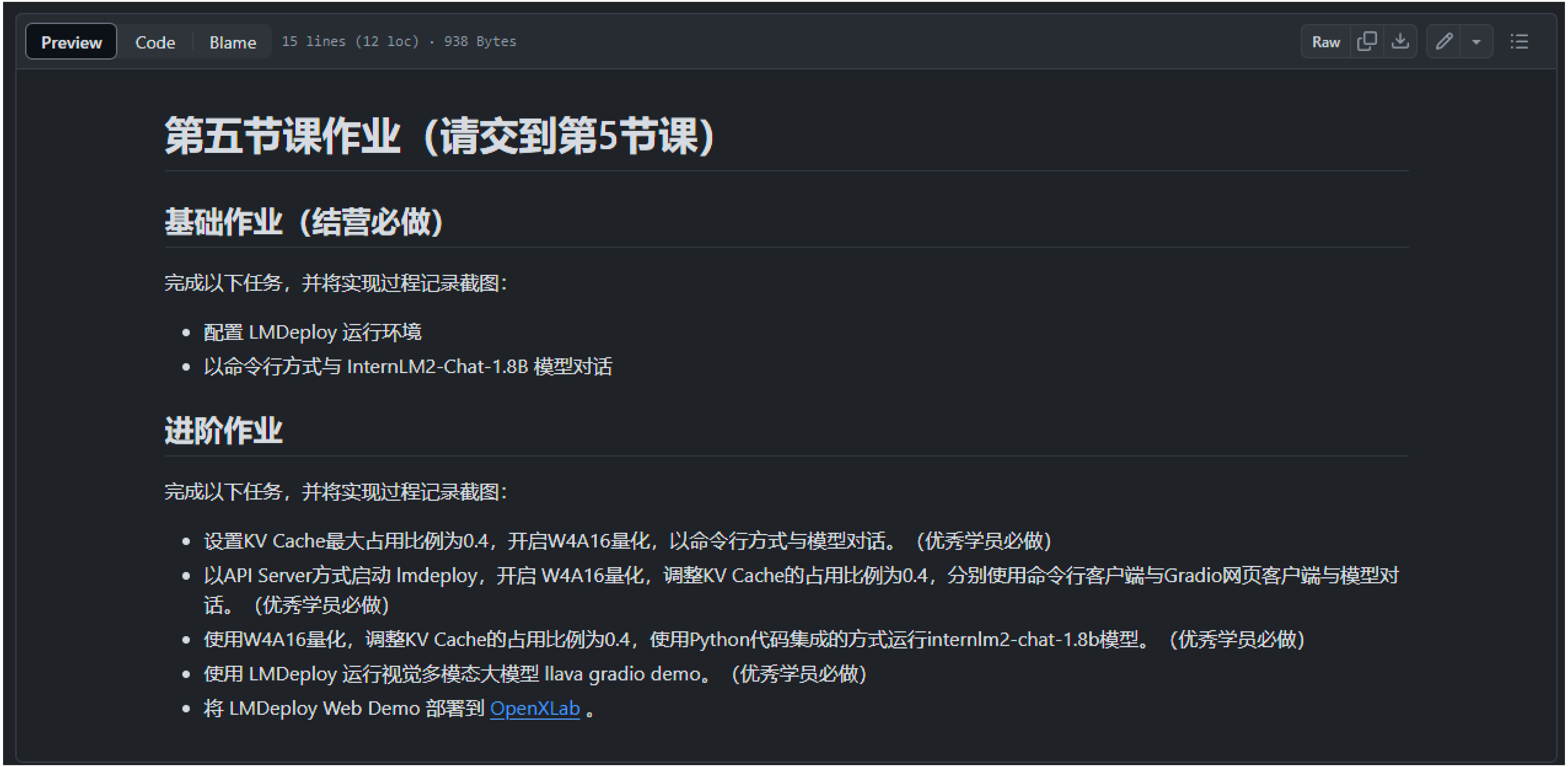


第五节课作业（请交到第5节课）

基础作业（结营必做）

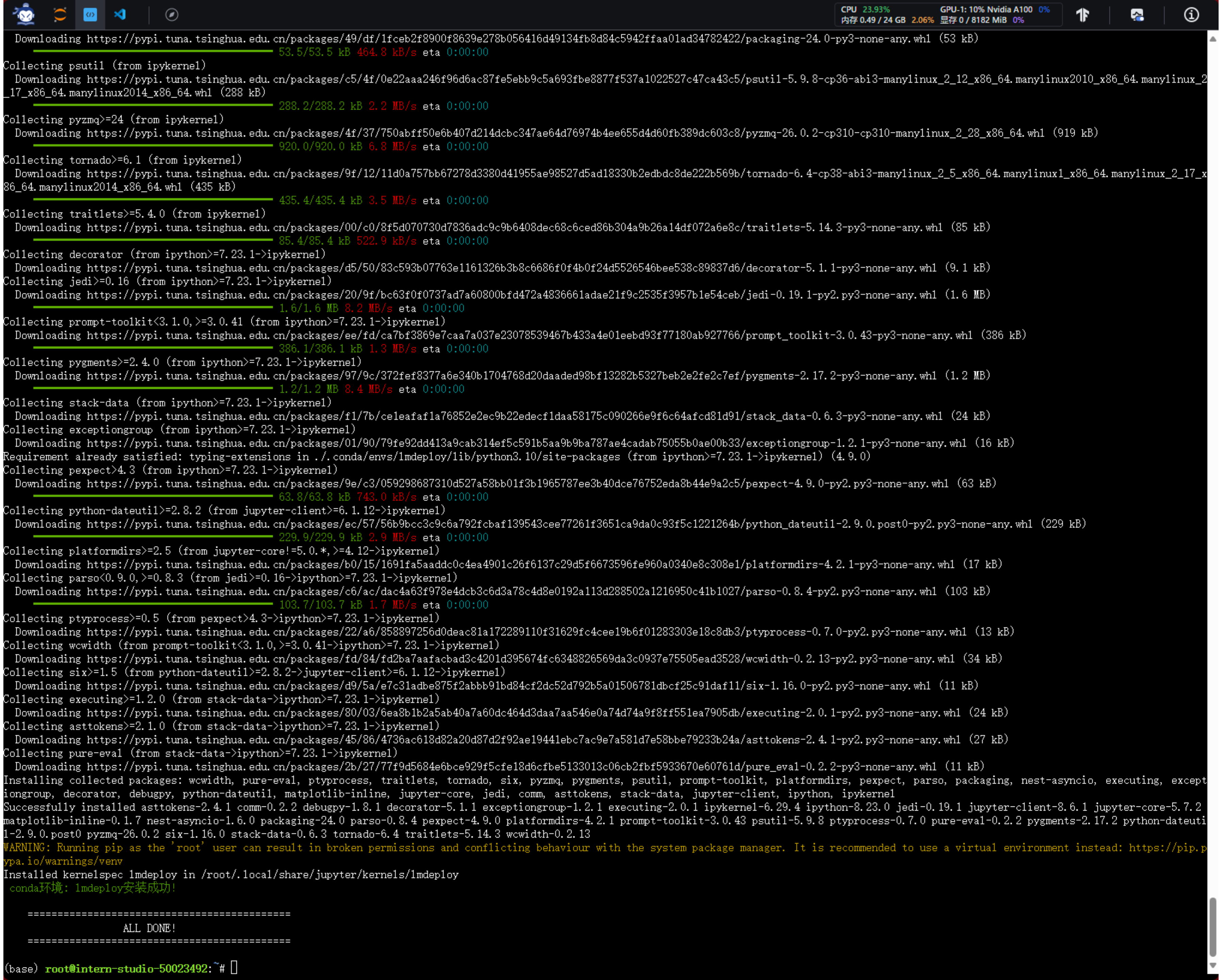
完成以下任务，并将实现过程记录截图：

- 配置 LMDeploy 运行环境
- 以命令行方式与 InternLM2-Chat-1.8B 模型对话

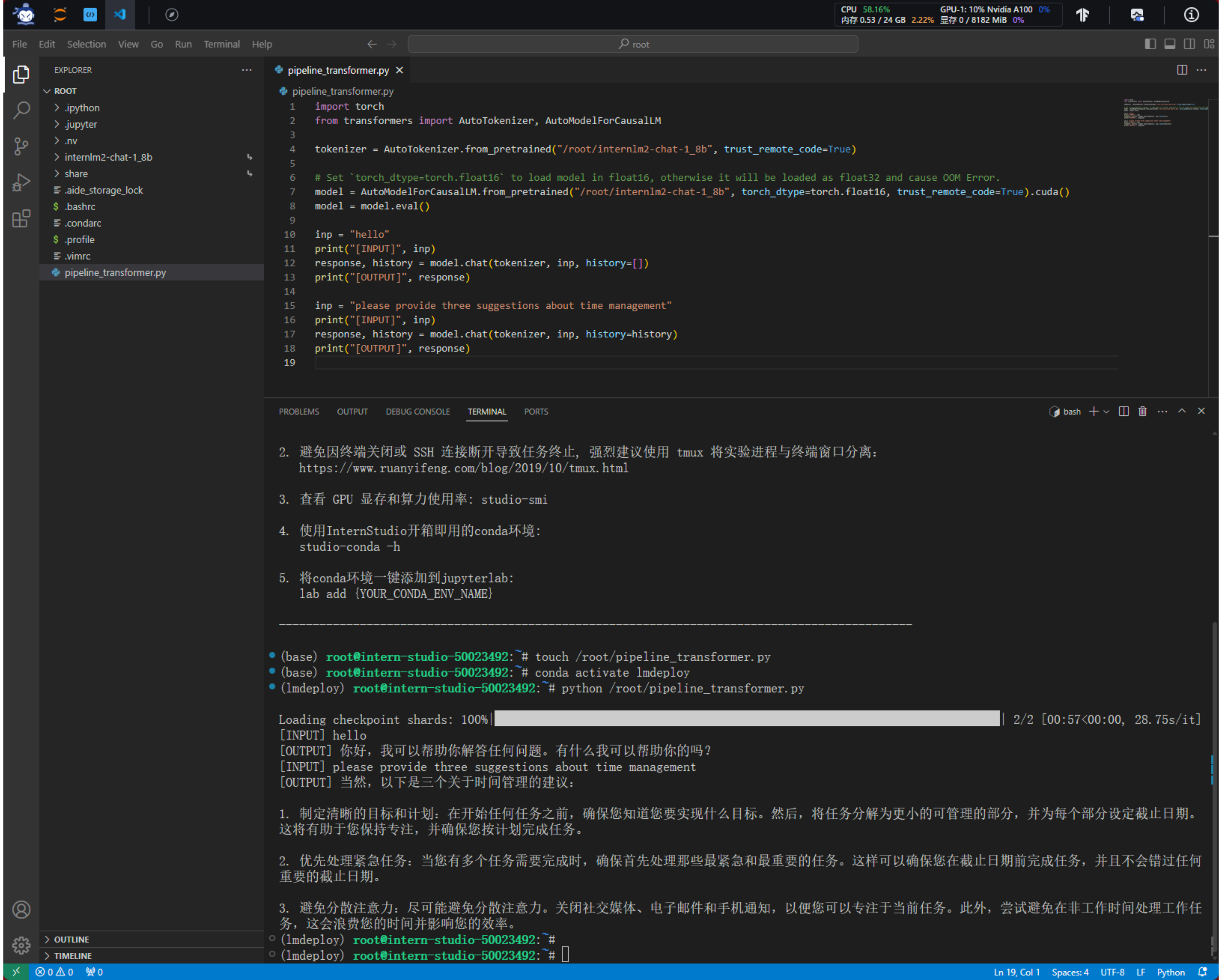
进阶作业

完成以下任务，并将实现过程记录截图：

- 设置KV Cache最大占用比例为0.4，开启W4A16量化，以命令行方式与模型对话。（优秀学员必做）
- 以API Server方式启动 lmdeploy，开启 W4A16量化，调整KV Cache的占用比例为0.4，分别使用命令行客户端与Gradio网页客户端与模型对话。（优秀学员必做）
- 使用W4A16量化，调整KV Cache的占用比例为0.4，使用Python代码集成的方式运行internlm2-chat-1.8b模型。（优秀学员必做）
- 使用 LMDeploy 运行视觉多模态大模型 llava gradio demo。（优秀学员必做）
- 将 LMDeploy Web Demo 部署到 OpenXLab 。

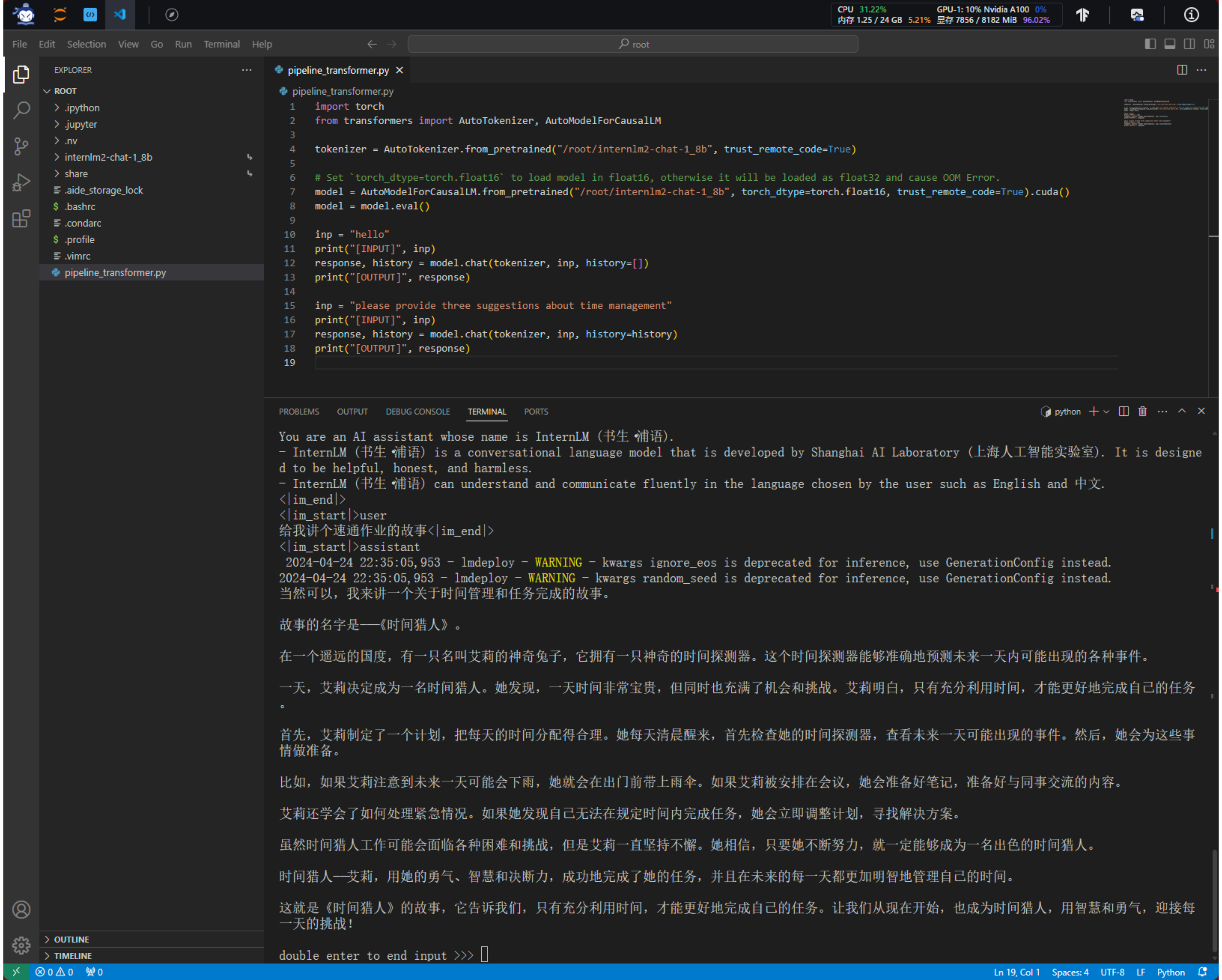


```
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/49/df/1fceb2f8900f8639e278b056416d49134fb8d84c5942ffaa01ad34782422/packaging-24.0-py3-none-any.whl (53 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 53.5/53.5 kB 464.8 kB/s eta 0:00:00
Collecting psutil (from ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/c5/4f/0e22aaa246f96d6ac87fe5ebb9c5a693fbe8877f537a1022527c47ca43c5/psutil-5.9.8-cp36-abi3-manylinux_2_12_x86_64.manylinux2010_x86_64.manylinux_2
_17_x86_64.manylinux2014_x86_64.whl (288 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 288.2/288.2 kB 2.2 MB/s eta 0:00:00
Collecting pyzmq>=24 (from ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/4f/37/750abff50e6b407d214dcbc347ae64d76974b4ee655d4d60fb389dc603c8/pyzmq-26.0.2-cp310-cp310-manylinux_2_28_x86_64.whl (919 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 920.0/920.0 kB 6.8 MB/s eta 0:00:00
Collecting tornado>=6.1 (from ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/9f/12/11d0a757bb67278d3380d41955ae98527d5ad18330b2edbdc8de222b569b/tornado-6.4-cp38-abi3-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x
86_64.manylinux2014_x86_64.whl (435 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 435.4/435.4 kB 3.5 MB/s eta 0:00:00
Collecting traitlets>=5.4.0 (from ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/00/c0/8f5d070730d7836adc9c9b6408dec68c6ced86b304a9b26a14df072a6e8c/traitlets-5.14.3-py3-none-any.whl (85 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 85.4/85.4 kB 522.9 kB/s eta 0:00:00
Collecting decorator (from ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/d5/50/83c593b07763e1161326b3b8c6686f0f4b0f24d5526546bee538c89837d6/decorator-5.1.1-py3-none-any.whl (9.1 kB)
Collecting jedi>=0.16 (from ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/20/9f/bc63f0f0737ad7a60800bfd472a4836661adae21f9c2535f3957b1e54ceb/jedi-0.19.1-py2.py3-none-any.whl (1.6 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.6/1.6 MB 8.2 MB/s eta 0:00:00
Collecting prompt-toolkit<3.1.0,>=3.0.41 (from ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/ee/fd/ca7bf3869e7caa7a037e23078539467b433a4e01eebd93f77180ab927766/prompt_toolkit-3.0.43-py3-none-any.whl (386 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 386.1/386.1 kB 1.3 MB/s eta 0:00:00
Collecting pygments>=2.4.0 (from ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/97/9c/372fef8377a6e340b1704768d20daaded98bf13282b5327beb2e2fe2c7ef/pygments-2.17.2-py3-none-any.whl (1.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.2/1.2 MB 8.4 MB/s eta 0:00:00
Collecting stack-data (from ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/f1/7b/ce1eafaf1a76852e2ec9b22edecf1daa58175c090266e9f6c64afcd81d91/stack_data-0.6.3-py3-none-any.whl (24 kB)
Collecting exceptiongroup (from ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/01/90/79fe92dd413a9cab314ef5c591b5aa9b9ba787ae4cadab75055b0ae00b33/exceptiongroup-1.2.1-py3-none-any.whl (16 kB)
Requirement already satisfied: typing-extensions in ./.conda/envs/lmdeploy/lib/python3.10/site-packages (from ipython>=7.23.1->ipykernel) (4.9.0)
Collecting pexpect>4.3 (from ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/9e/c3/059298687310d527a58bb01f3b1965787ee3b40dce76752eda8b44e9a2c5/pexpect-4.9.0-py2.py3-none-any.whl (63 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 63.8/63.8 kB 743.0 kB/s eta 0:00:00
Collecting python-dateutil>=2.8.2 (from jupyter-client>=6.1.12->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/ec/57/56b9bcc3c9c6a792fcbaf139543cee77261f3651ca9da0c93f5c1221264b/python_dateutil-2.9.0.post0-py2.py3-none-any.whl (229 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 229.9/229.9 kB 2.9 MB/s eta 0:00:00
Collecting platformdirs>=2.5 (from jupyter-core!=5.0.*,>=4.12->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/b0/15/1691fa5aaddc0c4ea4901c26f6137c29d5f6673596fe960a0340e8c308e1/platformdirs-4.2.1-py3-none-any.whl (17 kB)
Collecting parso<0.9.0,>=0.8.3 (from jedi>=0.16->ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/c6/ac/dac4a63f978e4dcb3c6d3a78c4d8e0192a113d288502a1216950c41b1027/parso-0.8.4-py2.py3-none-any.whl (103 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 103.7/103.7 kB 1.7 MB/s eta 0:00:00
Collecting ptyprocess>=0.5 (from pexpect>4.3->ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/22/a6/858897256d0deac81a172289110f31629fc4cee19b6f01283303e18c8db3/ptyprocess-0.7.0-py2.py3-none-any.whl (13 kB)
Collecting wcwidth (from prompt-toolkit<3.1.0,>=3.0.41->ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/fd/84/fd2ba7aafacbad3c4201d395674fc6348826569da3c0937e75505ead3528/wcwidth-0.2.13-py2.py3-none-any.whl (34 kB)
Collecting six>=1.5 (from python-dateutil>=2.8.2->jupyter-client>=6.1.12->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/d9/5a/e7c31adbe875f2abbb91bd84cf2dc52d792b5a01506781dbcf25c91daf11/six-1.16.0-py2.py3-none-any.whl (11 kB)
Collecting executing>=1.2.0 (from stack-data->ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/80/03/6ea8b1b2a5ab40a7a60dc464d3daa7aa546e0a74d74a9f8ff551ea7905db/executing-2.0.1-py2.py3-none-any.whl (24 kB)
Collecting asttokens>=2.1.0 (from stack-data->ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/45/86/4736ac618d82a20d87d2f92ae19441ebc7ac9e7a581d7e58bbe79233b24a/asttokens-2.4.1-py2.py3-none-any.whl (27 kB)
Collecting pure-eval (from stack-data->ipython>=7.23.1->ipykernel)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/2b/27/77f9d5684e6bce929f5cfe18d6cfbe5133013c06cb2fbf5933670e60761d/pure_eval-0.2.2-py3-none-any.whl (11 kB)
Installing collected packages: wcwidth, pure-eval, ptyprocess, traitlets, tornado, six, pyzmq, pygments, psutil, prompt-toolkit, platformdirs, pexpect, parso, packaging, nest-asyncio, executing, except
iongroup, decorator, debugpy, python-dateutil, matplotlib-inline, jupyter-core, jedi, comm, asttokens, stack-data, jupyter-client, ipython, ipykernel
Successfully installed asttokens-2.4.1 comm-0.2.2 debugpy-1.8.1 decorator-5.1.1 exceptiongroup-1.2.1 executing-2.0.1 ipykernel-6.29.4 ipython-8.23.0 jedi-0.19.1 jupyter-client-8.6.1 jupyter-core-5.7.2
matplotlib-inline-0.1.7 nest-asyncio-1.6.0 packaging-24.0 parso-0.8.4 pexpect-4.9.0 platformdirs-4.2.1 prompt-toolkit-3.0.43 psutil-5.9.8 ptyprocess-0.7.0 pure-eval-0.2.2 pygments-2.17.2 python-dateuti
l-2.9.0.post0 pyzmq-26.0.2 six-1.16.0 stack-data-0.6.3 tornado-6.4 traitlets-5.14.3 wcwidth-0.2.13
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.p
ypa.io/warnings/venv
Installed kernelspec lmdeploy in /root/.local/share/jupyter/kernels/lmdeploy
 conda环境: lmdeploy安装成功!

    ============================================
                    ALL DONE!
    ============================================

(base) root@intern-studio-50023492:~# 
```



```python
import torch
from transformers import AutoTokenizer, AutoModelForCausalLM

tokenizer = AutoTokenizer.from_pretrained("/root/internlm2-chat-1_8b", trust_remote_code=True)

# Set `torch_dtype=torch.float16` to load model in float16, otherwise it will be loaded as float32 and cause OOM Error.
model = AutoModelForCausalLM.from_pretrained("/root/internlm2-chat-1_8b", torch_dtype=torch.float16, trust_remote_code=True).cuda()
model = model.eval()

inp = "hello"
print("[INPUT]", inp)
response, history = model.chat(tokenizer, inp, history=[])
print("[OUTPUT]", response)

inp = "please provide three suggestions about time management"
print("[INPUT]", inp)
response, history = model.chat(tokenizer, inp, history=history)
print("[OUTPUT]", response)

```

```
2. 避免因终端关闭或 SSH 连接断开导致任务终止，强烈建议使用 tmux 将实验进程与终端窗口分离：
   https://www.ruanyifeng.com/blog/2019/10/tmux.html

3. 查看 GPU 显存和算力使用率: studio-smi

4. 使用InternStudio开箱即用的conda环境：
   studio-conda -h

5. 将conda环境一键添加到jupyterlab：
   lab add {YOUR_CONDA_ENV_NAME}

------------------------------------------------------------------------------------------------

● (base) root@intern-studio-50023492:~# touch /root/pipeline_transformer.py
● (base) root@intern-studio-50023492:~# conda activate lmdeploy
● (lmdeploy) root@intern-studio-50023492:~# python /root/pipeline_transformer.py

Loading checkpoint shards: 100%|██████████████████████████████████████| 2/2 [00:57<00:00, 28.75s/it]
[INPUT] hello
[OUTPUT] 你好，我可以帮助你解答任何问题。有什么我可以帮助你的吗？
[INPUT] please provide three suggestions about time management
[OUTPUT] 当然，以下是三个关于时间管理的建议：

1. 制定清晰的目标和计划：在开始任何任务之前，确保您知道您要实现什么目标。然后，将任务分解为更小的可管理的部分，并为每个部分设定截止日期。
这将有助于您保持专注，并确保您按计划完成任务。

2. 优先处理紧急任务：当您有多个任务需要完成时，确保首先处理那些最紧急和最重要的任务。这样可以确保您在截止日期前完成任务，并且不会错过任何
重要的截止日期。

3. 避免分散注意力：尽可能避免分散注意力。关闭社交媒体、电子邮件和手机通知，以便您可以专注于当前任务。此外，尝试避免在非工作时间处理工作任
务，这会浪费您的时间并影响您的效率。
○ (lmdeploy) root@intern-studio-50023492:~#
○ (lmdeploy) root@intern-studio-50023492:~# 
```



pipeline_transformer.py

```python
import torch
from transformers import AutoTokenizer, AutoModelForCausalLM

tokenizer = AutoTokenizer.from_pretrained("/root/internlm2-chat-1_8b", trust_remote_code=True)

# Set `torch_dtype=torch.float16` to load model in float16, otherwise it will be loaded as float32 and cause OOM Error.
model = AutoModelForCausalLM.from_pretrained("/root/internlm2-chat-1_8b", torch_dtype=torch.float16, trust_remote_code=True).cuda()
model = model.eval()

inp = "hello"
print("[INPUT]", inp)
response, history = model.chat(tokenizer, inp, history=[])
print("[OUTPUT]", response)

inp = "please provide three suggestions about time management"
print("[INPUT]", inp)
response, history = model.chat(tokenizer, inp, history=history)
print("[OUTPUT]", response)

```

```
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 designe
d 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.
<|im_end|>
<|im_start|>user
给我讲个速通作业的故事<|im_end|>
<|im_start|>assistant
 2024-04-24 22:35:05,953 - lmdeploy - WARNING - kwargs ignore_eos is deprecated for inference, use GenerationConfig instead.
2024-04-24 22:35:05,953 - lmdeploy - WARNING - kwargs random_seed is deprecated for inference, use GenerationConfig instead.
当然可以，我来讲一个关于时间管理和任务完成的故事。

故事的名字是——《时间猎人》。

在一个遥远的国度，有一只名叫艾莉的神奇兔子，它拥有一只神奇的时间探测器。这个时间探测器能够准确地预测未来一天内可能出现的各种事件。

一天，艾莉决定成为一名时间猎人。她发现，一天时间非常宝贵，但同时也充满了机会和挑战。艾莉明白，只有充分利用时间，才能更好地完成自己的任务
。

首先，艾莉制定了一个计划，把每天的时间分配得合理。她每天清晨醒来，首先检查她的时间探测器，查看未来一天可能出现的事件。然后，她会为这些事
情做准备。

比如，如果艾莉注意到未来一天可能会下雨，她就会在出门前带上雨伞。如果艾莉被安排在会议，她会准备好笔记，准备好与同事交流的内容。

艾莉还学会了如何处理紧急情况。如果她发现自己无法在规定时间内完成任务，她会立即调整计划，寻找解决方案。

虽然时间猎人工作可能会面临各种困难和挑战，但是艾莉一直坚持不懈。她相信，只要她不断努力，就一定能够成为一名出色的时间猎人。

时间猎人——艾莉，用她的勇气、智慧和决断力，成功地完成了她的任务，并且在未来的每一天都更加明智地管理自己的时间。

这就是《时间猎人》的故事，它告诉我们，只有充分利用时间，才能更好地完成自己的任务。让我们从现在开始，也成为时间猎人，用智慧和勇气，迎接每
一天的挑战！

double enter to end input >>> 
```