## 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)

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS 4
                                                                                                                                                  ┌ 🕡 pyth
  了一个英雄。他也意识到,冒险和探险不仅仅是为了获得财富
更是为了寻找真正的自己。
                                                                         > --server_port 23333 \
                                                                                                                                                  └ 🕟 pyth
                                                                         > --instance_num 64 \
                                                                         > ---tp 1
从此以后,小明继续他的冒险之旅,他不断地探索新的世界,发现新的宝藏,也发现了更多的自己。
double enter to end input >>> 重新讲一个300字的小故事
                                                                         model_source: workspace
                                                                         WARNING: Can not find tokenizer.json. It may take long t
                                                                         ime to initialize the tokenizer.
[WARNING] gemm_config.in is not found; using default GEM
从前,有一个小女孩叫做小红。她非常喜欢吃糖果,每天都会吃很多糖果。但是,她总是感到不舒服,因为她吃太多的糖果,导致她的牙齿蛀牙了。
                                                                         M algo
                                                                         HINT: Please open http://0.0.0.0:23333 in a browser f or detailed api usage!!!
HINT: Please open http://0.0.0:23333 in a browser f or detailed api usage!!!
小红的妈妈带她去看牙医,牙医告诉她,如果她再吃太多的糖果,她的牙齿就会烂掉。小红听了很害怕,她决定要改变自己的饮
,她的,
食习惯。
                                                                         HINT: Please open <a href="http://0.0.0.0:23333">http://0.0.0:23333</a> in a browser for detailed api usage!!!
小红开始减少吃糖果的频率,每天只吃一两个糖果。她开始吃更
多的蔬菜和水果,这些食物对牙齿健康有益。
                                                                         INFO:
                                                                                     Started server process [128623]
                                                                          INFO:
                                                                                     Waiting for application startup.
                                                                          INFO:
                                                                                      Application startup complete.
小红的牙齿开始变得健康了,她的笑容也变得更加灿烂。她意识到,健康的生活方式比吃糖果更重要。
                                                                                     Uvicorn running on http://0.0.0.0:23333 (Press
                                                                          INFO:
                                                                          CTRL+C to quit)
                                                                          INFO:
                                                                                      127.0.0.1:37294 - "POST /v1/chat/interactive H
小红继续保持健康的生活方式,她的牙齿一直保持健康,她也变得更加自信和快乐。她知道,只要坚持健康的生活方式,她就可以拥有健康和幸福的生活。
                                                                          TTP/1.1"
                                                                                      127.0.0.1:36462 - "POST /v1/chat/interactive H
                                                                          TTP/1.1" 200 OK
double enter to end input >>>
```

#### TurboMind 服务作为后端 + gradio

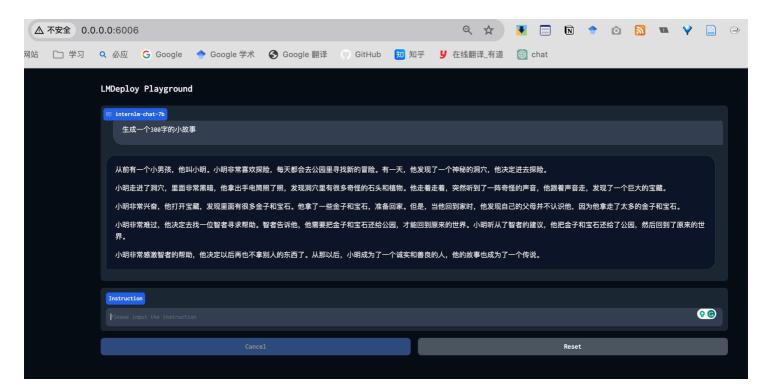
```
ime to initialize the tokenizer.
[WARNING] gemm_config.in is not found; using default GEM
(internlm-demo) root@intern-studio:~# lmdeploy serve grad
io http://0.0.0.0:23333 \
> --server_name 0.0.0.0 \
> --server_port 6006 \
> --restful_api True
                                                                      HINT:
                                                                                 Please open http://0.0.0.0:23333 in a browser f
                                                                      or detailed api usage!!!
                                                                      HTNT:
                                                                                Please open http://0.0.0.0:23333 in a browser f
/root/.conda/envs/internlm-demo/lib/python3.10/site-packa
                                                                      or detailed api usage!!!
ges/gradio/components/button.py:89: UserWarning: Using the update method is deprecated. Simply return a new object instead, e.g. `return gr.Button(...)` instead of `return gr.Button(...)`
                                                                      HINT:
                                                                                Please open http://0.0.0.0:23333 in a browser f
                                                                      or detailed api usage!!!
                                                                                  Started server process [128623] Waiting for application startup.
                                                                      INFO:
 gr.Button.update(...)`;
  warnings.warn(
                                                                      INFO:
                                                                                  Application startup complete.
server is gonna mount on: http://0.0.0.0:6006
                                                                      INFO:
                                                                                  Uvicorn running on http://0.0.0.0:23333 (Press
Running on local URL: http://0.0.0.0:6006
                                                                       CTRL+C to quit)
                                                                                  127.0.0.1:37294 - "POST /v1/chat/interactive H
                                                                      INFO:
                                                                      TTP/1.1"
Could not create share link. Missing file: /root/.conda/e
                                                                                200 0
                                                                                  127.0.0.1:36462 - "POST /v1/chat/interactive H
nvs/internlm-demo/lib/python3.10/site-packages/gradio/frp
                                                                      INFO:
c_linux_amd64_v0.2.
                                                                      TTP/1.1"
                                                                      INFO:
                                                                                  127.0.0.1:54638 - "POST /v1/chat/interactive H
Please check your internet connection. This can happen if
                                                                      TTP/1.1"
                                                                                  127.0.0.1:37758 - "GET /v1/models HTTP/1.1" 20
 your antivirus software blocks the download of this file
                                                                      INFO:
  You can install manually by following these steps:
                                                                      INFO:
                                                                                  127.0.0.1:59846 - "POST /v1/chat/interactive H
                                                                      TTP/1.1" 200 OK

    Download this file: https://cdn-media.huggingface.co/f

rpc-gradio-0.2/frpc linux amd64
```



### TurboMind 推理作为后端 + gradio



# 进阶作业 (可选做)

无量化 显存14G

