:55<01:39,  7.60it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 24%|        | 244/1000 [00:56<01:30,  8.38it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 25%|        | 246/1000 [00:56<01:24,  8.96it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 25%|        | 248/1000 [00:56<02:02,  6.14it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 25%|        | 250/1000 [00:57<01:59,  6.29it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 25%|        | 253/1000 [00:58<02:56,  4.24it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 26%|        | 256/1000 [00:59<04:10,  2.97it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 26%|        | 258/1000 [00:59<03:27,  3.57it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
 26%|         | 261/1000 [01:01<05:07,  2.40it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 26%|         | 263/1000 [01:02<04:35,  2.67it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 26%|         | 265/1000 [01:02<03:46,  3.24it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 27%|         | 266/1000 [01:02<03:16,  3.73it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 27%|         | 268/1000 [01:03<03:40,  3.31it/s]Input length of input_ids is
63, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 27%|         | 269/1000 [01:03<03:11,  3.82it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 27%|         | 271/1000 [01:03<02:26,  4.98it/s]Input length of input_ids is
58, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 27%|         | 273/1000 [01:04<02:49,  4.29it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 28%|         | 275/1000 [01:04<02:15,  5.34it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 28%|         | 278/1000 [01:04<02:00,  5.97it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 33, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 28%|         | 280/1000 [01:04<01:41,  7.09it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 28%|         | 282/1000 [01:05<01:28,  8.11it/s]Input length of input_ids is
63, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 28%|         | 284/1000 [01:05<01:29,  7.99it/s]Input length of input_ids is
```

```
50, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 29%|         | 287/1000 [01:06<02:32,  4.69it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 29%|         | 289/1000 [01:06<01:59,  5.96it/s]Input length of input_ids is
67, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 29%|         | 290/1000 [01:06<01:50,  6.44it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 29%|         | 292/1000 [01:06<01:32,  7.65it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 29%|         | 293/1000 [01:06<01:27,  8.04it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 30%|         | 295/1000 [01:06<01:18,  9.01it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 30%|         | 298/1000 [01:08<03:04,  3.80it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 30%|         | 302/1000 [01:09<03:36,  3.23it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 30%|         | 303/1000 [01:09<03:01,  3.84it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 30%|         | 305/1000 [01:09<02:36,  4.45it/s]Input length of input_ids is
63, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 31%|         | 307/1000 [01:10<02:53,  3.99it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 46, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 31%|         | 309/1000 [01:10<02:16,  5.06it/s]Input length of input_ids is
76, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
 31%|         | 311/1000 [01:10<02:08,  5.38it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 31%|        | 313/1000 [01:10<01:43,  6.66it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 31%|        | 314/1000 [01:11<01:35,  7.15it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 32%|        | 316/1000 [01:11<02:45,  4.12it/s]Input length of input_ids is
75, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 32%|        | 317/1000 [01:11<02:27,  4.62it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 32%|        | 320/1000 [01:12<02:24,  4.70it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 48, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 32%|        | 322/1000 [01:12<01:53,  5.96it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 32%|        | 324/1000 [01:12<01:35,  7.08it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 33%|        | 328/1000 [01:14<03:23,  3.30it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 33%|        | 330/1000 [01:15<03:56,  2.84it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 33%|        | 332/1000 [01:15<02:54,  3.84it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 55, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 34%|        | 335/1000 [01:16<03:02,  3.64it/s]Input length of input_ids is
64, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
 34%|        | 336/1000 [01:16<02:38,  4.19it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 34%|        | 339/1000 [01:17<04:06,  2.69it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 31, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 34%|        | 341/1000 [01:17<02:55,  3.75it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 34, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 34%|        | 345/1000 [01:19<04:27,  2.45it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 35%|        | 348/1000 [01:19<03:03,  3.55it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 35%|        | 350/1000 [01:20<02:16,  4.77it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 35%|        | 353/1000 [01:21<03:27,  3.12it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 36%|        | 355/1000 [01:21<02:25,  4.44it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 36%|        | 356/1000 [01:21<02:07,  5.07it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 74, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 36%|        | 358/1000 [01:21<01:41,  6.32it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 36%|        | 360/1000 [01:22<02:44,  3.90it/s]Input length of input_ids is
75, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 36%|        | 364/1000 [01:23<03:47,  2.80it/s]Input length of input_ids is
```

```
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 37%|        | 366/1000 [01:24<03:23,  3.11it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 37%|        | 367/1000 [01:24<02:49,  3.74it/s]Input length of input_ids is
60, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 37%|        | 369/1000 [01:25<02:41,  3.90it/s]Input length of input_ids is
68, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 37%|        | 371/1000 [01:25<02:50,  3.68it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 37%|        | 373/1000 [01:25<02:04,  5.03it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 38%|        | 375/1000 [01:25<01:39,  6.31it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 43, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 38%|        | 377/1000 [01:26<01:25,  7.27it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 38%|        | 382/1000 [01:27<02:02,  5.06it/s]Input length of input_ids is
59, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 38%|        | 384/1000 [01:27<01:36,  6.40it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 39%|        | 386/1000 [01:27<01:21,  7.53it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 46, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 39%|        | 388/1000 [01:27<01:14,  8.25it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
```

```
unexpected behavior. You should consider increasing `max_new_tokens`.
 39%|        | 391/1000 [01:28<01:25,  7.14it/s]Input length of input_ids is
69, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 39%|        | 392/1000 [01:28<01:21,  7.42it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 39%|        | 394/1000 [01:28<01:12,  8.38it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 72, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 40%|        | 396/1000 [01:28<01:08,  8.81it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 31, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 40%|        | 398/1000 [01:28<01:04,  9.39it/s]Input length of input_ids is
61, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 40%|        | 400/1000 [01:29<01:02,  9.66it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 40%|        | 402/1000 [01:29<00:59,  9.99it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 72, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 40%|        | 404/1000 [01:29<00:59, 10.07it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 41%|        | 410/1000 [01:30<01:20,  7.31it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 41%|        | 413/1000 [01:30<01:33,  6.27it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 41%|        | 414/1000 [01:31<01:26,  6.78it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
Input length of input_ids is 72, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 42%|        | 417/1000 [01:31<02:24,  4.05it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 42%|        | 419/1000 [01:32<01:50,  5.24it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 42%|        | 421/1000 [01:32<01:30,  6.37it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 42%|        | 423/1000 [01:32<01:50,  5.20it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 42%|        | 424/1000 [01:32<01:41,  5.69it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 34, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 43%|        | 426/1000 [01:33<01:24,  6.82it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 43%|        | 428/1000 [01:33<01:13,  7.75it/s]Input length of input_ids is
30, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 43%|        | 431/1000 [01:34<02:04,  4.56it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 43%|        | 433/1000 [01:35<03:11,  2.97it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 44%|        | 435/1000 [01:35<02:29,  3.77it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 44%|        | 437/1000 [01:36<02:35,  3.62it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 44%|        | 440/1000 [01:36<02:40,  3.49it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 67, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
```

```
 44%|        | 444/1000 [01:37<01:29,  6.21it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 45%|        | 446/1000 [01:37<01:15,  7.30it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 45%|        | 448/1000 [01:38<02:37,  3.52it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 45%|        | 449/1000 [01:38<02:17,  4.02it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 45%|        | 451/1000 [01:39<02:41,  3.39it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 45%|        | 452/1000 [01:39<02:20,  3.91it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 45%|        | 454/1000 [01:39<01:47,  5.09it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 46%|        | 458/1000 [01:40<02:39,  3.39it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 46%|        | 461/1000 [01:41<02:43,  3.31it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 67, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 46%|        | 463/1000 [01:41<02:01,  4.42it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 46%|        | 465/1000 [01:42<02:08,  4.15it/s]Input length of input_ids is
72, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 47%|        | 467/1000 [01:42<02:38,  3.36it/s]Input length of input_ids is
62, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 47%|        | 468/1000 [01:43<02:10,  4.08it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 47%|        | 469/1000 [01:43<01:48,  4.89it/s]Input length of input_ids is
```

```
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 65, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 47%|        | 471/1000 [01:43<01:23,  6.31it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 47%|        | 474/1000 [01:44<03:01,  2.90it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 48%|        | 476/1000 [01:44<02:03,  4.23it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 48%|        | 478/1000 [01:45<01:35,  5.48it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 48%|        | 480/1000 [01:45<01:18,  6.62it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 48%|        | 482/1000 [01:45<01:07,  7.64it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 48%|        | 484/1000 [01:45<01:01,  8.44it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 49%|        | 486/1000 [01:45<00:56,  9.12it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 44, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 49%|        | 488/1000 [01:46<00:53,  9.50it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 49%|        | 492/1000 [01:48<02:55,  2.90it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 49%|        | 494/1000 [01:48<03:13,  2.62it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
 50%|        | 496/1000 [01:50<03:59,  2.10it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 32, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 50%|        | 498/1000 [01:50<02:55,  2.87it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 58, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 50%|        | 500/1000 [01:50<02:13,  3.74it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 50%|        | 502/1000 [01:50<01:46,  4.69it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 50%|        | 505/1000 [01:51<02:37,  3.15it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 51%|        | 506/1000 [01:51<02:10,  3.78it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 51%|        | 508/1000 [01:52<01:35,  5.14it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 51%|        | 509/1000 [01:52<01:25,  5.76it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 51%|        | 512/1000 [01:53<01:55,  4.23it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 51%|        | 514/1000 [01:53<01:28,  5.51it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 52%|        | 518/1000 [01:55<03:44,  2.15it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 53, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 52%|        | 520/1000 [01:55<02:34,  3.10it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
 52%|        | 522/1000 [01:55<02:04,  3.84it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 53%|        | 527/1000 [01:58<03:04,  2.56it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 53%|        | 529/1000 [01:58<02:07,  3.70it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 53%|        | 531/1000 [01:58<01:36,  4.87it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 53%|        | 533/1000 [01:58<01:27,  5.33it/s]Input length of input_ids is
64, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 48, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 54%|        | 535/1000 [01:58<01:13,  6.35it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 54%|        | 537/1000 [01:59<01:03,  7.26it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 48, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 54%|        | 539/1000 [01:59<00:57,  8.03it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 67, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 54%|        | 541/1000 [01:59<00:53,  8.57it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 54%|        | 543/1000 [02:00<01:19,  5.78it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 55%|        | 545/1000 [02:01<02:11,  3.45it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
 55%|       | 546/1000 [02:01<01:52,  4.03it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 42, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 55%|       | 549/1000 [02:02<02:30,  2.99it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 44, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 55%|       | 551/1000 [02:02<01:49,  4.09it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 55%|       | 553/1000 [02:02<01:25,  5.24it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 56%|       | 555/1000 [02:02<01:10,  6.34it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 56%|       | 559/1000 [02:05<03:58,  1.85it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 56%|       | 561/1000 [02:06<02:44,  2.66it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 56%|       | 564/1000 [02:07<03:21,  2.17it/s]Input length of input_ids is
66, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 57%|       | 566/1000 [02:07<02:23,  3.03it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 57%|       | 568/1000 [02:08<02:06,  3.41it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 47, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 57%|       | 571/1000 [02:08<01:33,  4.57it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 57%|       | 572/1000 [02:08<01:22,  5.19it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 58, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
```

```
 57%|       | 575/1000 [02:10<02:33,  2.76it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 58%|       | 577/1000 [02:10<01:49,  3.87it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 58%|       | 578/1000 [02:10<01:35,  4.42it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 58%|       | 580/1000 [02:11<01:42,  4.11it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 58%|       | 582/1000 [02:11<01:19,  5.24it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 58%|       | 584/1000 [02:11<01:06,  6.27it/s]Input length of input_ids is
65, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 58%|       | 585/1000 [02:11<01:02,  6.65it/s]Input length of input_ids is
65, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 59%|       | 586/1000 [02:11<00:58,  7.03it/s]Input length of input_ids is
56, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 59%|       | 589/1000 [02:13<02:47,  2.46it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 59%|       | 591/1000 [02:13<01:50,  3.69it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 59%|       | 593/1000 [02:13<01:22,  4.95it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 60%|       | 595/1000 [02:13<01:05,  6.17it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 68, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 60%|       | 599/1000 [02:15<02:03,  3.24it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 60%|       | 602/1000 [02:16<02:00,  3.31it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 60%|       | 604/1000 [02:16<01:37,  4.07it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 61%|       | 606/1000 [02:17<02:05,  3.13it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 53, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 61%|       | 608/1000 [02:17<01:32,  4.22it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 54, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 61%|       | 611/1000 [02:17<01:18,  4.96it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 61%|       | 613/1000 [02:18<01:02,  6.16it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 62%|       | 615/1000 [02:18<01:04,  5.98it/s]Input length of input_ids is
52, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 62%|       | 616/1000 [02:18<00:59,  6.47it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 62%|       | 618/1000 [02:19<02:04,  3.08it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 62%|       | 620/1000 [02:19<01:33,  4.06it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 62%|       | 623/1000 [02:21<02:03,  3.04it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 63%|       | 626/1000 [02:21<01:48,  3.46it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 63%|       | 628/1000 [02:21<01:17,  4.82it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 63%|       | 632/1000 [02:23<02:18,  2.66it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 64%|       | 635/1000 [02:24<01:45,  3.45it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 64%|       | 637/1000 [02:24<01:17,  4.71it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 64%|       | 639/1000 [02:24<01:27,  4.12it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 64%|       | 642/1000 [02:25<01:21,  4.37it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 64%|       | 644/1000 [02:26<01:35,  3.72it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 64%|       | 645/1000 [02:26<01:23,  4.25it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 65%|       | 647/1000 [02:26<01:04,  5.46it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 65%|       | 649/1000 [02:27<01:24,  4.14it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 65%|       | 650/1000 [02:27<01:15,  4.63it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 51, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 65%|       | 652/1000 [02:27<01:00,  5.78it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
 65%|       | 654/1000 [02:28<01:21,  4.25it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 66%|       | 656/1000 [02:28<01:05,  5.27it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 66%|       | 658/1000 [02:28<00:54,  6.30it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 66%|       | 660/1000 [02:29<01:04,  5.24it/s]Input length of input_ids is
67, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 66%|       | 661/1000 [02:29<01:00,  5.62it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 66%|       | 662/1000 [02:29<00:55,  6.14it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 66%|       | 664/1000 [02:29<00:46,  7.30it/s]Input length of input_ids is
66, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 67%|       | 666/1000 [02:30<01:31,  3.63it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 67%|       | 667/1000 [02:30<01:17,  4.30it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 51, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 67%|       | 669/1000 [02:30<00:57,  5.73it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 67%|       | 671/1000 [02:31<01:00,  5.41it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 67%|       | 674/1000 [02:31<00:59,  5.49it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 68%|       | 675/1000 [02:31<00:53,  6.06it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 55, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
```

```
 68%|       | 677/1000 [02:31<00:44,  7.25it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 68%|       | 679/1000 [02:31<00:38,  8.23it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 68%|       | 681/1000 [02:32<00:35,  8.90it/s]Input length of input_ids is
50, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 68%|       | 684/1000 [02:33<01:15,  4.19it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 69%|       | 686/1000 [02:33<01:22,  3.81it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 69%|       | 687/1000 [02:33<01:09,  4.48it/s]Input length of input_ids is
50, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 69%|       | 689/1000 [02:34<00:52,  5.91it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 25, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 69%|       | 691/1000 [02:34<00:43,  7.11it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 69%|       | 693/1000 [02:34<00:38,  8.07it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 70%|       | 695/1000 [02:34<00:49,  6.13it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 70%|       | 696/1000 [02:35<00:46,  6.58it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 33, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 70%|       | 699/1000 [02:35<01:06,  4.51it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
 70%|       | 701/1000 [02:36<01:15,  3.95it/s]Input length of input_ids is
57, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 70%|       | 703/1000 [02:36<01:04,  4.58it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 71%|       | 707/1000 [02:37<01:23,  3.49it/s]Input length of input_ids is
62, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 71%|       | 709/1000 [02:38<01:02,  4.68it/s]Input length of input_ids is
64, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 71%|       | 711/1000 [02:38<01:04,  4.51it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 71%|      | 713/1000 [02:39<01:19,  3.62it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 72%|      | 715/1000 [02:39<00:56,  5.03it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 58, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 72%|      | 717/1000 [02:39<00:45,  6.29it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 25, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 72%|      | 720/1000 [02:39<00:44,  6.23it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 72%|      | 722/1000 [02:40<00:37,  7.36it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 72%|      | 725/1000 [02:41<01:33,  2.95it/s]Input length of input_ids is
22, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 73%|      | 727/1000 [02:42<01:08,  3.99it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 73%|      | 729/1000 [02:42<00:53,  5.10it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 25, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 73%|      | 732/1000 [02:42<00:46,  5.82it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 73%|      | 734/1000 [02:42<00:38,  6.91it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 68, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 74%|      | 736/1000 [02:43<00:34,  7.67it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 74%|      | 737/1000 [02:43<00:33,  7.85it/s]Input length of input_ids is
46, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 74%|      | 740/1000 [02:43<00:43,  5.94it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 74%|      | 742/1000 [02:43<00:36,  7.05it/s]Input length of input_ids is
56, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 74%|      | 743/1000 [02:44<00:34,  7.49it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 75%|      | 746/1000 [02:45<01:26,  2.95it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 75%|      | 747/1000 [02:45<01:12,  3.51it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 75%|      | 749/1000 [02:45<00:51,  4.83it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
Input length of input_ids is 34, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 75%|      | 752/1000 [02:46<01:18,  3.17it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 75%|      | 754/1000 [02:47<01:10,  3.49it/s]Input length of input_ids is
70, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 76%|      | 755/1000 [02:47<01:01,  4.00it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 76%|      | 757/1000 [02:47<00:46,  5.22it/s]Input length of input_ids is
53, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 76%|      | 758/1000 [02:47<00:41,  5.79it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 76%|      | 760/1000 [02:47<00:34,  7.05it/s]Input length of input_ids is
66, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 76%|      | 761/1000 [02:47<00:31,  7.47it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 33, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 76%|      | 763/1000 [02:48<00:27,  8.56it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 77%|      | 769/1000 [02:50<01:37,  2.37it/s]Input length of input_ids is
73, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 77%|      | 770/1000 [02:50<01:20,  2.85it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 68, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 77%|      | 773/1000 [02:51<00:57,  3.95it/s]Input length of input_ids is
34, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 48, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
```

```
 78%|      | 775/1000 [02:51<00:43,  5.23it/s]Input length of input_ids is
50, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 78%|      | 779/1000 [02:52<00:39,  5.58it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 78%|      | 781/1000 [02:52<00:31,  6.86it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 78%|      | 784/1000 [02:53<01:08,  3.13it/s]Input length of input_ids is
50, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 46, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 79%|      | 786/1000 [02:53<00:50,  4.22it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 30, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 79%|      | 791/1000 [02:55<01:13,  2.83it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 79%|      | 793/1000 [02:55<00:51,  4.04it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 80%|      | 796/1000 [02:57<01:42,  1.99it/s]Input length of input_ids is
84, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 80%|      | 798/1000 [02:58<01:18,  2.56it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 80%|      | 802/1000 [02:59<01:09,  2.86it/s]Input length of input_ids is
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 80%|      | 804/1000 [02:59<00:47,  4.09it/s]Input length of input_ids is
69, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 81%|      | 806/1000 [03:00<01:20,  2.40it/s]Input length of input_ids is
```

```
26, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 25, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 81%|      | 808/1000 [03:00<00:54,  3.50it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 44, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 81%|      | 810/1000 [03:00<00:40,  4.65it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 43, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 81%|      | 812/1000 [03:01<00:32,  5.80it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 81%|      | 813/1000 [03:01<00:29,  6.32it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 33, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 82%|      | 815/1000 [03:01<00:25,  7.39it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 82%|      | 817/1000 [03:02<00:38,  4.78it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 82%|      | 819/1000 [03:02<00:31,  5.83it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 82%|      | 821/1000 [03:02<00:26,  6.83it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 82%|      | 823/1000 [03:02<00:25,  6.81it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 82%|      | 825/1000 [03:02<00:22,  7.69it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 83%|      | 827/1000 [03:04<00:46,  3.73it/s]Input length of input_ids is
69, but `max_length` is set to 20. This can lead to unexpected behavior. You
```

```
should consider increasing `max_new_tokens`.
 83%|      | 829/1000 [03:04<00:52,  3.24it/s]Input length of input_ids is
66, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 83%|      | 830/1000 [03:04<00:45,  3.76it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 83%|      | 832/1000 [03:05<00:33,  5.03it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 83%|      | 834/1000 [03:05<00:42,  3.94it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 43, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 84%|      | 836/1000 [03:05<00:31,  5.21it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 53, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 84%|      | 838/1000 [03:06<00:25,  6.44it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 84%|      | 840/1000 [03:07<00:45,  3.55it/s]Input length of input_ids is
44, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 84%|      | 841/1000 [03:07<00:39,  4.06it/s]Input length of input_ids is
25, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 84%|      | 844/1000 [03:07<00:33,  4.63it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 85%|      | 848/1000 [03:10<01:17,  1.95it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 24, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 85%|      | 850/1000 [03:10<00:52,  2.87it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 86%|      | 855/1000 [03:12<00:55,  2.62it/s]Input length of input_ids is
55, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 86%|      | 856/1000 [03:12<00:44,  3.23it/s]Input length of input_ids is
```

```
66, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 86%|      | 857/1000 [03:12<00:36,  3.91it/s]Input length of input_ids is
32, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 45, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 86%|      | 860/1000 [03:13<00:39,  3.56it/s]Input length of input_ids is
65, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 86%|      | 861/1000 [03:13<00:33,  4.14it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 86%|      | 863/1000 [03:13<00:24,  5.55it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 86%|      | 865/1000 [03:14<00:20,  6.72it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 87%|      | 870/1000 [03:15<00:25,  5.06it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 67, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 87%|      | 873/1000 [03:16<00:44,  2.86it/s]Input length of input_ids is
56, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 88%|      | 878/1000 [03:18<01:03,  1.93it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 88%|      | 880/1000 [03:18<00:46,  2.58it/s]Input length of input_ids is
24, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 88%|      | 882/1000 [03:19<00:32,  3.63it/s]Input length of input_ids is
33, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 89%|      | 886/1000 [03:20<00:46,  2.45it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 52, but `max_length` is set to 20. This can lead to
```

unexpected behavior. You should consider increasing `max_new_tokens`.
 89%| | 889/1000 [03:21<00:32, 3.41it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 29, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 89%| | 892/1000 [03:22<00:37, 2.89it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 89%| | 894/1000 [03:23<00:47, 2.25it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 20, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 90%| | 899/1000 [03:25<00:43, 2.32it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 90%| | 902/1000 [03:26<00:34, 2.80it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 90%| | 904/1000 [03:26<00:23, 4.17it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 37, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 91%| | 906/1000 [03:26<00:17, 5.47it/s]Input length of input_ids is
20, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 91%| | 908/1000 [03:27<00:26, 3.47it/s]Input length of input_ids is
36, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 91%| | 911/1000 [03:28<00:28, 3.08it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 91%| | 914/1000 [03:30<00:36, 2.33it/s]Input length of input_ids is
40, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 92%| | 916/1000 [03:30<00:29, 2.80it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 92%| | 919/1000 [03:31<00:26, 3.04it/s]Input length of input_ids is
43, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 22, but `max_length` is set to 20. This can lead to

```
unexpected behavior. You should consider increasing `max_new_tokens`.
 92%|      | 921/1000 [03:31<00:19,  4.10it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 41, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 92%|      | 924/1000 [03:32<00:17,  4.34it/s]Input length of input_ids is
27, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 93%|      | 926/1000 [03:32<00:22,  3.26it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 35, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 93%|      | 928/1000 [03:33<00:16,  4.26it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 93%|      | 930/1000 [03:33<00:13,  5.32it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 52, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 93%|      | 932/1000 [03:33<00:10,  6.34it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 93%|      | 934/1000 [03:34<00:17,  3.81it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 94%|      | 935/1000 [03:34<00:15,  4.29it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 94%|      | 937/1000 [03:34<00:11,  5.42it/s]Input length of input_ids is
47, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 94%|      | 939/1000 [03:35<00:12,  4.88it/s]Input length of input_ids is
57, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 94%|      | 941/1000 [03:35<00:12,  4.91it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 94%|      | 943/1000 [03:35<00:09,  5.77it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 39, but `max_length` is set to 20. This can lead to
```

```
unexpected behavior. You should consider increasing `max_new_tokens`.
 95%|      | 946/1000 [03:37<00:18,  2.96it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 65, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 95%|      | 948/1000 [03:37<00:13,  3.98it/s]Input length of input_ids is
39, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 95%|      | 950/1000 [03:37<00:11,  4.47it/s]Input length of input_ids is
30, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 95%|      | 951/1000 [03:37<00:09,  5.00it/s]Input length of input_ids is
28, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 23, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 95%|      | 953/1000 [03:37<00:07,  6.23it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 25, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 96%|      | 955/1000 [03:38<00:06,  7.27it/s]Input length of input_ids is
68, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 96%|      | 956/1000 [03:38<00:05,  7.58it/s]Input length of input_ids is
52, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 96%|      | 958/1000 [03:38<00:08,  5.04it/s]Input length of input_ids is
45, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 96%|      | 959/1000 [03:39<00:07,  5.61it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 36, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 96%|      | 961/1000 [03:39<00:05,  6.87it/s]Input length of input_ids is
48, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 96%|      | 963/1000 [03:39<00:04,  7.86it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 97%|      | 966/1000 [03:39<00:05,  5.84it/s]Input length of input_ids is
38, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 97%|      | 968/1000 [03:40<00:04,  6.50it/s]Input length of input_ids is
```

```
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 97%|      | 969/1000 [03:40<00:04,  6.98it/s]Input length of input_ids is
41, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 26, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 97%|      | 972/1000 [03:41<00:09,  2.97it/s]Input length of input_ids is
23, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 38, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 98%|      | 976/1000 [03:43<00:09,  2.42it/s]Input length of input_ids is
54, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 42, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 98%|      | 980/1000 [03:44<00:07,  2.80it/s]Input length of input_ids is
37, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 98%|      | 982/1000 [03:45<00:09,  1.88it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 98%|      | 984/1000 [03:46<00:07,  2.06it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 98%|      | 985/1000 [03:46<00:05,  2.67it/s]Input length of input_ids is
29, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 99%|      | 988/1000 [03:47<00:05,  2.24it/s]Input length of input_ids is
51, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 21, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 99%|      | 990/1000 [03:48<00:03,  3.31it/s]Input length of input_ids is
35, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
Input length of input_ids is 27, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
 99%|      | 992/1000 [03:48<00:01,  4.46it/s]Input length of input_ids is
49, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
 99%|      | 993/1000 [03:48<00:01,  5.00it/s]Input length of input_ids is
31, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
100%|      | 995/1000 [03:48<00:00,  6.29it/s]Input length of input_ids is
21, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
```

```
Input length of input_ids is 28, but `max_length` is set to 20. This can lead to
unexpected behavior. You should consider increasing `max_new_tokens`.
100%|      | 997/1000 [03:48<00:00,  7.40it/s]Input length of input_ids is
42, but `max_length` is set to 20. This can lead to unexpected behavior. You
should consider increasing `max_new_tokens`.
100%|      | 1000/1000 [03:49<00:00,  4.35it/s]

Average BLEU score on the test set: 2.3069713109025516e-05
```

Analyzing the BLEU Scores, we see they are both very low and this can be attriburted to the fact that story generations have creative freedom meaning they will not exactly match the expected output, as the BLEU score takes n-sized grams of text and does a similarity check, we do not expect these to exactly match. Building on that since these are short stories the brevity check is likely to be a low value, but what we can do is compare the values from the baseline to the checkpoint model, and comparing these 2 values, we see that our score went up by 8% which is a significant increase from where we started.

2.5 Related Work

The project draws on previous research in the areas of generative storytelling, NLP, and user experience design as found in the Hierarchical Neural Story Generation paper. We position ourselves in a similar context by evolving LLM applications for creative content generation by applying parameter-efficient fine-tuning methods such as qLoRA to a trained LLaMA 2 model, but in a similar vein, people have created story generation LLMs through GPT-2.

2.6 Limitations

As with any machine learning model, the potential for perpetuating biases exists, and the degree of creativity remains bounded by the input data. Moreover, the complexity of processing natural language feedback to adapt storylines is yet to be fully gauged.

2.7 Improvement

Exploring additional fine-tuning methods and extending the dataset diversity could further improve the model. Incorporating feedback loops from readers to create more interactive and adaptive stories might also enhance the model's capability.

2.8 Strengths of the Paper

The paper's strength lies in its novel approach to personalized storytelling using LLMs, its rigorous evaluation process combining various quantitative and qualitative measures, and the exploration of fine-tuning techniques tailored to enhancing engaging and understandable narrative generation. Accompanying that, the use of a very large dataset of around 300,000 stories helps ensure proper identification of prompt-story correlations and trends.

2.9 Weaknesses of the Paper

The weaknesses include the challenge of scaling the personalized narrative generation process and the need for further investigation into real-time adaptability and feedback integration. Furthermore,

a notable constraint that impacts the paper's comprehensiveness is the limitation in GPU and RAM resources, hindering the exploration of other models such as Mistral. Without these comparisons, we could not understand how the proposed approach fares against alternative methodologies. The initial weakness in the inability to directly use traditional testing loss methods is a weakness, but we are working towards countering that by applying new methods of perplexity, BLEU, and ROUGE.

2.10 TODOs

- Further fine-tuning of the Llama 2 model is required
- Evals on all the checkpoints are needed, along with sample outputs for qualititative and quantitative analyses

2.11 Conclusion

This project represents a significant step towards creating more immersive, personalized, and coherent storytelling experiences facilitated by LLMs, with potential applications in entertainment, education, and beyond.