

第五次课笔记（作业截图附后）

2024年1月27日 23:08

环境配置:

```
• (lmdeploy) root@intern-studio:~# python
Python 3.10.13 (main, Sep 11 2023, 13:44:35) [GCC 11
.2.0] on linux
Type "help", "copyright", "credits" or "license" for
more information.
>>> import lmdeploy
>>> exit()
File "<stdin>", line 1
    exit()
```

本地一键运行:

```
>>> exit()
• (lmdeploy) root@intern-studio:~# lmdeploy chat turbo
mind /share/temp/model_repos/internlm-chat-7b/ --mo
del-name internlm-chat-7b
```

运行结果:

```
<|User|>:please introduce yourself
<|Bot|>: Hello! I am InternLM, a conversational language model developed by Sh
anghai AI Laboratory. I am designed to be helpful, honest, and harmless. How m
ay I assist you today?

double enter to end input >>> 你是谁

<|User|>:你是谁
<|Bot|>: 你好！我是书生·浦语，上海人工智能实验室开发的人机交互语言模型。我致力
于通过执行常见的基于语言的任务和提供建议来帮助人类。请问有什么我可以帮助你的吗
？

double enter to end input >>> 
```

Turbo推理+命令行本地对话

```
• (lmdeploy) root@intern-studio:~# # Turbomind + Bash Local Chat
• (lmdeploy) root@intern-studio:~# lmdeploy chat turbomind ./workspace
```

```

◦ (lmdeploy) root@intern-studio:~# # Turbomind + Bash
Local Chat
◦ (lmdeploy) root@intern-studio:~# lmdeploy chat turbo
mind ./workspace
model_source: workspace
WARNING: Can not find tokenizer.json. It may take lo
ng time to initialize the tokenizer.
[WARNING] gemm_config.in is not found; using default
GEMM algo
session 1

double enter to end input >>> introduce you

<|System|>:You are an AI assistant whose name is Int
ernLM (书生·浦语).
- InternLM (书生·浦语) is a conversational language
model that is developed by Shanghai AI Laboratory (
上海人工智能实验室). It is designed to be helpful, h
onest, and harmless.
- InternLM (书生·浦语) can understand and communicat
e fluently in the language chosen by the user such a
s English and 中文.

<|User|>:introduce you
<|Bot|>: Hello! My name is InternLM (书生·浦语), and
I am a conversational language model developed by S
hanghai AI Laboratory. I am designed to be helpful,
honest, and harmless. How may I assist you today?

double enter to end input >>> █

```



Responses

Curl

```
curl -X 'GET' \
  'http://localhost:23333/v1/models' \
  -H 'accept: application/json'
```

Request URL

```
http://localhost:23333/v1/models
```

Server response

Code Details

200

Response body

```
{
  "object": "list",
  "data": [
    {
      "id": "internlm-chat-7b",
      "object": "model",
      "created": 1706417834,
      "owned_by": "lmdeploy",
      "root": "internlm-chat-7b",
      "parent": null,
      "permission": [
        {
          "id": "modelperm-3Dabx35ak9W9Mh5qeuKzBt",
          "object": "model_permission",
          "created": 1706417834,
          "allow_create_engine": false,
          "allow_sampling": true,
          "allow_logprobs": true,
          "allow_search_indices": true,
          "allow_view": true,
          "allow_fine_tuning": false,
          "organization": "*",
          "group": null,
          "is_blocking": false
        }
      ]
    }
  ]
}
```

Response headers

```
content-length: 454
content-type: application/json
date: Sun, 28 Jan 2024 04:57:14 GMT
server: uvicorn
```

Responses

Code	Description	Links
200	Successful Response	No links

The request should be a JSON object with the following fields:

- model: model name. Available from /v1/models.
- messages: string prompt or chat history in OpenAI format.
- temperature (float): to modulate the next token probability
- top_p (float): If set to float < 1, only the smallest set of most probable tokens with probabilities that add up to top_p or higher are kept for generation.
- n (int): How many chat completion choices to generate for each input message. Only support one here.
- stream: whether to stream the results or not. Default to false.
- max_tokens (int): output token nums
- repetition_penalty (float): The parameter for repetition penalty. 1.0 means no penalty

Additional arguments supported by LMDeploy:

- ignore_eos (bool): indicator for ignoring eos
- session_id (int): if not specified, will set random value

Currently we do not support the following features:

- function_call (Users should implement this by themselves)
- logit_bias (not supported yet)
- presence_penalty (replaced with repetition_penalty)
- frequency_penalty (replaced with repetition_penalty)

Parameters

Cancel

Reset

No parameters

Request body required

application/json

```
{
  "model": "internlm-chat-7b",
  "messages": "please introduce yourself",
  "temperature": 0.7,
  "top_p": 1,
  "n": 1,
  "max_tokens": 512,
  "stop": false,
  "stream": false,
  "presence_penalty": 0,
  "frequency_penalty": 0,
  "user": "string",
  "repetition_penalty": 1,
  "session_id": -1,
  "ignore_eos": false
}
```

Execute

LMDeploy Playground

internlm-chat-7b

hello

Hello! How can I assist you today?

Instruction

Please input the instruction

Cancel

Reset

Use via API  · 使用Gradio构建 

```
(lmdeploy) root@intern-s
o tudio:~# # Gradio+ApiSer
Client 时 Gradio 为
(lmdeploy) root@intern-s
o tudio:~# lmdeploy serve
gradio http://0.0.0.0:23
333 \
> --server_name 0.0.0.0
\
> --server_port 6006 \
> --restful_api True
█
```

LMDeploy Playground

internlm-chat-7b

nihao

你好，有什么我可以帮助你的吗？

Instruction

Please input the instruction

Cancel

Reset

Use via API  · 使用Gradio构建 

作业：


```
double enter to end input >>> 生成一个300字的小故事

<|System|>:You are an AI assistant whose name is InternLM (书生·浦语).
- InternLM (书生·浦语) is a conversational language model that is developed by
  Shanghai AI Laboratory (上海人工智能实验室). It is designed to be helpful, ho
  nest, 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|>:生成一个300字的小故事
<|Bot|>: 从前，有一个小女孩叫做艾米。她非常喜欢探险，总是在寻找新的冒险。有一
  天，她听说了一个传说，说在森林深处有一个神秘的宝藏。艾米决定要去寻找这个宝藏。

  她穿过了茂密的森林，爬过了陡峭的山峰，终于来到了宝藏的所在地。但是，她发现宝藏
  被一只凶猛的野兽守护着。艾米不想放弃，于是她想出了一个计划。

  她找到了一些绳子，绑在了野兽的脖子上，然后放开了绳子。绳子被野兽拉扯着，让野兽
  无法继续守护宝藏。艾米成功地拿到了宝藏，并带着它回到了家。

  艾米的朋友们都非常惊讶，问她怎么能够打败野兽。艾米告诉他们，她并不是为了打败野
  兽，而是为了保护宝藏。她认为，每个人都应该保护自己的宝藏，无论是财富、知识还是
  友谊。

  从此以后，艾米成为了一个受人尊敬的探险家，她的勇气和智慧也成为了她的标志。

double enter to end input
>>> █
```

API& webdemo部分忘记截图了（