193.69.10.14

Home | ToT (4) | Qwen2_5 (1) | [BUG] RuntimeError: Can't pickle local object 'patch_vl... | RuntimeError: Can't get local object 'patch_vllm_comp...

jupyter ToT (4) Last Checkpoint: 11 hours ago

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab | Python 3 (ipykernel)

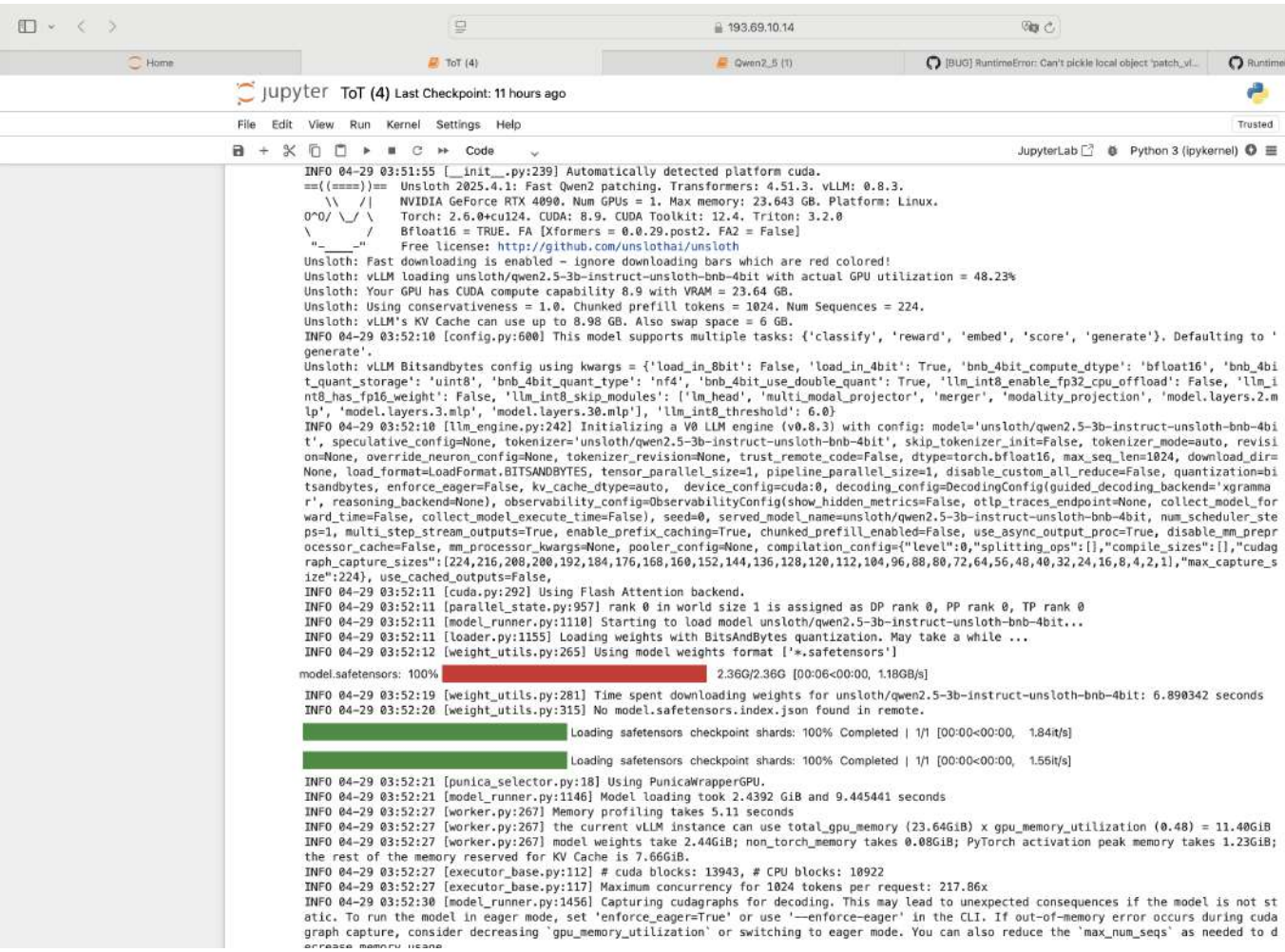
[1]:
```
import os
os.environ["VLLM_USE_V1"] = "0"
```

[2]:
```
import os
if "COLAB_" not in "".join(os.environ.keys()):
    !pip install unsloth vllm==0.8.3
else:
    # [NOTE] Do the below ONLY in Colab! Use [[pip install unsloth vllm]]
    !pip install --no-deps unsloth vllm
```

```
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.11/site-packages (from pandas->datasets>=2.16.0->unsloth) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in /opt/conda/lib/python3.11/site-packages (from pandas->datasets>=2.16.0->unsloth) (2025.2)
Requirement already satisfied: mdurl~=0.1 in /opt/conda/lib/python3.11/site-packages (from markdown-it-py>=2.2.0->rich->trl!=0.15.0,!=0.9.0,!=0.9.1,!=0.9.2,!=0.9.3,<=0.15.2,>=0.7.9->unsloth) (0.1.2)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.11/site-packages (from python-dateutil>=2.8.2->pandas->datasets>=2.16.0->unsloth) (1.16.0)
Requirement already satisfied: shellingham>=1.3.0 in /opt/conda/lib/python3.11/site-packages (from typer>=0.12.3->fastapi-cli>=0.0.5->fastapi-cli[standard]>=0.0.5; extra == "standard"->fastapi[standard]>=0.115.0->vllm==0.8.3) (1.5.4)
Using cached vllm-0.8.3-cp38-abi3-manylinux1_x86_64.whl (294.0 MB)
Installing collected packages: vllm
  Attempting uninstall: vllm
    Found existing installation: vllm 0.8.4
    Uninstalling vllm-0.8.4:
      Successfully uninstalled vllm-0.8.4
Successfully installed vllm-0.8.3
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager, possibly rendering your system unusable.It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv. Use the --root-user-action option if you know what you are doing and want to suppress this warning.
```

[3]:
```
from unsloth import FastLanguageModel, is_bfloat16_supported
import torch
max_seq_length = 1024 # Can increase for longer reasoning traces
lora_rank = 64 # Larger rank = smarter, but slower

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name = "Qwen/Qwen2.5-3B-Instruct",
    max_seq_length = max_seq_length,
    load_in_4bit = True, # False for LoRA 16bit
    fast_inference = True, # Enable vLLM fast inference
    max_lora_rank = lora_rank,
    gpu_memory_utilization = 0.5, # Reduce if out of memory
)
```

```
🦥 Unsloth: Will patch your computer to enable 2x faster free finetuning.
Unsloth: Failed to patch SmolVLMForConditionalGeneration forward function.
🦥 Unsloth Zoo will now patch everything to make training faster!
INFO 04-29 03:51:55 [__init__.py:239] Automatically detected platform cuda.
==((====))==  Unsloth 2025.4.1: Fast Qwen2 patching. Transformers: 4.51.3. vLLM: 0.8.3.
```


193.69.10.14
Home
ToT (4)
Qwen2_5 (1)
[BUG] RuntimeError: Can't pickle local object 'patch_vl...
RuntimeError: Can't get local object 'patch_vllm_comp...
jupyter ToT (4) Last Checkpoint: 11 hours ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
INFO 04-29 03:51:55 [__init__.py:239] Automatically detected platform cuda.
==((====))== Unsloth 2025.4.1: Fast Qwen2 patching. Transformers: 4.51.3. vLLM: 0.8.3.
\\ /| NVIDIA GeForce RTX 4090. Num GPUs = 1. Max memory: 23.643 GB. Platform: Linux.
O^O/ _/ \ Torch: 2.6.0+cu124. CUDA: 8.9. CUDA Toolkit: 12.4. Triton: 3.2.0
\ / Bfloat16 = TRUE. FA [Xformers = 0.0.29.post2. FA2 = False]
"-____-" Free license: http://github.com/unslothai/unsloth
Unsloth: Fast downloading is enabled - ignore downloading bars which are red colored!
Unsloth: vLLM loading unsloth/qwen2.5-3b-instruct-unsloth-bnb-4bit with actual GPU utilization = 48.23%
Unsloth: Your GPU has CUDA compute capability 8.9 with VRAM = 23.64 GB.
Unsloth: Using conservativeness = 1.0. Chunked prefill tokens = 1024. Num Sequences = 224.
Unsloth: vLLM's KV Cache can use up to 8.98 GB. Also swap space = 6 GB.
INFO 04-29 03:52:10 [config.py:600] This model supports multiple tasks: {'classify', 'reward', 'embed', 'score', 'generate'}. Defaulting to 'generate'.
Unsloth: vLLM Bitsandbytes config using kwargs = {'load_in_8bit': False, 'load_in_4bit': True, 'bnb_4bit_compute_dtype': 'bfloat16', 'bnb_4bit_quant_storage': 'uint8', 'bnb_4bit_quant_type': 'nf4', 'bnb_4bit_use_double_quant': True, 'llm_int8_enable_fp32_cpu_offload': False, 'llm_int8_has_fp16_weight': False, 'llm_int8_skip_modules': ['lm_head', 'multi_modal_projector', 'merger', 'modality_projection', 'model.layers.2.mlp', 'model.layers.3.mlp', 'model.layers.30.mlp'], 'llm_int8_threshold': 6.0}
INFO 04-29 03:52:10 [llm_engine.py:242] Initializing a V0 LLM engine (v0.8.3) with config: model='unsloth/qwen2.5-3b-instruct-unsloth-bnb-4bit', speculative_config=None, tokenizer='unsloth/qwen2.5-3b-instruct-unsloth-bnb-4bit', skip_tokenizer_init=False, tokenizer_mode=auto, revision=None, override_neuron_config=None, tokenizer_revision=None, trust_remote_code=False, dtype=torch.bfloat16, max_seq_len=1024, download_dir=None, load_format=LoadFormat.BITSANDBYTES, tensor_parallel_size=1, pipeline_parallel_size=1, disable_custom_all_reduce=False, quantization=bitsandbytes, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda:0, decoding_config=DecodingConfig(guided_decoding_backend='xgrammar', reasoning_backend=None), observability_config=ObservabilityConfig(show_hidden_metrics=False, otlp_traces_endpoint=None, collect_model_forward_time=False, collect_model_execute_time=False), seed=0, served_model_name=unsloth/qwen2.5-3b-instruct-unsloth-bnb-4bit, num_scheduler_steps=1, multi_step_stream_outputs=True, enable_prefix_caching=True, chunked_prefill_enabled=False, use_async_output_proc=True, disable_mm_preprocessor_cache=False, mm_processor_kwargs=None, pooler_config=None, compilation_config={"level":0,"splitting_ops":[],"compile_sizes":[],"cudagraph_capture_sizes":[224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,72,64,56,48,40,32,24,16,8,4,2,1],"max_capture_size":224}, use_cached_outputs=False,
INFO 04-29 03:52:11 [cuda.py:292] Using Flash Attention backend.
INFO 04-29 03:52:11 [parallel_state.py:957] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0
INFO 04-29 03:52:11 [model_runner.py:1110] Starting to load model unsloth/qwen2.5-3b-instruct-unsloth-bnb-4bit...
INFO 04-29 03:52:11 [loader.py:1155] Loading weights with BitsAndBytes quantization. May take a while ...
INFO 04-29 03:52:12 [weight_utils.py:265] Using model weights format ['*.safetensors']
model.safetensors: 100% 2.36G/2.36G [00:06<00:00, 1.18GB/s]
INFO 04-29 03:52:19 [weight_utils.py:281] Time spent downloading weights for unsloth/qwen2.5-3b-instruct-unsloth-bnb-4bit: 6.890342 seconds
INFO 04-29 03:52:20 [weight_utils.py:315] No model.safetensors.index.json found in remote.
Loading safetensors checkpoint shards: 100% Completed | 1/1 [00:00<00:00, 1.84it/s]
Loading safetensors checkpoint shards: 100% Completed | 1/1 [00:00<00:00, 1.55it/s]
INFO 04-29 03:52:21 [punica_selector.py:18] Using PunicaWrapperGPU.
INFO 04-29 03:52:21 [model_runner.py:1146] Model loading took 2.4392 GiB and 9.445441 seconds
INFO 04-29 03:52:27 [worker.py:267] Memory profiling takes 5.11 seconds
INFO 04-29 03:52:27 [worker.py:267] the current vLLM instance can use total_gpu_memory (23.64GiB) x gpu_memory_utilization (0.48) = 11.40GiB
INFO 04-29 03:52:27 [worker.py:267] model weights take 2.44GiB; non_torch_memory takes 0.08GiB; PyTorch activation peak memory takes 1.23GiB; the rest of the memory reserved for KV Cache is 7.66GiB.
INFO 04-29 03:52:27 [executor_base.py:112] # cuda blocks: 13943, # CPU blocks: 10922
INFO 04-29 03:52:27 [executor_base.py:117] Maximum concurrency for 1024 tokens per request: 217.86x
INFO 04-29 03:52:30 [model_runner.py:1456] Capturing cudagraphs for decoding. This may lead to unexpected consequences if the model is not static. To run the model in eager mode, set 'enforce_eager=True' or use '--enforce-eager' in the CLI. If out-of-memory error occurs during cuda graph capture, consider decreasing `gpu_memory_utilization` or switching to eager mode. You can also reduce the `max_num_seqs` as needed to decrease memory usage

193.69.10.14

Home | ToT (4) | Qwen2_5 (1) | [BUG] RuntimeError: Can't pickle local object 'patch_vl... | RuntimeError: Can't get local object 'patch_vllm_comp...

jupyter ToT (4) Last Checkpoint: 11 hours ago

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab | Python 3 (ipykernel)

```
ecrease memory usage.
Capturing CUDA graph shapes: 100%|██████████| 31/31 [00:35<00:00,  1.14s/it]
INFO 04-29 03:53:05 [model_runner.py:1598] Graph capturing finished in 35 secs, took 0.64 GiB
INFO 04-29 03:53:05 [llm_engine.py:448] init engine (profile, create kv cache, warmup model) took 44.32 seconds
```

[4]:
```
model = FastLanguageModel.get_peft_model(
    model,
    r = lora_rank, # Choose any number > 0 ! Suggested 8, 16, 32, 64, 128
    target_modules = [
        "q_proj", "k_proj", "v_proj", "o_proj",
        "gate_proj", "up_proj", "down_proj",
    ], # Remove QKVO if out of memory
    lora_alpha = lora_rank,
    use_gradient_checkpointing = "unsloth", # Enable long context finetuning
    random_state = 3407,
)
```

```
Unsloth 2025.4.1 patched 36 layers with 36 QKV layers, 36 O layers and 36 MLP layers.
```

[5]:
```
# 3️⃣ Build the prompt exactly as before
SYSTEM_PROMPT = """
Respond in the following format:
<reasoning>
...
</reasoning>
<answer>
...
</answer>
"""
prompt = tokenizer.apply_chat_template(
    [
        {"role": "system", "content": SYSTEM_PROMPT},
        {"role": "user",   "content": "How many r's are in strawberry?"},
    ],
    tokenize=False,
    add_generation_prompt=True,
)
```

[6]:
```
text = tokenizer.apply_chat_template([
    {"role" : "system", "content" : SYSTEM_PROMPT},
    {"role" : "user", "content" : "How many r's are in strawberry?"},
], tokenize = False, add_generation_prompt = True)

from vllm import SamplingParams
sampling_params = SamplingParams(
    temperature = 0.8,
    top_p = 0.95,
    max_tokens = 1024,
```

193.69.10.14

Home | ToT (4) | Qwen2_5 (1) | [BUG] RuntimeError: Can't pickle local object 'patch_vl... | RuntimeError: Can't get local object 'patch_vllm_comp...

jupyter ToT (4) Last Checkpoint: 11 hours ago

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab Python 3 (ipykernel)

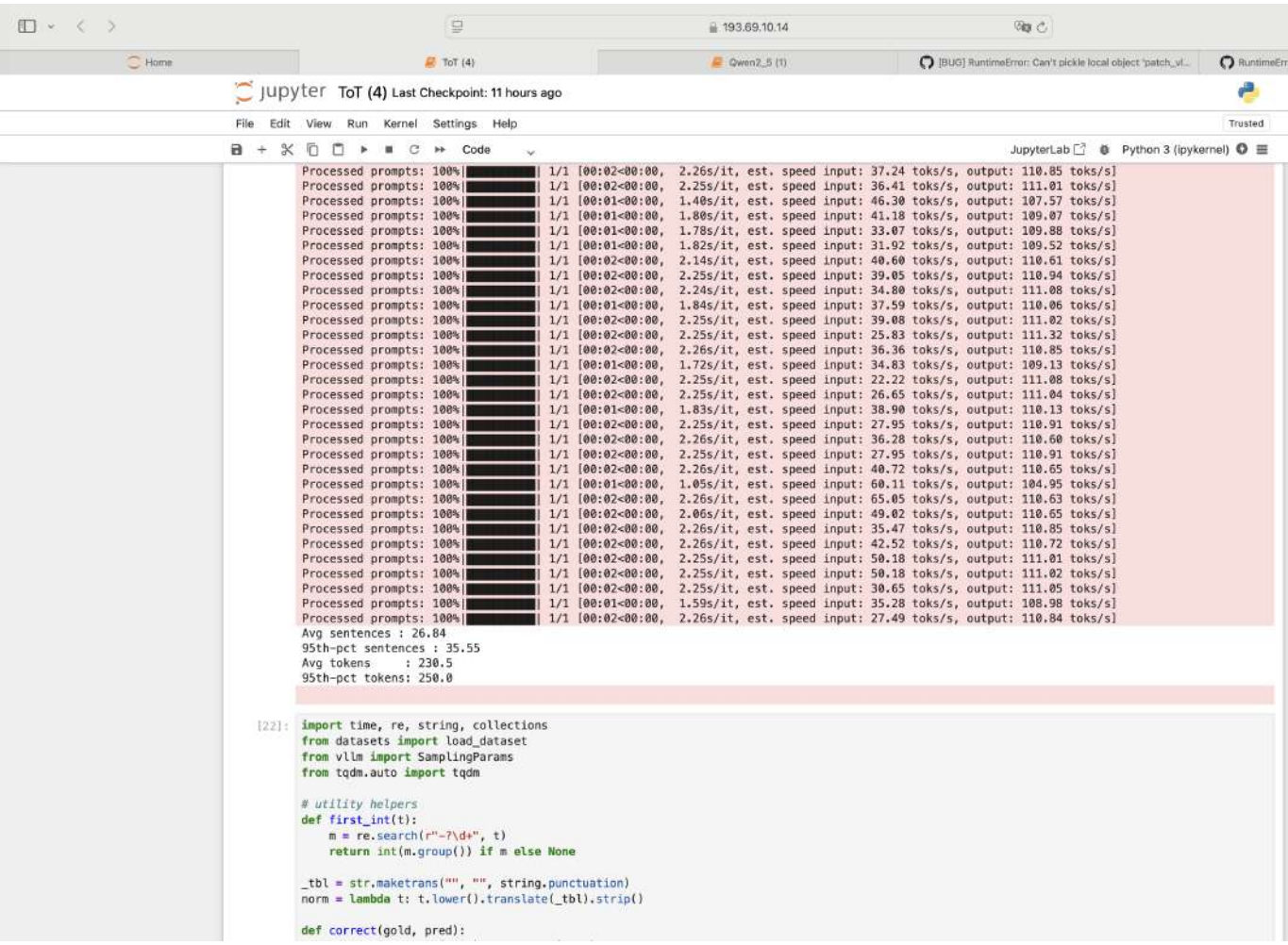
```
    max_tokens = 1024,
)
output = model.fast_generate(
    text,
    sampling_params = sampling_params,
    lora_request = model.load_lora("grpo_saved_lora"),
)[0].outputs[0].text

output
```

```
Processed prompts: 100%|██████████| 1/1 [00:00<00:00,  2.16it/s, est. speed input: 93.28 toks/s, output: 123.65 toks/s]
```

[6]:
```
'<reasoning>\nTo determine how many r\'s are in the word "strawberry," I need to count each occurrence of the letter \'r\' within the word.\n</reasoning>\n<answer>\nThere are three \'r\'s in the word "strawberry."'
```

[14]:
```
# Measure sentence & token stats for the model's own ToT outputs on 50 test Qs
from datasets import load_dataset
import re, statistics, numpy as np
from vllm import SamplingParams

N            = 50
sample_ds    = load_dataset("openai/gsm8k", "main", split=f"test[:{N}]")
sent_lens    = []
token_lens   = []

sp = SamplingParams(temperature=0.3, top_p=1.0, max_tokens=250)
for ex in sample_ds:
    q  = ex["question"]
    p  = tokenizer.apply_chat_template(
            [{"role":"system","content":"You are a helpful assistant."},
             {"role":"user","content":q}],
            tokenize=False, add_generation_prompt=True
         )
    txt = model.fast_generate(p, sampling_params=sp,
                              lora_request=lora_handle)[0].outputs[0].text
    # crude split: one sentence ≈ ends with period or newline
    sent_lens.append(len(re.split(r"[.\n]", txt)))
    token_lens.append(len(tokenizer(txt).input_ids))

print("Avg sentences :", statistics.mean(sent_lens))
print("95th-pct sentences :", np.percentile(sent_lens, 95))
print("Avg tokens    :", statistics.mean(token_lens))
print("95th-pct tokens:", np.percentile(token_lens, 95))
```

```
Processed prompts: 100%|██████████| 1/1 [00:02<00:00,  2.26s/it, est. speed input: 37.19 toks/s, output: 110.68 toks/s]
Processed prompts: 100%|██████████| 1/1 [00:01<00:00,  1.32s/it, est. speed input: 34.20 toks/s, output: 107.93 toks/s]
Processed prompts: 100%|██████████| 1/1 [00:02<00:00,  2.25s/it, est. speed input: 33.77 toks/s, output: 111.08 toks/s]
Processed prompts: 100%|██████████| 1/1 [00:01<00:00,  1.97s/it, est. speed input: 27.38 toks/s, output: 110.04 toks/s]
Processed prompts: 100%|██████████| 1/1 [00:02<00:00,  2.25s/it, est. speed input: 57.68 toks/s, output: 110.93 toks/s]
Processed prompts: 100%|██████████| 1/1 [00:02<00:00,  2.26s/it, est. speed input: 33.20 toks/s, output: 110.65 toks/s]
```


193.69.10.14
Home
ToT (4)
Qwen2_5 (1)
[BUG] RuntimeError: Can't pickle local object 'patch_vl...
RuntimeError: Can't get local object 'patch_vllm_comp...
jupyter ToT (4) Last Checkpoint: 11 hours ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Processed prompts: 100%| 1/1 [00:02<00:00, 2.26s/it, est. speed input: 37.24 toks/s, output: 110.85 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.25s/it, est. speed input: 36.41 toks/s, output: 111.01 toks/s]
Processed prompts: 100%| 1/1 [00:01<00:00, 1.40s/it, est. speed input: 46.30 toks/s, output: 107.57 toks/s]
Processed prompts: 100%| 1/1 [00:01<00:00, 1.80s/it, est. speed input: 41.18 toks/s, output: 109.07 toks/s]
Processed prompts: 100%| 1/1 [00:01<00:00, 1.78s/it, est. speed input: 33.07 toks/s, output: 109.88 toks/s]
Processed prompts: 100%| 1/1 [00:01<00:00, 1.82s/it, est. speed input: 31.92 toks/s, output: 109.52 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.14s/it, est. speed input: 40.60 toks/s, output: 110.61 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.25s/it, est. speed input: 39.05 toks/s, output: 110.94 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.24s/it, est. speed input: 34.80 toks/s, output: 111.08 toks/s]
Processed prompts: 100%| 1/1 [00:01<00:00, 1.84s/it, est. speed input: 37.59 toks/s, output: 110.06 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.25s/it, est. speed input: 39.08 toks/s, output: 111.02 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.25s/it, est. speed input: 25.83 toks/s, output: 111.32 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.26s/it, est. speed input: 36.36 toks/s, output: 110.85 toks/s]
Processed prompts: 100%| 1/1 [00:01<00:00, 1.72s/it, est. speed input: 34.83 toks/s, output: 109.13 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.25s/it, est. speed input: 22.22 toks/s, output: 111.08 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.25s/it, est. speed input: 26.65 toks/s, output: 111.04 toks/s]
Processed prompts: 100%| 1/1 [00:01<00:00, 1.83s/it, est. speed input: 38.90 toks/s, output: 110.13 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.25s/it, est. speed input: 27.95 toks/s, output: 110.91 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.26s/it, est. speed input: 36.28 toks/s, output: 110.60 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.25s/it, est. speed input: 27.95 toks/s, output: 110.91 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.26s/it, est. speed input: 40.72 toks/s, output: 110.65 toks/s]
Processed prompts: 100%| 1/1 [00:01<00:00, 1.05s/it, est. speed input: 60.11 toks/s, output: 104.95 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.26s/it, est. speed input: 65.05 toks/s, output: 110.63 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.06s/it, est. speed input: 49.02 toks/s, output: 110.65 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.26s/it, est. speed input: 35.47 toks/s, output: 110.85 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.26s/it, est. speed input: 42.52 toks/s, output: 110.72 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.25s/it, est. speed input: 50.18 toks/s, output: 111.01 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.25s/it, est. speed input: 50.18 toks/s, output: 111.02 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.25s/it, est. speed input: 30.65 toks/s, output: 111.05 toks/s]
Processed prompts: 100%| 1/1 [00:01<00:00, 1.59s/it, est. speed input: 35.28 toks/s, output: 108.98 toks/s]
Processed prompts: 100%| 1/1 [00:02<00:00, 2.26s/it, est. speed input: 27.49 toks/s, output: 110.84 toks/s]
Avg sentences : 26.84
95th-pct sentences : 35.55
Avg tokens : 230.5
95th-pct tokens: 250.0
[22]:
import time, re, string, collections
from datasets import load_dataset
from vllm import SamplingParams
from tqdm.auto import tqdm
# utility helpers
def first_int(t):
m = re.search(r"-?\d+", t)
return int(m.group()) if m else None
_tbl = str.maketrans("", "", string.punctuation)
norm = lambda t: t.lower().translate(_tbl).strip()
def correct(gold, pred):

193.69.10.14

Home | ToT (4) | Qwen2_5 (1) | [BUG] RuntimeError: Can't pickle local object 'patch_vl... | RuntimeError: Can't get local object 'patch_vllm_comp...

jupyter ToT (4) Last Checkpoint: 11 hours ago

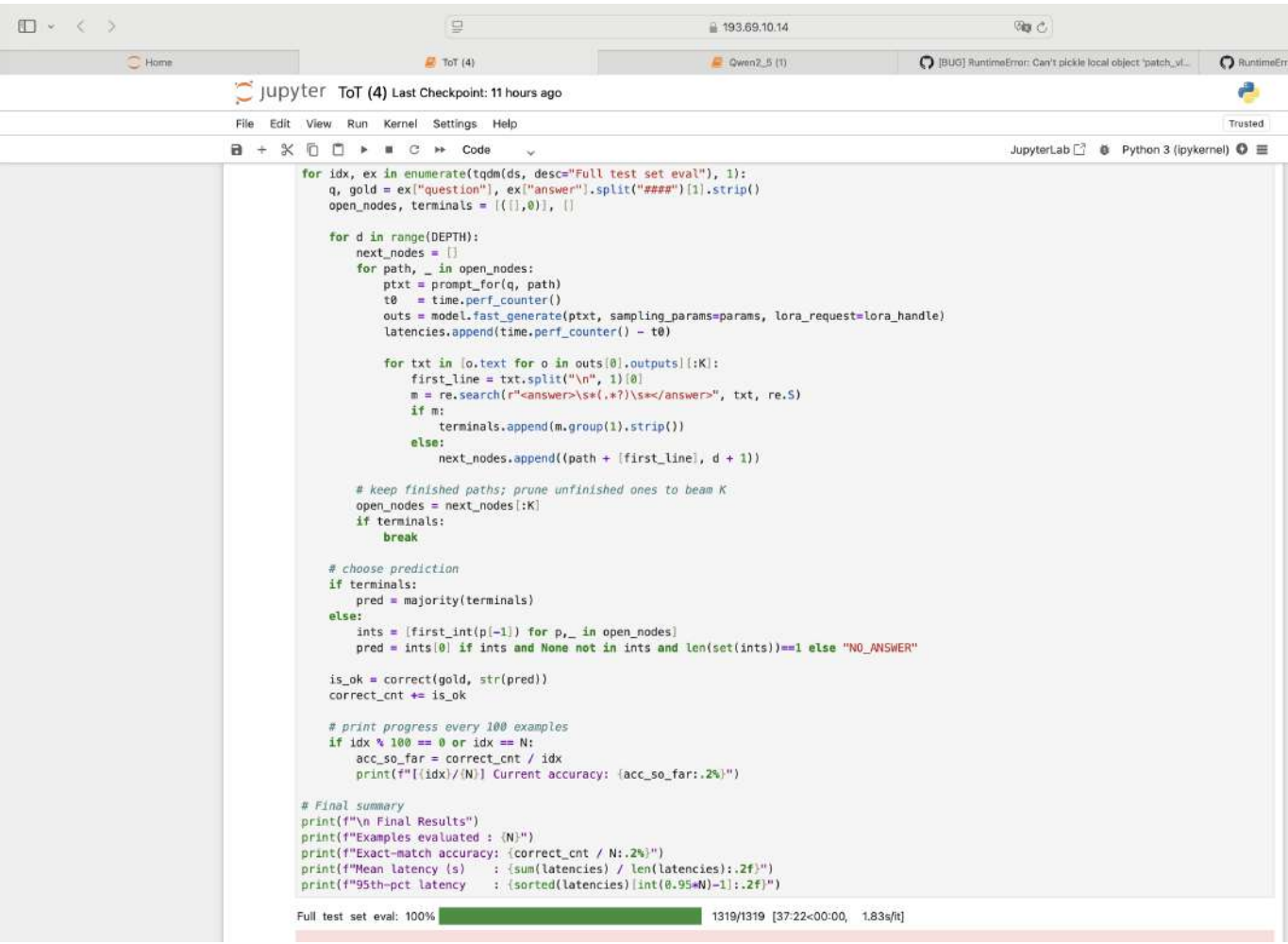
File Edit View Run Kernel Settings Help

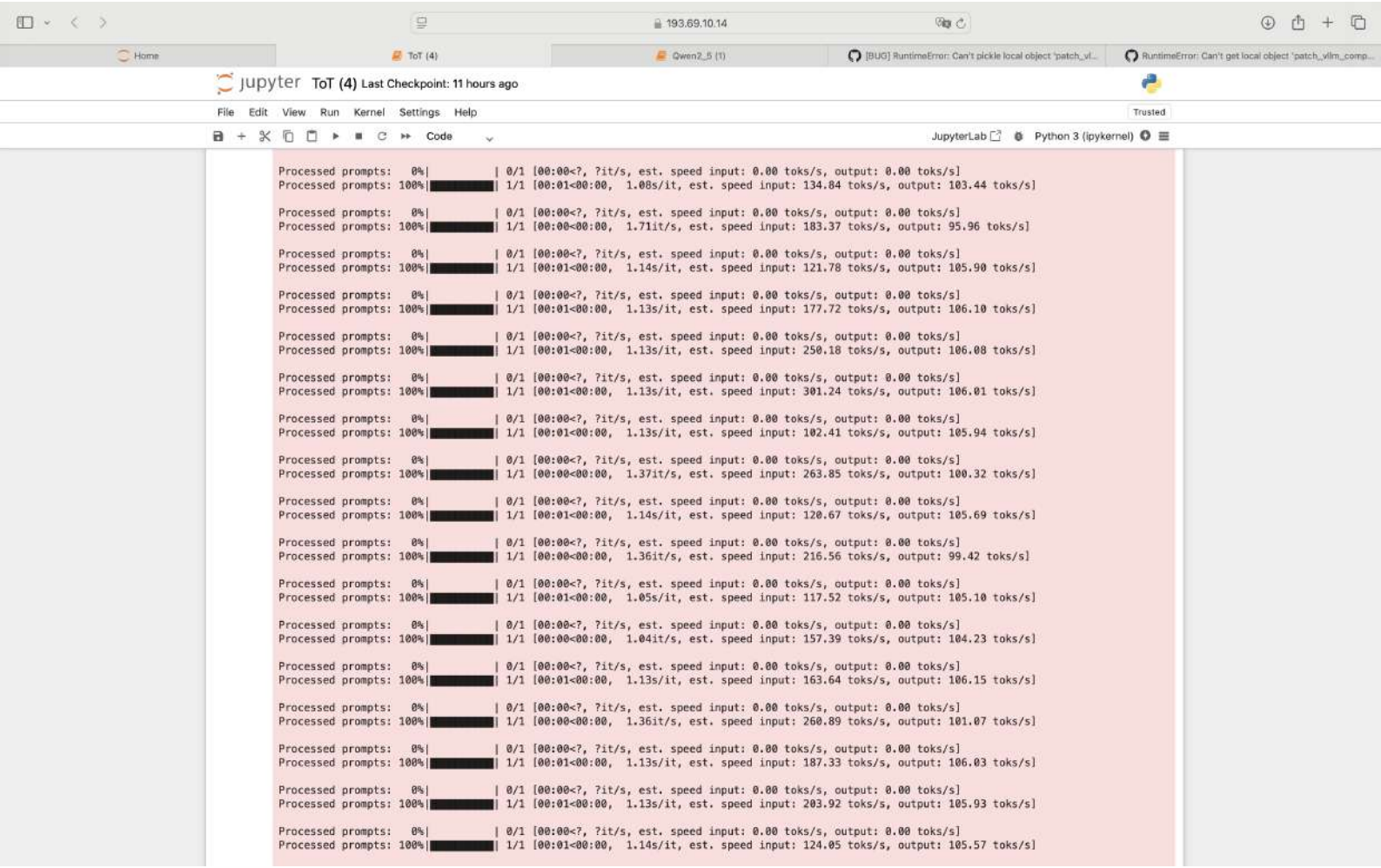
Trusted

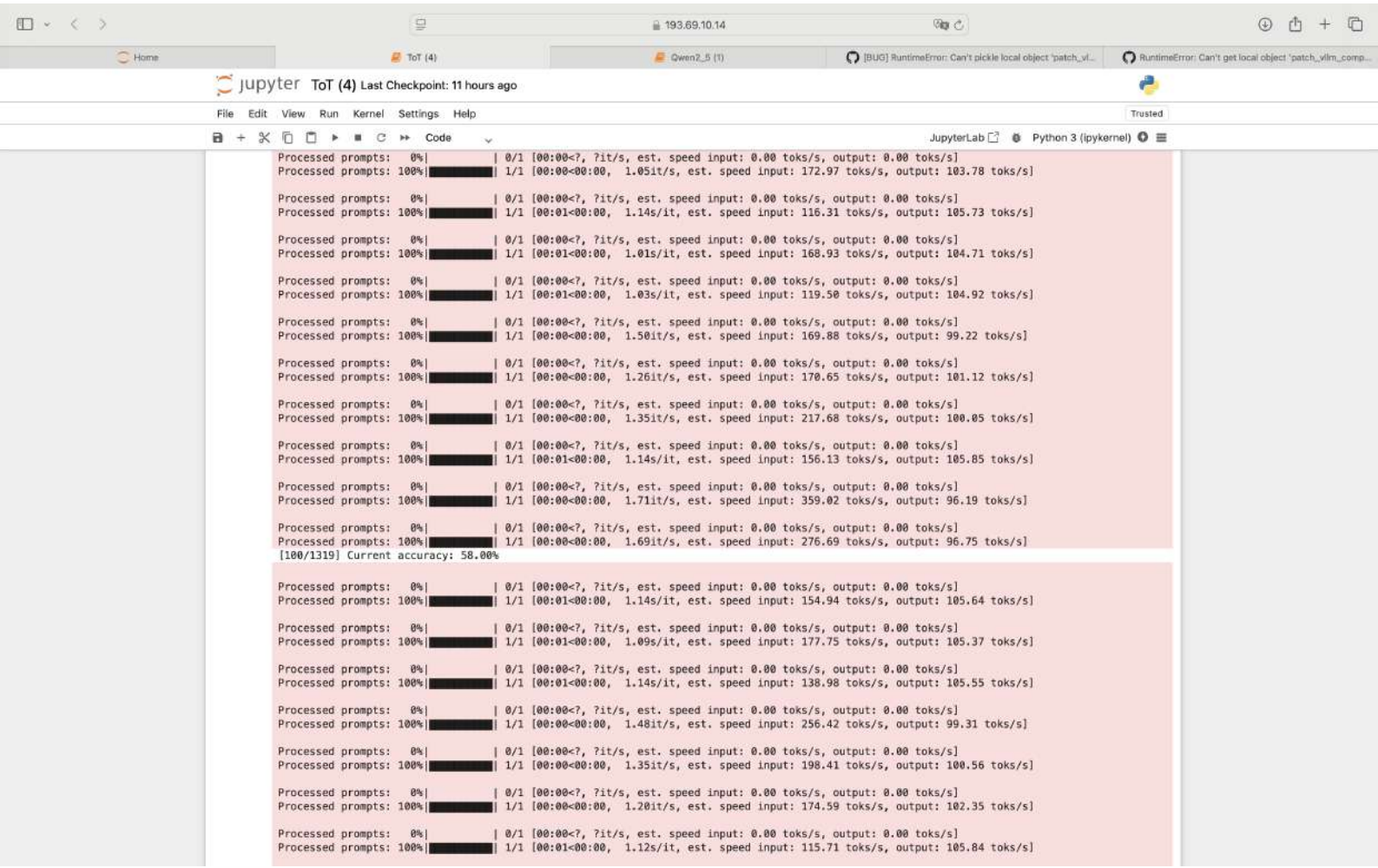
Code

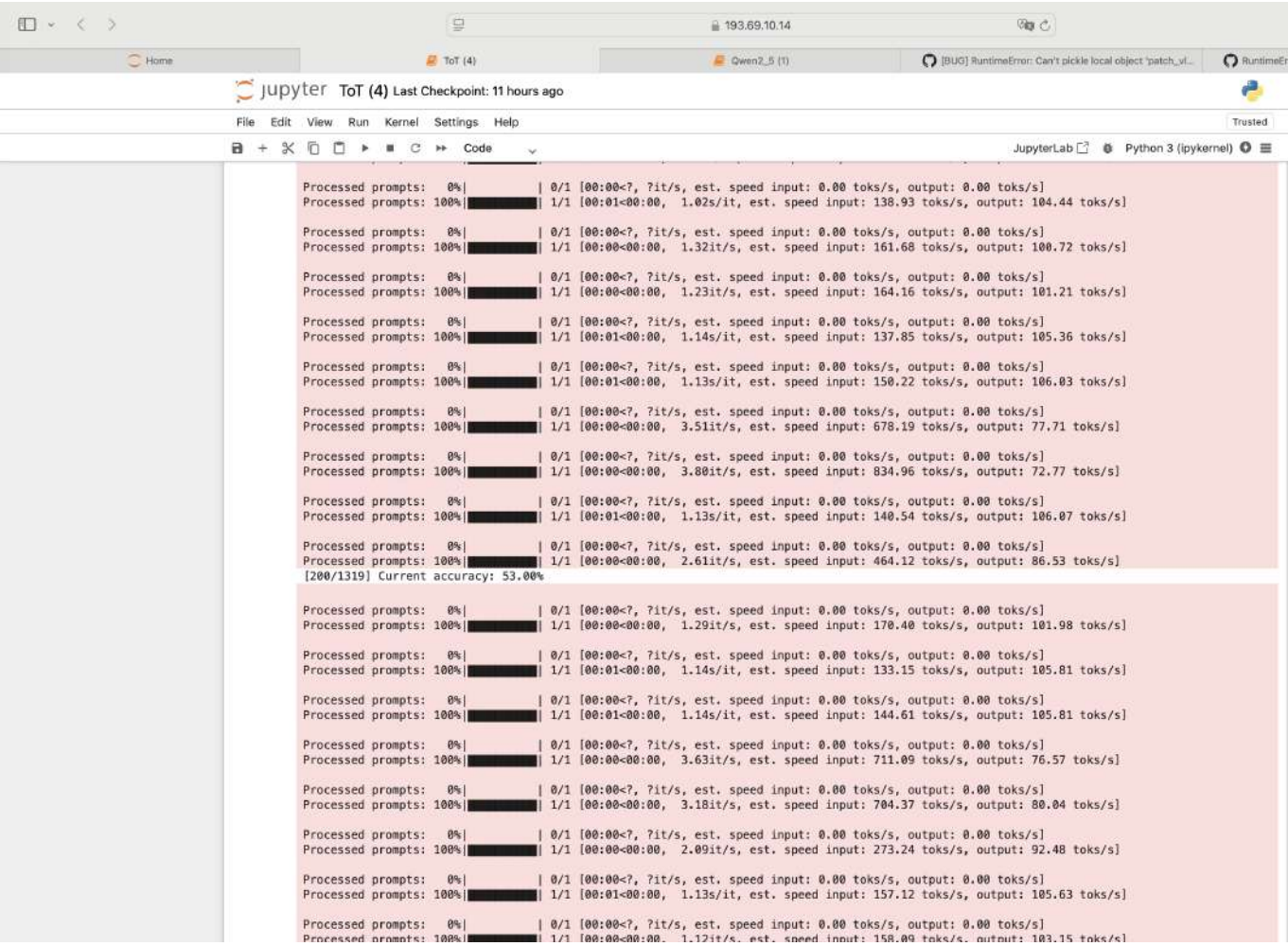
JupyterLab | Python 3 (ipykernel)

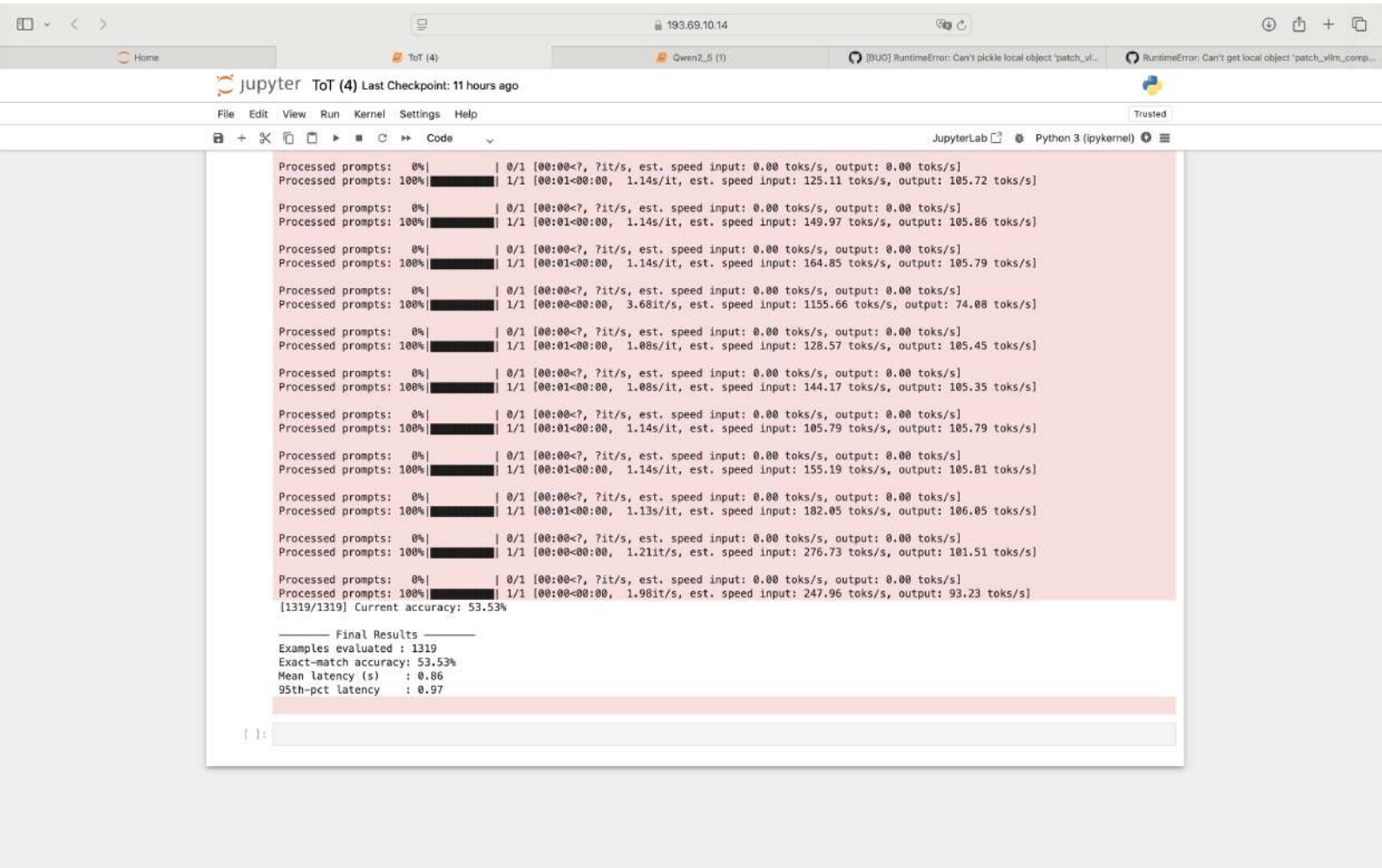
```python
def correct(gold, pred):
    g, p = first_int(gold), first_int(pred)
    return (g is not None and g == p) or norm(gold) in norm(pred)

def majority(ans):
    nums = [first_int(a) for a in ans]
    if all(n is not None for n in nums):
        return collections.Counter(nums).most_common(1)[0][0]
    return collections.Counter([norm(a) for a in ans]).most_common(1)[0][0]

# search parameters
DEPTH          = 4          # thought levels
K              = 3          # beam width
TOKENS         = 120        # tokens per thought
TEMP           = 0.3
SYS_PROMPT = (
    "You are a careful problem-solving assistant.\n"
    "For each question, think step-by-step - one line per thought.\n"
    "When you are ready, output one final line EXACTLY in this form:\n"
    "<answer>42</answer>\n"
    "…replacing 42 with the numeric answer. Do not write anything after </answer>."
)
TH_MARK = "Thought: "

def prompt_for(question, thoughts):
    """Assemble full text prompt ending with 'Thought: ' ready for generation."""
    msgs = [{"role": "system", "content": SYS_PROMPT},
            {"role": "user",   "content": question}]
    for t in thoughts:
        msgs.append({"role": "assistant", "content": t})
    text = tokenizer.apply_chat_template(
        msgs, tokenize=False, add_generation_prompt=True
    )
    return text + TH_MARK

params = SamplingParams(
    temperature = TEMP,
    top_p       = 1.0,
    max_tokens  = TOKENS,
)

# load LoRA once
lora_handle = globals().get("lora_handle") or model.load_lora("grpo_saved_lora")

# load the FULL test set
ds = load_dataset("openai/gsm8k", "main", split="test")
N = len(ds)

correct_cnt, latencies = 0, []
```


193.69.10.14
Home
ToT (4)
Qwen2_5 (1)
[BUG] RuntimeError: Can't pickle local object 'patch_vl...
RuntimeError: Can't get local object 'patch_vllm_comp...
jupyter ToT (4) Last Checkpoint: 11 hours ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
for idx, ex in enumerate(tqdm(ds, desc="Full test set eval"), 1):
q, gold = ex["question"], ex["answer"].split("####")[1].strip()
open_nodes, terminals = [([],0)], []
for d in range(DEPTH):
next_nodes = []
for path, _ in open_nodes:
ptxt = prompt_for(q, path)
t0 = time.perf_counter()
outs = model.fast_generate(ptxt, sampling_params=params, lora_request=lora_handle)
latencies.append(time.perf_counter() - t0)
for txt in [o.text for o in outs[0].outputs][:K]:
first_line = txt.split("\n", 1)[0]
m = re.search(r"<answer>\s*(.*?)\s*</answer>", txt, re.S)
if m:
terminals.append(m.group(1).strip())
else:
next_nodes.append((path + [first_line], d + 1))
# keep finished paths; prune unfinished ones to beam K
open_nodes = next_nodes[:K]
if terminals:
break
# choose prediction
if terminals:
pred = majority(terminals)
else:
ints = [first_int(p[-1]) for p,_ in open_nodes]
pred = ints[0] if ints and None not in ints and len(set(ints))==1 else "NO_ANSWER"
is_ok = correct(gold, str(pred))
correct_cnt += is_ok
# print progress every 100 examples
if idx % 100 == 0 or idx == N:
acc_so_far = correct_cnt / idx
print(f"[{idx}/{N}] Current accuracy: {acc_so_far:.2%}")
# Final summary
print(f"\n Final Results")
print(f"Examples evaluated : {N}")
print(f"Exact-match accuracy: {correct_cnt / N:.2%}")
print(f"Mean latency (s) : {sum(latencies) / len(latencies):.2f}")
print(f"95th-pct latency : {sorted(latencies)[int(0.95*N)-1]:.2f}")
Full test set eval: 100%
1319/1319 [37:22<00:00, 1.83s/it]


193.69.10.14
Home
ToT (4)
Qwen2_5 (1)
[BUG] RuntimeError: Can't pickle local object 'patch_vl...
RuntimeError: Can't get local object 'patch_vllm_comp...
jupyter ToT (4) Last Checkpoint: 11 hours ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.08s/it, est. speed input: 134.84 toks/s, output: 103.44 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.71it/s, est. speed input: 183.37 toks/s, output: 95.96 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.14s/it, est. speed input: 121.78 toks/s, output: 105.90 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.13s/it, est. speed input: 177.72 toks/s, output: 106.10 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.13s/it, est. speed input: 250.18 toks/s, output: 106.08 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.13s/it, est. speed input: 301.24 toks/s, output: 106.01 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.13s/it, est. speed input: 102.41 toks/s, output: 105.94 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.37it/s, est. speed input: 263.85 toks/s, output: 100.32 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.14s/it, est. speed input: 120.67 toks/s, output: 105.69 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.36it/s, est. speed input: 216.56 toks/s, output: 99.42 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.05s/it, est. speed input: 117.52 toks/s, output: 105.10 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.04it/s, est. speed input: 157.39 toks/s, output: 104.23 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.13s/it, est. speed input: 163.64 toks/s, output: 106.15 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.36it/s, est. speed input: 260.89 toks/s, output: 101.07 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.13s/it, est. speed input: 187.33 toks/s, output: 106.03 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.13s/it, est. speed input: 203.92 toks/s, output: 105.93 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.14s/it, est. speed input: 124.05 toks/s, output: 105.57 toks/s]


193.69.10.14
Home
ToT (4)
Qwen2_5 (1)
[BUG] RuntimeError: Can't pickle local object 'patch_vl...
RuntimeError: Can't get local object 'patch_vllm_comp...
jupyter ToT (4) Last Checkpoint: 11 hours ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.05it/s, est. speed input: 172.97 toks/s, output: 103.78 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.14s/it, est. speed input: 116.31 toks/s, output: 105.73 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.01s/it, est. speed input: 168.93 toks/s, output: 104.71 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.03s/it, est. speed input: 119.50 toks/s, output: 104.92 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.50it/s, est. speed input: 169.88 toks/s, output: 99.22 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.26it/s, est. speed input: 170.65 toks/s, output: 101.12 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.35it/s, est. speed input: 217.68 toks/s, output: 100.05 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.14s/it, est. speed input: 156.13 toks/s, output: 105.85 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.71it/s, est. speed input: 359.02 toks/s, output: 96.19 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.69it/s, est. speed input: 276.69 toks/s, output: 96.75 toks/s]
[100/1319] Current accuracy: 58.00%
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.14s/it, est. speed input: 154.94 toks/s, output: 105.64 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.09s/it, est. speed input: 177.75 toks/s, output: 105.37 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.14s/it, est. speed input: 138.98 toks/s, output: 105.55 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.48it/s, est. speed input: 256.42 toks/s, output: 99.31 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.35it/s, est. speed input: 198.41 toks/s, output: 100.56 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.20it/s, est. speed input: 174.59 toks/s, output: 102.35 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.12s/it, est. speed input: 115.71 toks/s, output: 105.84 toks/s]


193.69.10.14
Home
ToT (4)
Qwen2_5 (1)
[BUG] RuntimeError: Can't pickle local object 'patch_vl...
RuntimeError: Can't get local object 'patch_vllm_comp...
jupyter ToT (4) Last Checkpoint: 11 hours ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.02s/it, est. speed input: 138.93 toks/s, output: 104.44 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.32it/s, est. speed input: 161.68 toks/s, output: 100.72 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.23it/s, est. speed input: 164.16 toks/s, output: 101.21 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.14s/it, est. speed input: 137.85 toks/s, output: 105.36 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.13s/it, est. speed input: 150.22 toks/s, output: 106.03 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 3.51it/s, est. speed input: 678.19 toks/s, output: 77.71 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 3.80it/s, est. speed input: 834.96 toks/s, output: 72.77 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.13s/it, est. speed input: 140.54 toks/s, output: 106.07 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 2.61it/s, est. speed input: 464.12 toks/s, output: 86.53 toks/s]
[200/1319] Current accuracy: 53.00%
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 1.29it/s, est. speed input: 170.40 toks/s, output: 101.98 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.14s/it, est. speed input: 133.15 toks/s, output: 105.81 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.14s/it, est. speed input: 144.61 toks/s, output: 105.81 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 3.63it/s, est. speed input: 711.09 toks/s, output: 76.57 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 3.18it/s, est. speed input: 704.37 toks/s, output: 80.04 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:00<00:00, 2.09it/s, est. speed input: 273.24 toks/s, output: 92.48 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%| | 1/1 [00:01<00:00, 1.13s/it, est. speed input: 157.12 toks/s, output: 105.63 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]


193.69.10.14
Home
ToT (4)
Qwen2_5 (1)
[BUG] RuntimeError: Can't pickle local object 'patch_vl...
RuntimeError: Can't get local object 'patch_vllm_comp...
jupyter ToT (4) Last Checkpoint: 11 hours ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%|| 1/1 [00:01<00:00, 1.14s/it, est. speed input: 125.11 toks/s, output: 105.72 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%|| 1/1 [00:01<00:00, 1.14s/it, est. speed input: 149.97 toks/s, output: 105.86 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%|| 1/1 [00:01<00:00, 1.14s/it, est. speed input: 164.85 toks/s, output: 105.79 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%|| 1/1 [00:00<00:00, 3.68it/s, est. speed input: 1155.66 toks/s, output: 74.08 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%|| 1/1 [00:01<00:00, 1.08s/it, est. speed input: 128.57 toks/s, output: 105.45 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%|| 1/1 [00:01<00:00, 1.08s/it, est. speed input: 144.17 toks/s, output: 105.35 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%|| 1/1 [00:01<00:00, 1.14s/it, est. speed input: 105.79 toks/s, output: 105.79 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%|| 1/1 [00:01<00:00, 1.14s/it, est. speed input: 155.19 toks/s, output: 105.81 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%|| 1/1 [00:01<00:00, 1.13s/it, est. speed input: 182.05 toks/s, output: 106.05 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%|| 1/1 [00:00<00:00, 1.21it/s, est. speed input: 276.73 toks/s, output: 101.51 toks/s]
Processed prompts: 0%| | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Processed prompts: 100%|| 1/1 [00:00<00:00, 1.98it/s, est. speed input: 247.96 toks/s, output: 93.23 toks/s]
[1319/1319] Current accuracy: 53.53%
———— Final Results ————
Examples evaluated : 1319
Exact-match accuracy: 53.53%
Mean latency (s) : 0.86
95th-pct latency : 0.97
[ ]: