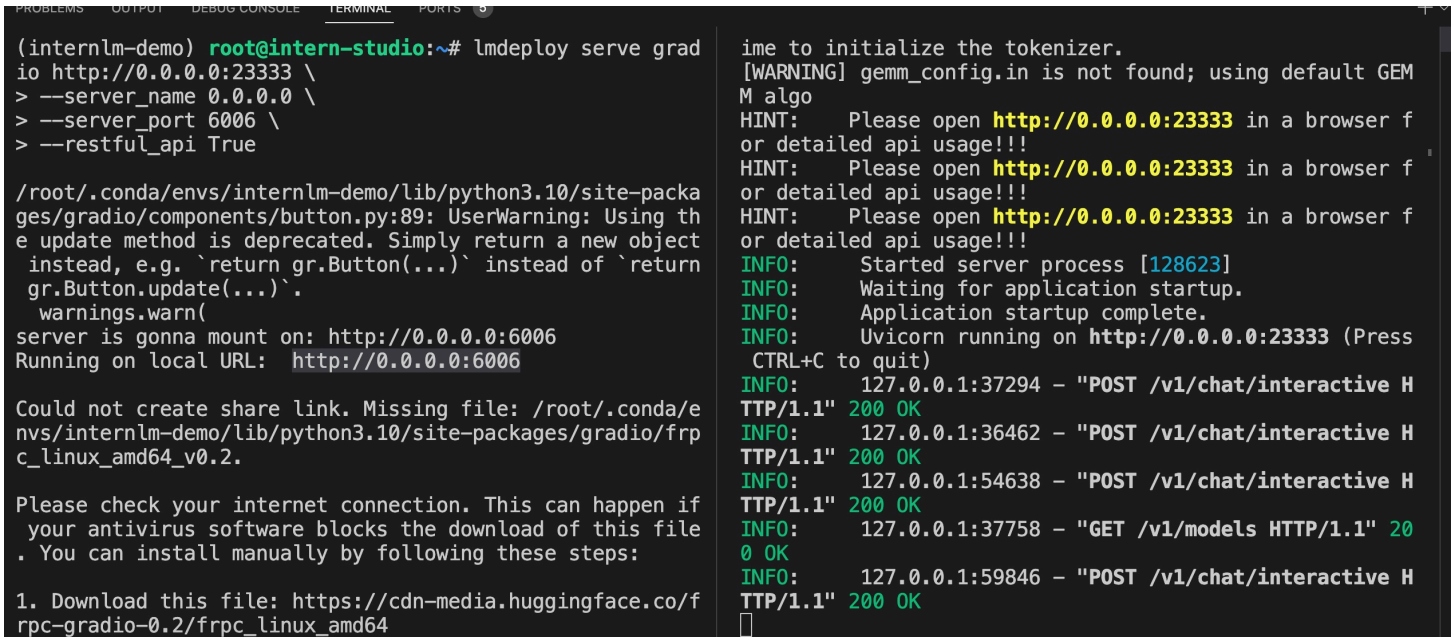
基础作业：

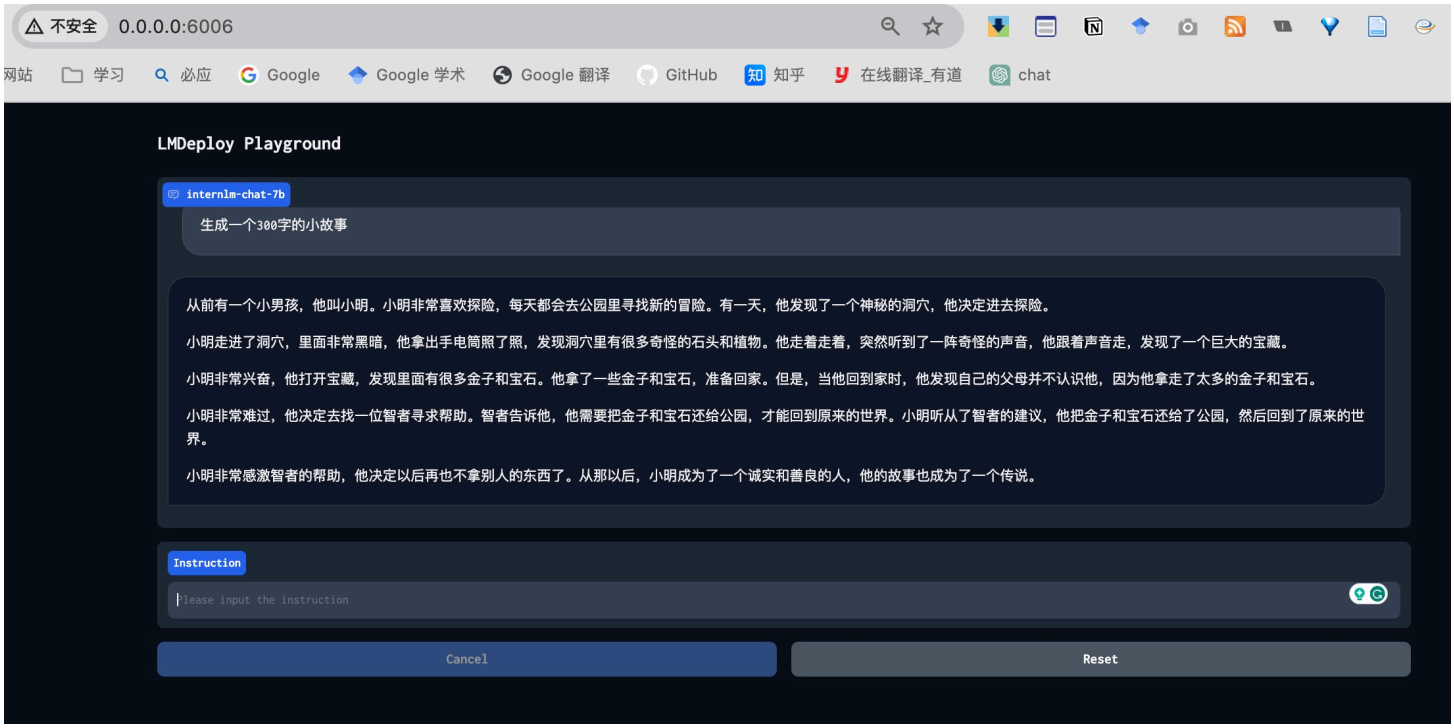
使用 LMDeploy 以本地对话、网页Gradio、API服务中的一种方式部署 InternLM-Chat-7B 模型，生成 300 字的小故事（需截图）

本地对话 (TurboMind推理+API服务)

TurboMind 服务作为后端 + gradio

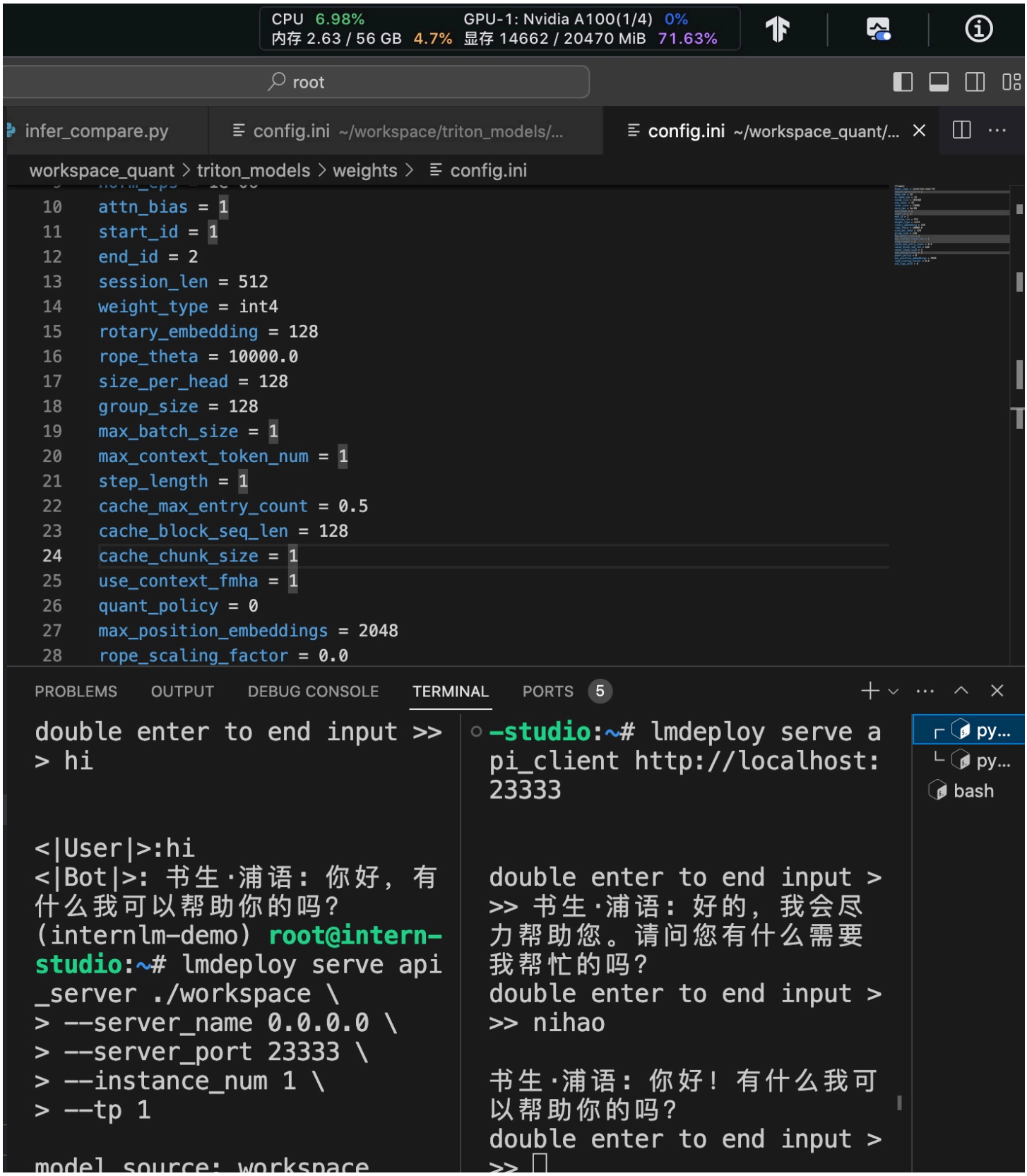


TurboMind 推理作为后端 + gradio

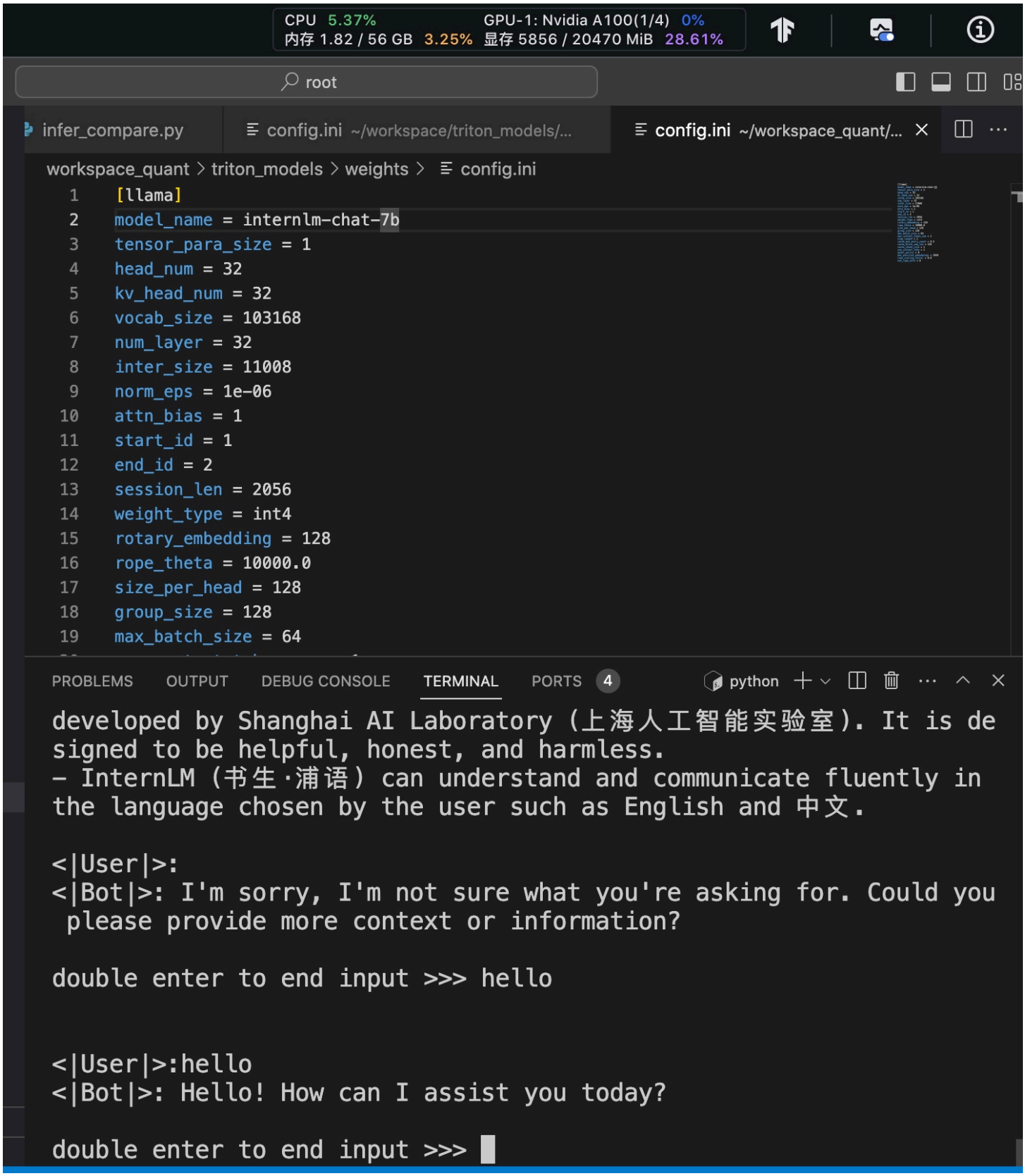


进阶作业（可选做）

无量化 显存14G


CPU 6.98%
内存 2.63 / 56 GB 4.7%
GPU-1: Nvidia A100(1/4) 0%
显存 14662 / 20470 MiB 71.63%
root
infer_compare.py
config.ini ~/workspace/triton_models/...
config.ini ~/workspace_quant/...
workspace_quant > triton_models > weights > config.ini
attn_bias = 1
start_id = 1
end_id = 2
session_len = 512
weight_type = int4
rotary_embedding = 128
rope_theta = 10000.0
size_per_head = 128
group_size = 128
max_batch_size = 1
max_context_token_num = 1
step_length = 1
cache_max_entry_count = 0.5
cache_block_seq_len = 128
cache_chunk_size = 1
use_context_fmha = 1
quant_policy = 0
max_position_embeddings = 2048
rope_scaling_factor = 0.0
PROBLEMS
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS 5
double enter to end input >>
> hi
<|User|>:hi
<|Bot|>: 书生·浦语：你好，有什么我可以帮助你的吗？
(internlm-demo) root@intern-studio:~# lmdeploy serve api_server ./workspace \
> --server_name 0.0.0.0 \
> --server_port 23333 \
> --instance_num 1 \
> --tp 1
model source: workspace
-studio:~# lmdeploy serve api_client http://localhost:23333
double enter to end input >>
>> 书生·浦语：好的，我会尽力帮助您。请问您有什么需要我帮忙的吗？
double enter to end input >>
>> nihao
书生·浦语：你好！有什么我可以帮助你的吗？
double enter to end input >>
py...
py...
bash

W4A16量化


CPU 5.37%
内存 1.82 / 56 GB 3.25%
GPU-1: Nvidia A100(1/4) 0%
显存 5856 / 20470 MiB 28.61%
root
infer_compare.py
config.ini ~/workspace/triton_models/...
config.ini ~/workspace_quant/...
workspace_quant > triton_models > weights > config.ini
[llama]
model_name = internlm-chat-7b
tensor_para_size = 1
head_num = 32
kv_head_num = 32
vocab_size = 103168
num_layer = 32
inter_size = 11008
norm_eps = 1e-06
attn_bias = 1
start_id = 1
end_id = 2
session_len = 2056
weight_type = int4
rotary_embedding = 128
rope_theta = 10000.0
size_per_head = 128
group_size = 128
max_batch_size = 64
PROBLEMS
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS 4
python
developed by Shanghai AI Laboratory (上海人工智能实验室). It is de
signed to be helpful, honest, and harmless.
- InternLM (书生·浦语) can understand and communicate fluently in
the language chosen by the user such as English and 中文.
<|User|>:
<|Bot|>: I'm sorry, I'm not sure what you're asking for. Could you
please provide more context or information?
double enter to end input >>> hello
<|User|>:hello
<|Bot|>: Hello! How can I assist you today?
double enter to end input >>>