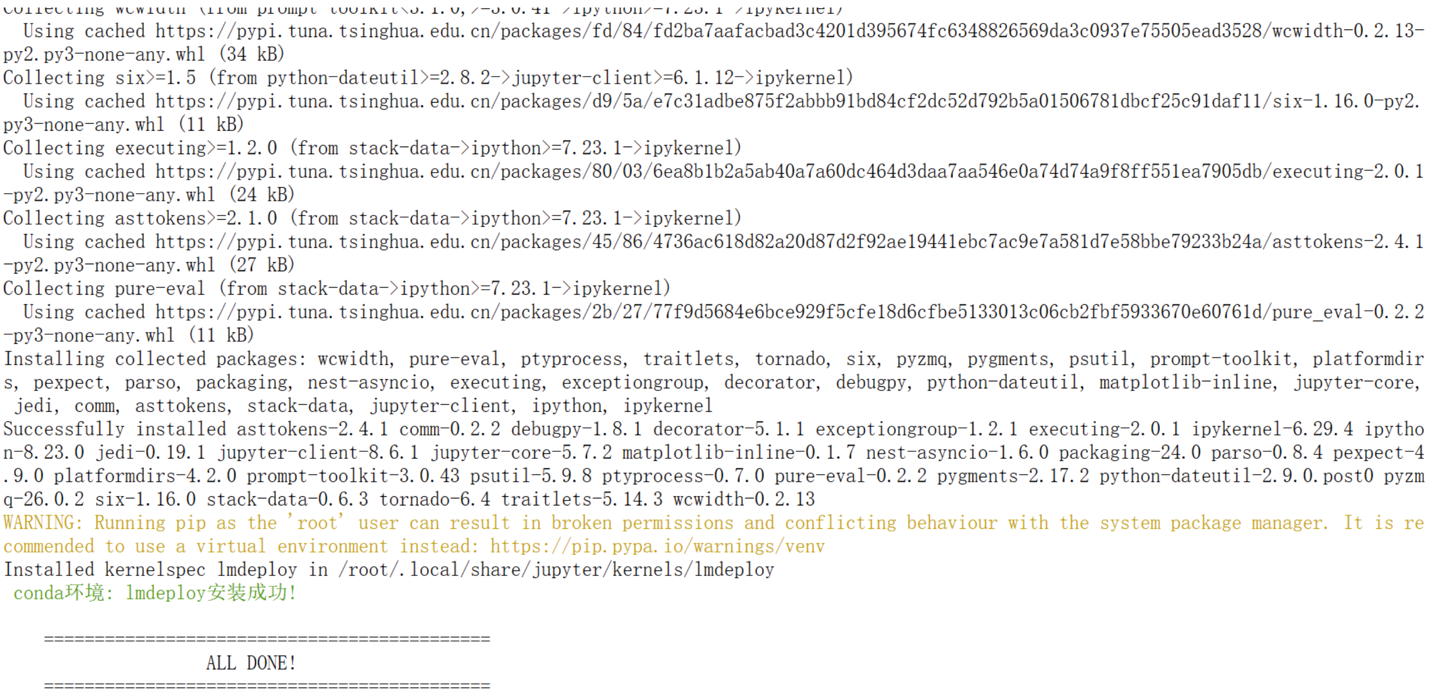
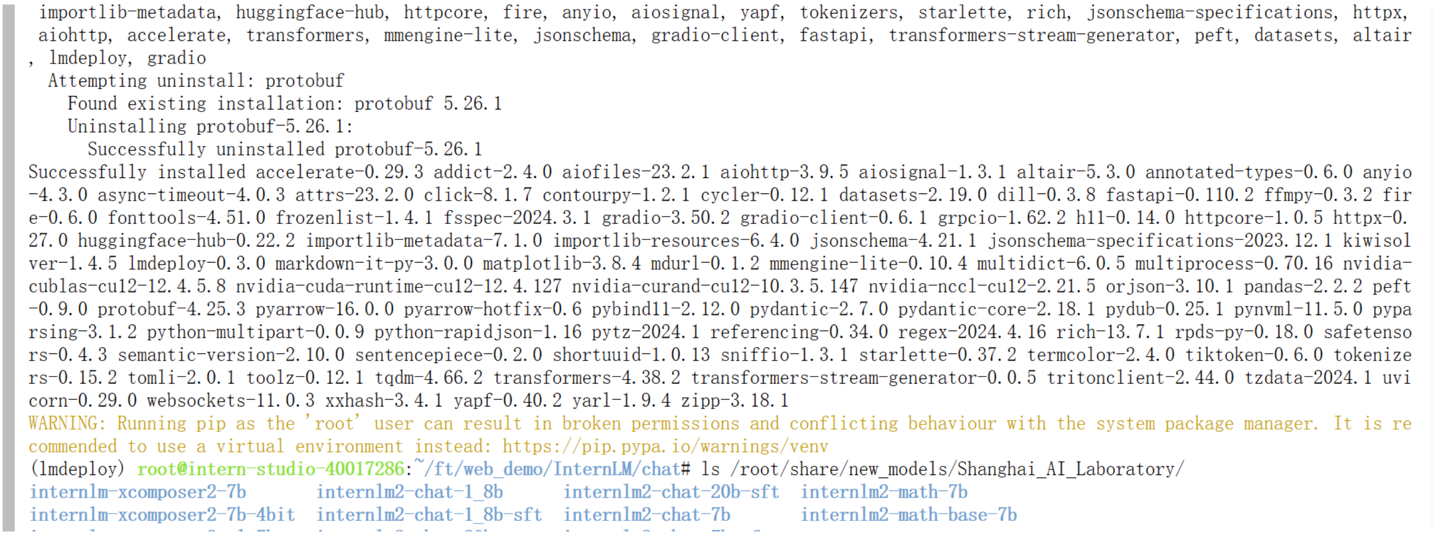
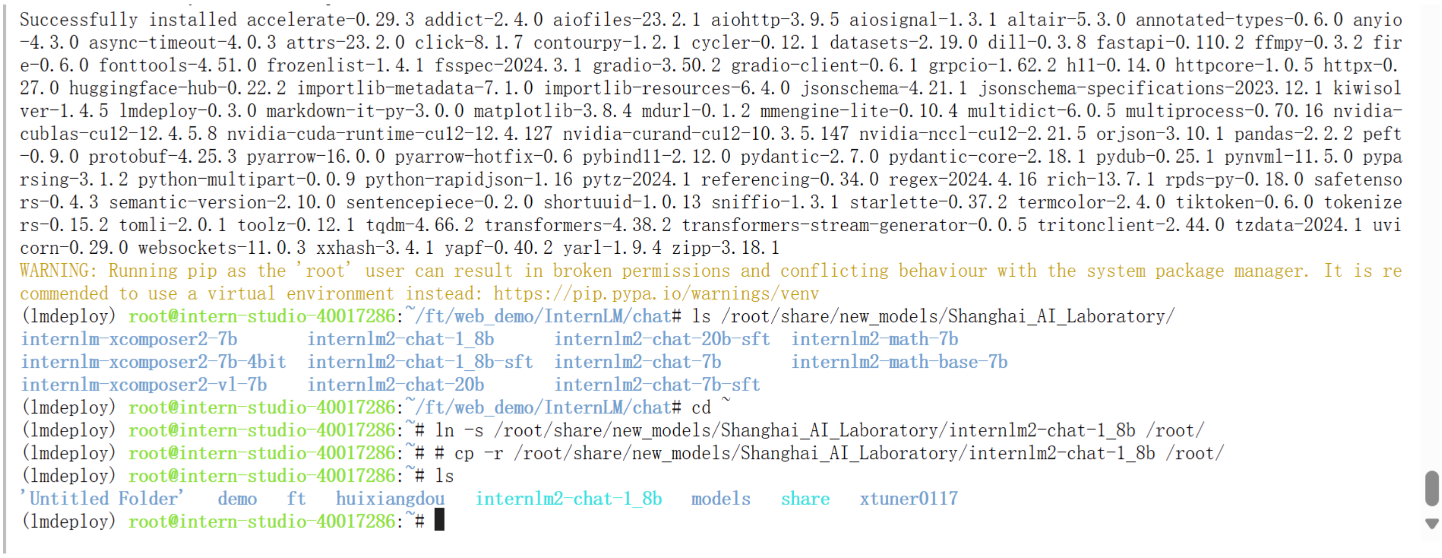
# 书生二期第五课作业



```
Collecting wcwidth (from prompt-toolkit<3.1.0,>=3.0.41->ipython>=7.23.1->ipykernel)
  Using cached 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)
  Using cached 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)
  Using cached 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)
  Using cached 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)
  Using cached 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, platformdir
s, pexpect, parso, packaging, nest-asyncio, executing, exceptiongroup, 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 ipytho
n-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-4.2.0 prompt-toolkit-3.0.43 psutil-5.9.8 ptyprocess-0.7.0 pure-eval-0.2.2 pygments-2.17.2 python-dateutil-2.9.0.post0 pyzm
q-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 re
commended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
Installed kernelspec lmdeploy in /root/.local/share/jupyter/kernels/lmdeploy
 conda环境: lmdeploy安装成功!

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



```
 importlib-metadata, huggingface-hub, httpcore, fire, anyio, aiosignal, yapf, tokenizers, starlette, rich, jsonschema-specifications, httpx,
 aiohttp, accelerate, transformers, mmengine-lite, jsonschema, gradio-client, fastapi, transformers-stream-generator, peft, datasets, altair
, lmdeploy, gradio
  Attempting uninstall: protobuf
    Found existing installation: protobuf 5.26.1
    Uninstalling protobuf-5.26.1:
      Successfully uninstalled protobuf-5.26.1
Successfully installed accelerate-0.29.3 addict-2.4.0 aiofiles-23.2.1 aiohttp-3.9.5 aiosignal-1.3.1 altair-5.3.0 annotated-types-0.6.0 anyio
-4.3.0 async-timeout-4.0.3 attrs-23.2.0 click-8.1.7 contourpy-1.2.1 cycler-0.12.1 datasets-2.19.0 dill-0.3.8 fastapi-0.110.2 ffmpy-0.3.2 fir
e-0.6.0 fonttools-4.51.0 frozenlist-1.4.1 fsspec-2024.3.1 gradio-3.50.2 gradio-client-0.6.1 grpcio-1.62.2 h11-0.14.0 httpcore-1.0.5 httpx-0.
27.0 huggingface-hub-0.22.2 importlib-metadata-7.1.0 importlib-resources-6.4.0 jsonschema-4.21.1 jsonschema-specifications-2023.12.1 kiwisol
ver-1.4.5 lmdeploy-0.3.0 markdown-it-py-3.0.0 matplotlib-3.8.4 mdurl-0.1.2 mmengine-lite-0.10.4 multidict-6.0.5 multiprocess-0.70.16 nvidia-
cublas-cu12-12.4.5.8 nvidia-cuda-runtime-cu12-12.4.127 nvidia-curand-cu12-10.3.5.147 nvidia-nccl-cu12-2.21.5 orjson-3.10.1 pandas-2.2.2 peft
-0.9.0 protobuf-4.25.3 pyarrow-16.0.0 pyarrow-hotfix-0.6 pybind11-2.12.0 pydantic-2.7.0 pydantic-core-2.18.1 pydub-0.25.1 pynvml-11.5.0 pypa
rsing-3.1.2 python-multipart-0.0.9 python-rapidjson-1.16 pytz-2024.1 referencing-0.34.0 regex-2024.4.16 rich-13.7.1 rpds-py-0.18.0 safetenso
rs-0.4.3 semantic-version-2.10.0 sentencepiece-0.2.0 shortuuid-1.0.13 sniffio-1.3.1 starlette-0.37.2 termcolor-2.4.0 tiktoken-0.6.0 tokenize
rs-0.15.2 tomli-2.0.1 toolz-0.12.1 tqdm-4.66.2 transformers-4.38.2 transformers-stream-generator-0.0.5 tritonclient-2.44.0 tzdata-2024.1 uvi
corn-0.29.0 websockets-11.0.3 xxhash-3.4.1 yapf-0.40.2 yarl-1.9.4 zipp-3.18.1
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is re
commended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
(lmdeploy) root@intern-studio-40017286:~/ft/web_demo/InternLM/chat# ls /root/share/new_models/Shanghai_AI_Laboratory/
internlm-xcomposer2-7b       internlm2-chat-1_8b      internlm2-chat-20b-sft  internlm2-math-7b
internlm-xcomposer2-7b-4bit  internlm2-chat-1_8b-sft  internlm2-chat-7b       internlm2-math-base-7b
```



```
Successfully installed accelerate-0.29.3 addict-2.4.0 aiofiles-23.2.1 aiohttp-3.9.5 aiosignal-1.3.1 altair-5.3.0 annotated-types-0.6.0 anyio
-4.3.0 async-timeout-4.0.3 attrs-23.2.0 click-8.1.7 contourpy-1.2.1 cycler-0.12.1 datasets-2.19.0 dill-0.3.8 fastapi-0.110.2 ffmpy-0.3.2 fir
e-0.6.0 fonttools-4.51.0 frozenlist-1.4.1 fsspec-2024.3.1 gradio-3.50.2 gradio-client-0.6.1 grpcio-1.62.2 h11-0.14.0 httpcore-1.0.5 httpx-0.
27.0 huggingface-hub-0.22.2 importlib-metadata-7.1.0 importlib-resources-6.4.0 jsonschema-4.21.1 jsonschema-specifications-2023.12.1 kiwisol
ver-1.4.5 lmdeploy-0.3.0 markdown-it-py-3.0.0 matplotlib-3.8.4 mdurl-0.1.2 mmengine-lite-0.10.4 multidict-6.0.5 multiprocess-0.70.16 nvidia-
cublas-cu12-12.4.5.8 nvidia-cuda-runtime-cu12-12.4.127 nvidia-curand-cu12-10.3.5.147 nvidia-nccl-cu12-2.21.5 orjson-3.10.1 pandas-2.2.2 peft
-0.9.0 protobuf-4.25.3 pyarrow-16.0.0 pyarrow-hotfix-0.6 pybind11-2.12.0 pydantic-2.7.0 pydantic-core-2.18.1 pydub-0.25.1 pynvml-11.5.0 pypa
rsing-3.1.2 python-multipart-0.0.9 python-rapidjson-1.16 pytz-2024.1 referencing-0.34.0 regex-2024.4.16 rich-13.7.1 rpds-py-0.18.0 safetenso
rs-0.4.3 semantic-version-2.10.0 sentencepiece-0.2.0 shortuuid-1.0.13 sniffio-1.3.1 starlette-0.37.2 termcolor-2.4.0 tiktoken-0.6.0 tokenize
rs-0.15.2 tomli-2.0.1 toolz-0.12.1 tqdm-4.66.2 transformers-4.38.2 transformers-stream-generator-0.0.5 tritonclient-2.44.0 tzdata-2024.1 uvi
corn-0.29.0 websockets-11.0.3 xxhash-3.4.1 yapf-0.40.2 yarl-1.9.4 zipp-3.18.1
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is re
commended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
(lmdeploy) root@intern-studio-40017286:~/ft/web_demo/InternLM/chat# ls /root/share/new_models/Shanghai_AI_Laboratory/
internlm-xcomposer2-7b       internlm2-chat-1_8b      internlm2-chat-20b-sft  internlm2-math-7b
internlm-xcomposer2-7b-4bit  internlm2-chat-1_8b-sft  internlm2-chat-7b       internlm2-math-base-7b
internlm-xcomposer2-vl-7b    internlm2-chat-20b       internlm2-chat-7b-sft
(lmdeploy) root@intern-studio-40017286:~/ft/web_demo/InternLM/chat# cd ~
(lmdeploy) root@intern-studio-40017286:~# ln -s /root/share/new_models/Shanghai_AI_Laboratory/internlm2-chat-1_8b /root/
(lmdeploy) root@intern-studio-40017286:~# # cp -r /root/share/new_models/Shanghai_AI_Laboratory/internlm2-chat-1_8b /root/
(lmdeploy) root@intern-studio-40017286:~# ls
'Untitled Folder'   demo   ft   huixiangdou   internlm2-chat-1_8b   models   share   xtuner0117
(lmdeploy) root@intern-studio-40017286:~# 
```

**报错如下：**

(lmdeploy) root@intern-studio-40017286:~# python /root/pipeline_transformer.py

Traceback (most recent call last):

 File "/root/pipeline_transformer.py", line 1, in <module>

  import torch

ModuleNotFoundError: No module named 'torch'

(lmdeploy) root@intern-studio-40017286:~# pip install torch

Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple

Collecting torch

 Downloading https://pypi.tuna.tsinghua.edu.cn/packages/35/3a/a39f354fa3119785be87e2f94ffa2620f8a270c8560f7356358ee62fb4c5/torch-2.3.0-cp311-cp311-manylinux1_x86_64.whl (779.2 MB)

———————————————————————————————————————

779.2/779.2 MB 10.4 MB/s eta 0:00:00

Collecting filelock (from torch)

 Using cached https://pypi.tuna.tsinghua.edu.cn/packages/41/24/0b023b6537dfc9bae2c779353998e3e99ac7dfff4222fc6126650e93c3f3/filelock-3.14.0-py3-none-any.whl (12 kB)

Collecting typing-extensions>=4.8.0 (from torch)

 Downloading https://pypi.tuna.tsinghua.edu.cn/packages/01/f3/936e209267d6ef7510322191003885de524fc48d1b43269810cd589ceaf5/typing_extensions-4.11.0-py3-none-any.whl (34 kB)

Collecting sympy (from torch)

```
  Using cached
https://pypi.tuna.tsinghua.edu.cn/packages/d2/05/e6600db80270777c4a64238a98d442f0fd07cc
8915be2a1c16da7f2b9e74/sympy-1.12-py3-none-any.whl (5.7 MB)

Collecting networkx (from torch)

  Downloading
https://pypi.tuna.tsinghua.edu.cn/packages/38/e9/5f72929373e1a0e8d142a130f3f97e6ff920070f
87f91c4e13e40e0fba5a/networkx-3.3-py3-none-any.whl (1.7 MB)

   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.7/1.7
MB 33.4 MB/s eta 0:00:00

Collecting jinja2 (from torch)

  Downloading
https://pypi.tuna.tsinghua.edu.cn/packages/30/6d/6de6be2d02603ab56e72997708809e8a5b0fbf
ee080735109b40a3564843/Jinja2-3.1.3-py3-none-any.whl (133 kB)

   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
133.2/133.2 kB 11.3 MB/s eta 0:00:00

Collecting fsspec (from torch)

  Using cached
https://pypi.tuna.tsinghua.edu.cn/packages/93/6d/66d48b03460768f523da62a57a7e14e5e95fdf
339d79e996ce3cecda2cdb/fsspec-2024.3.1-py3-none-any.whl (171 kB)

Collecting nvidia-cuda-nvrtc-cu12==12.1.105 (from torch)

  Using cached
https://pypi.tuna.tsinghua.edu.cn/packages/b6/9f/c64c03f49d6fbc56196664d05dba14e3a56103
8a81a638eeb47f4d4cfd48/nvidia_cuda_nvrtc_cu12-12.1.105-py3-none-manylinux1_x86_64.whl
(23.7 MB)

Collecting nvidia-cuda-runtime-cu12==12.1.105 (from torch)

  Downloading
https://pypi.tuna.tsinghua.edu.cn/packages/eb/d5/c68b1d2cdfcc59e72e8a5949a37ddb22ae6cad
e80cd4a57a84d4c8b55472/nvidia_cuda_runtime_cu12-12.1.105-py3-none-
manylinux1_x86_64.whl (823 kB)

   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
823.6/823.6 kB 26.2 MB/s eta 0:00:00

Collecting nvidia-cuda-cupti-cu12==12.1.105 (from torch)

  Using cached
https://pypi.tuna.tsinghua.edu.cn/packages/7e/00/6b218edd739ecfc60524e585ba8e6b00554dd
```

```
908de2c9c66c1af3e44e18d/nvidia_cuda_cupti_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (14.1 MB)
Collecting nvidia-cudnn-cu12==8.9.2.26 (from torch)
  Using cached
https://pypi.tuna.tsinghua.edu.cn/packages/ff/74/a2e2be7fb83aaedec84f391f082cf765dfb635e7caa9b49065f73e4835d8/nvidia_cudnn_cu12-8.9.2.26-py3-none-manylinux1_x86_64.whl (731.7 MB)
Collecting nvidia-cublas-cu12==12.1.3.1 (from torch)
  Downloading
https://pypi.tuna.tsinghua.edu.cn/packages/37/6d/121efd7382d5b0284239f4ab1fc1590d86d34ed4a4a2fdb13b30ca8e5740/nvidia_cublas_cu12-12.1.3.1-py3-none-manylinux1_x86_64.whl (410.6 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
410.6/410.6 MB 18.8 MB/s eta 0:00:00
Collecting nvidia-cufft-cu12==11.0.2.54 (from torch)
  Using cached
https://pypi.tuna.tsinghua.edu.cn/packages/86/94/eb540db023ce1d162e7bea9f8f5aa781d57c65aed513c33ee9a5123ead4d/nvidia_cufft_cu12-11.0.2.54-py3-none-manylinux1_x86_64.whl (121.6 MB)
Collecting nvidia-curand-cu12==10.3.2.106 (from torch)
  Downloading
https://pypi.tuna.tsinghua.edu.cn/packages/44/31/4890b1c9abc496303412947fc7dcea3d14861720642b49e8ceed89636705/nvidia_curand_cu12-10.3.2.106-py3-none-manylinux1_x86_64.whl (56.5 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
56.5/56.5 MB 69.2 MB/s eta 0:00:00
Collecting nvidia-cusolver-cu12==11.4.5.107 (from torch)
  Using cached
https://pypi.tuna.tsinghua.edu.cn/packages/bc/1d/8de1e5c67099015c834315e333911273a8c6aaba78923dd1d1e25fc5f217/nvidia_cusolver_cu12-11.4.5.107-py3-none-manylinux1_x86_64.whl (124.2 MB)
Collecting nvidia-cusparse-cu12==12.1.0.106 (from torch)
  Using cached
https://pypi.tuna.tsinghua.edu.cn/packages/65/5b/cfaeebf25cd9fdec14338ccb16f6b2c4c7fa9163aefcf057d86b9cc248bb/nvidia_cusparse_cu12-12.1.0.106-py3-none-manylinux1_x86_64.whl (196.0 MB)
```

```
Collecting nvidia-nccl-cu12==2.20.5 (from torch)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/4b/2a/0a131f572aa09f741c30ccd45a8e56316e8be8dfc7bc19bf0ab7cfef7b19/nvidia_nccl_cu12-2.20.5-py3-none-manylinux2014_x86_64.whl (176.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
176.2/176.2 MB 30.9 MB/s eta 0:00:00
Collecting nvidia-nvtx-cu12==12.1.105 (from torch)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/da/d3/8057f0587683ed2fcd4dbfbdfdfa807b9160b809976099d36b8f60d08f03/nvidia_nvtx_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (99 kB)
Collecting triton==2.3.0 (from torch)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/3c/00/84e0006f2025260fa111ddfc66194bd1af731b3ee18e2fd611a00f290b5e/triton-2.3.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (168.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
168.1/168.1 MB 29.9 MB/s eta 0:00:00
Collecting nvidia-nvjitlink-cu12 (from nvidia-cusolver-cu12==11.4.5.107->torch)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/ff/ff/847841bacfbefc97a00036e0fce5a0f086b640756dc38caea5e1bb002655/nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (21.1 MB)
Collecting MarkupSafe>=2.0 (from jinja2->torch)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/97/18/c30da5e7a0e7f4603abfc6780574131221d9148f323752c2755d48abad30/MarkupSafe-2.1.5-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (28 kB)
Collecting mpmath>=0.19 (from sympy->torch)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/43/e3/7d92a15f894aa0c9c4b49b8ee9ac9850d6e63b03c9c32c0367a13ae62209/mpmath-1.3.0-py3-none-any.whl (536 kB)
Installing collected packages: mpmath, typing-extensions, sympy, nvidia-nvtx-cu12, nvidia-nvjitlink-cu12, nvidia-nccl-cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia-cuda-runtime-cu12, nvidia-cuda-nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia-cublas-cu12, networkx,
MarkupSafe, fsspec, filelock, triton, nvidia-cusparse-cu12, nvidia-cudnn-cu12, jinja2, nvidia-cusolver-cu12, torch

Successfully installed MarkupSafe-2.1.5 filelock-3.14.0 fsspec-2024.3.1 jinja2-3.1.3 mpmath-1.3.0 networkx-3.3 nvidia-cublas-cu12-12.1.3.1 nvidia-cuda-cupti-cu12-12.1.105 nvidia-cuda-nvrtc-cu12-12.1.105 nvidia-cuda-runtime-cu12-12.1.105 nvidia-cudnn-cu12-8.9.2.26 nvidia-cufft-cu12-11.0.2.54 nvidia-curand-cu12-10.3.2.106 nvidia-cusolver-cu12-11.4.5.107 nvidia-cusparse-cu12-12.1.0.106 nvidia-nccl-cu12-2.20.5 nvidia-nvjitlink-cu12-12.4.127 nvidia-nvtx-cu12-12.1.105 sympy-1.12 torch-2.3.0 triton-2.3.0 typing-extensions-4.11.0

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.pypa.io/warnings/venv

(lmdeploy) root@intern-studio-40017286:~#

(lmdeploy) root@intern-studio-40017286:~# python /root/pipeline_transformer.py

Traceback (most recent call last):

  File "/root/pipeline_transformer.py", line 2, in <module>

    from transformers import AutoTokenizer, AutoModelForCausalLM

ModuleNotFoundError: No module named 'transformers'

(lmdeploy) root@intern-studio-40017286:~# pip install transformers

Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple

Collecting transformers

  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/cf/90/2596ac2ab49c4df6ff1fceaf7f5afb18401ba2f326348ce1a6261a65e7ed/transformers-4.40.1-py3-none-any.whl (9.0 MB)

     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 9.0/9.0 MB 73.1 MB/s eta 0:00:00

Requirement already satisfied: filelock in ./.conda/lib/python3.11/site-packages (from transformers) (3.14.0)

Collecting huggingface-hub<1.0,>=0.19.3 (from transformers)

  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/05/c0/779afbad8e75565c09ffa24a88b5dd7e293c92b74eb09df6435fc58ac986/huggingface_hub-0.22.2-py3-none-any.whl (388 kB)

Collecting numpy>=1.17 (from transformers)

  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/3a/d0/edc009c27b406c4f9cbc79274d6e46d634d139
```

```
075492ad055e3d68445925/numpy-1.26.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (18.3 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 18.3/18.3 MB 112.2 MB/s eta 0:00:00
Requirement already satisfied: packaging>=20.0 in ./.conda/lib/python3.11/site-packages (from transformers) (23.1)
Collecting pyyaml>=5.1 (from transformers)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/7b/5e/efd033ab7199a0b2044dab3b9f7a4f6670e6a52c089de572e928d2873b06/PyYAML-6.0.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (757 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 757.7/757.7 kB 15.1 MB/s eta 0:00:00
Collecting regex!=2019.12.17 (from transformers)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/52/21/22e993e8151c94e9adc9fc5f09848bad538d12c6390cec91f0fb1f6c8ba3/regex-2024.4.28-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (785 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 785.2/785.2 kB 19.7 MB/s eta 0:00:00
Requirement already satisfied: requests in ./.conda/lib/python3.11/site-packages (from transformers) (2.31.0)
Collecting tokenizers<0.20,>=0.19 (from transformers)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/a7/03/fb50fc03f86016b227a967c8d474f90230c885c0d18f78acdfda7a96ce56/tokenizers-0.19.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3.6 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3.6/3.6 MB 71.2 MB/s eta 0:00:00
Collecting safetensors>=0.4.1 (from transformers)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/d5/85/1e7d2804cbf82204cde462d16f1cb0ff5814b03f559fb46ceaa6b7020db4/safetensors-0.4.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.2 MB)
```

```
    ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.2/1.2
MB 70.7 MB/s eta 0:00:00
Requirement already satisfied: tqdm>=4.27 in ./.conda/lib/python3.11/site-packages (from
transformers) (4.65.0)
Requirement already satisfied: fsspec>=2023.5.0 in ./.conda/lib/python3.11/site-packages (from
huggingface-hub<1.0,>=0.19.3->transformers) (2024.3.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in ./.conda/lib/python3.11/site-
packages (from huggingface-hub<1.0,>=0.19.3->transformers) (4.11.0)
Requirement already satisfied: charset-normalizer<4,>=2 in ./.conda/lib/python3.11/site-
packages (from requests->transformers) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in ./.conda/lib/python3.11/site-packages (from
requests->transformers) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in ./.conda/lib/python3.11/site-packages (from
requests->transformers) (1.26.16)
Requirement already satisfied: certifi>=2017.4.17 in ./.conda/lib/python3.11/site-packages (from
requests->transformers) (2023.7.22)
Installing collected packages: safetensors, regex, pyyaml, numpy, huggingface-hub, tokenizers,
transformers
Successfully installed huggingface-hub-0.22.2 numpy-1.26.4 pyyaml-6.0.1 regex-2024.4.28
safetensors-0.4.3 tokenizers-0.19.1 transformers-4.40.1
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.pypa.io/warnings/venv
(lmdeploy) root@intern-studio-40017286:~# python /root/pipeline_transformer.py
Traceback (most recent call last):
  File "/root/pipeline_transformer.py", line 4, in <module>
    tokenizer = AutoTokenizer.from_pretrained("/root/internlm2-chat-1_8b",
trust_remote_code=True)

^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
^^
  File "/root/.conda/lib/python3.11/site-
packages/transformers/models/auto/tokenization_auto.py", line 843, in from_pretrained
    tokenizer_class = get_class_from_dynamic_module(class_ref,
pretrained_model_name_or_path, **kwargs)
```

```
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/root/.conda/lib/python3.11/site-packages/transformers/dynamic_module_utils.py", line 501, in get_class_from_dynamic_module
    return get_class_in_module(class_name, final_module)
           ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/root/.conda/lib/python3.11/site-packages/transformers/dynamic_module_utils.py", line 201, in get_class_in_module
    module = importlib.machinery.SourceFileLoader(name, module_path).load_module()
             ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "<frozen importlib._bootstrap_external>", line 605, in _check_name_wrapper
  File "<frozen importlib._bootstrap_external>", line 1120, in load_module
  File "<frozen importlib._bootstrap_external>", line 945, in load_module
  File "<frozen importlib._bootstrap>", line 290, in _load_module_shim
  File "<frozen importlib._bootstrap>", line 721, in _load
  File "<frozen importlib._bootstrap>", line 690, in _load_unlocked
  File "<frozen importlib._bootstrap_external>", line 940, in exec_module
  File "<frozen importlib._bootstrap>", line 241, in _call_with_frames_removed
  File "/root/.cache/huggingface/modules/transformers_modules/internlm2-chat-1_8b/tokenization_internlm2_fast.py", line 35, in <module>
    from .tokenization_internlm2 import InternLM2Tokenizer
  File "/root/.cache/huggingface/modules/transformers_modules/internlm2-chat-1_8b/tokenization_internlm2.py", line 23, in <module>
    import sentencepiece as spm
ModuleNotFoundError: No module named 'sentencepiece'
(lmdeploy) root@intern-studio-40017286:~# pip install sentencepiece
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Collecting sentencepiece
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/fb/12/2f5c8d4764b00033cf1c935b702d3bb878d10be9f0b87f0253495832d85f/sentencepiece-0.2.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.3 MB)
```

```
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.3/1.3 MB 26.9 MB/s eta 0:00:00
Installing collected packages: sentencepiece
Successfully installed sentencepiece-0.2.0
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.pypa.io/warnings/venv
(lmdeploy) root@intern-studio-40017286:~# python /root/pipeline_transformer.py
Traceback (most recent call last):
  File "/root/pipeline_transformer.py", line 4, in <module>
    tokenizer = AutoTokenizer.from_pretrained("/root/internlm2-chat-1_8b", trust_remote_code=True)
                ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/root/.conda/lib/python3.11/site-packages/transformers/models/auto/tokenization_auto.py", line 847, in from_pretrained
    return tokenizer_class.from_pretrained(
           ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/root/.conda/lib/python3.11/site-packages/transformers/tokenization_utils_base.py", line 2089, in from_pretrained
    return cls._from_pretrained(
           ^^^^^^^^^^^^^^^^^^^^^
  File "/root/.conda/lib/python3.11/site-packages/transformers/tokenization_utils_base.py", line 2311, in _from_pretrained
    tokenizer = cls(*init_inputs, **init_kwargs)
                ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/root/.cache/huggingface/modules/transformers_modules/internlm2-chat-1_8b/tokenization_internlm2_fast.py", line 131, in __init__
    super().__init__(
  File "/root/.conda/lib/python3.11/site-packages/transformers/tokenization_utils_fast.py", line 114, in __init__
    fast_tokenizer = convert_slow_tokenizer(slow_tokenizer)
```

```
           ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/root/.conda/lib/python3.11/site-packages/transformers/convert_slow_tokenizer.py", line 1534, in convert_slow_tokenizer
    return converter_class(transformer_tokenizer).converted()
        ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/root/.conda/lib/python3.11/site-packages/transformers/convert_slow_tokenizer.py", line 546, in __init__
    requires_backends(self, "protobuf")
  File "/root/.conda/lib/python3.11/site-packages/transformers/utils/import_utils.py", line 1438, in requires_backends
    raise ImportError("".join(failed))
ImportError:
InternLM2Converter requires the protobuf library but it was not found in your environment. Checkout the instructions on the
installation page of its repo: https://github.com/protocolbuffers/protobuf/tree/master/python#installation and follow the ones
that match your environment. Please note that you may need to restart your runtime after installation.


(lmdeploy) root@intern-studio-40017286:~# pip install protobuf
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Collecting protobuf
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/2c/2a/d2741cad35fa5f06d9c59dda3274e5727ca11075dfd7de3f69c100efdcad/protobuf-5.26.1-cp37-abi3-manylinux2014_x86_64.whl (302 kB)
  ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
302.8/302.8 kB 6.3 MB/s eta 0:00:00


Installing collected packages: protobuf
Successfully installed protobuf-5.26.1
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.pypa.io/warnings/venv

(lmdeploy) root@intern-studio-40017286:~#

(lmdeploy) root@intern-studio-40017286:~# python /root/pipeline_transformer.py

Traceback (most recent call last):

  File "/root/pipeline_transformer.py", line 7, in <module>

    model = AutoModelForCausalLM.from_pretrained("/root/internlm2-chat-1_8b", torch_dtype=torch.float16, trust_remote_code=True).cuda()


^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

  File "/root/.conda/lib/python3.11/site-packages/transformers/models/auto/auto_factory.py", line 550, in from_pretrained

    model_class = get_class_from_dynamic_module(

                  ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

  File "/root/.conda/lib/python3.11/site-packages/transformers/dynamic_module_utils.py", line 489, in get_class_from_dynamic_module

    final_module = get_cached_module_file(

                   ^^^^^^^^^^^^^^^^^^^^^^^^^

  File "/root/.conda/lib/python3.11/site-packages/transformers/dynamic_module_utils.py", line 315, in get_cached_module_file

    modules_needed = check_imports(resolved_module_file)

                     ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

  File "/root/.conda/lib/python3.11/site-packages/transformers/dynamic_module_utils.py", line 180, in check_imports

    raise ImportError(

ImportError: This modeling file requires the following packages that were not found in your environment: einops. Run pip install einops

(lmdeploy) root@intern-studio-40017286:~# pip install einops

pip install einops

Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple

Collecting einops

  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/44/5a/f0b9ad6c0a9017e62d4735daaeb11ba3b6c00
```

9d69a26141b258cd37b5588/einops-0.8.0-py3-none-any.whl (43 kB)

43.2/43.2 kB 2.7 MB/s eta 0:00:00

Installing collected packages: einops

Successfully installed einops-0.8.0

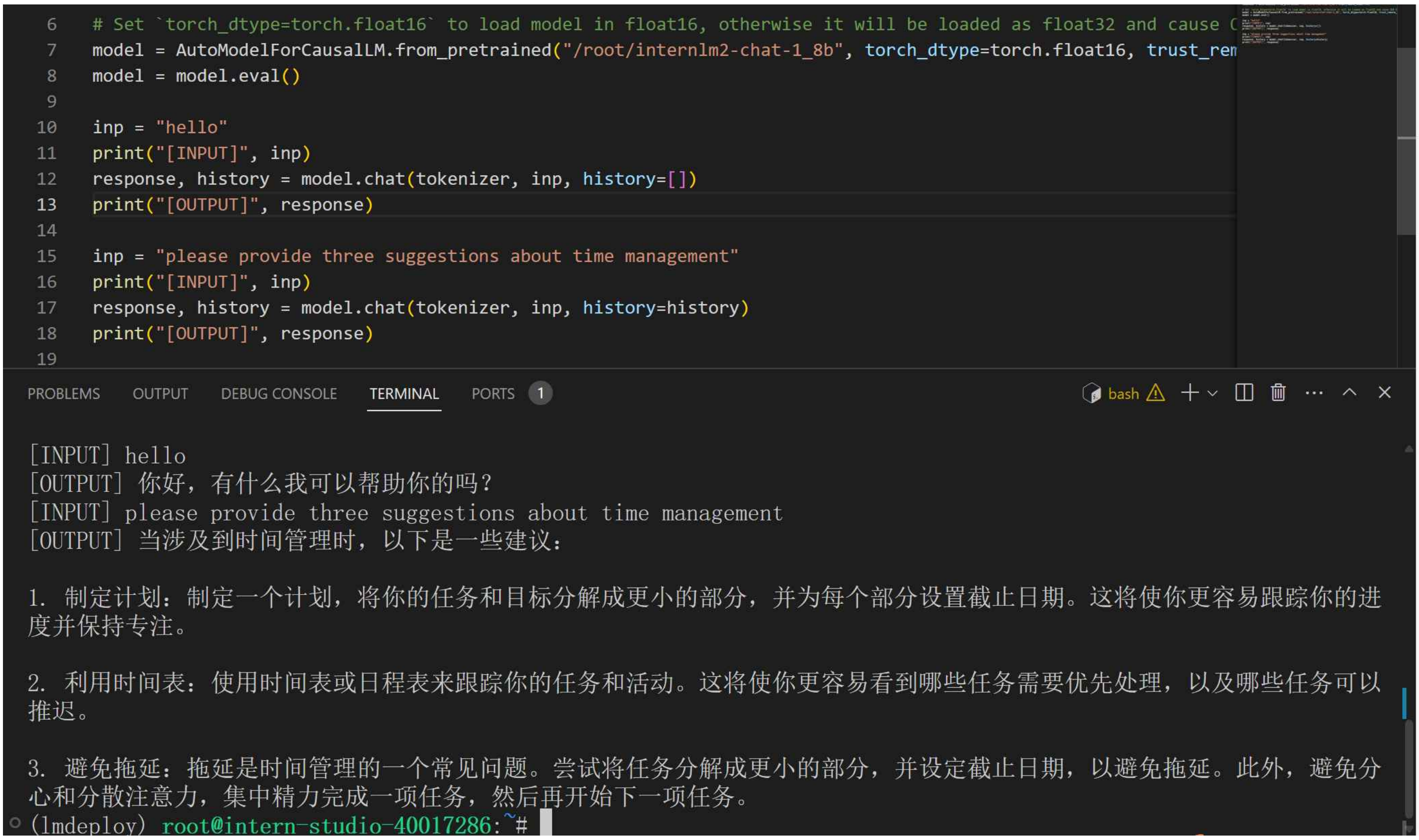
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

(lmdeploy) root@intern-studio-40017286:~# pip install einops

Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple

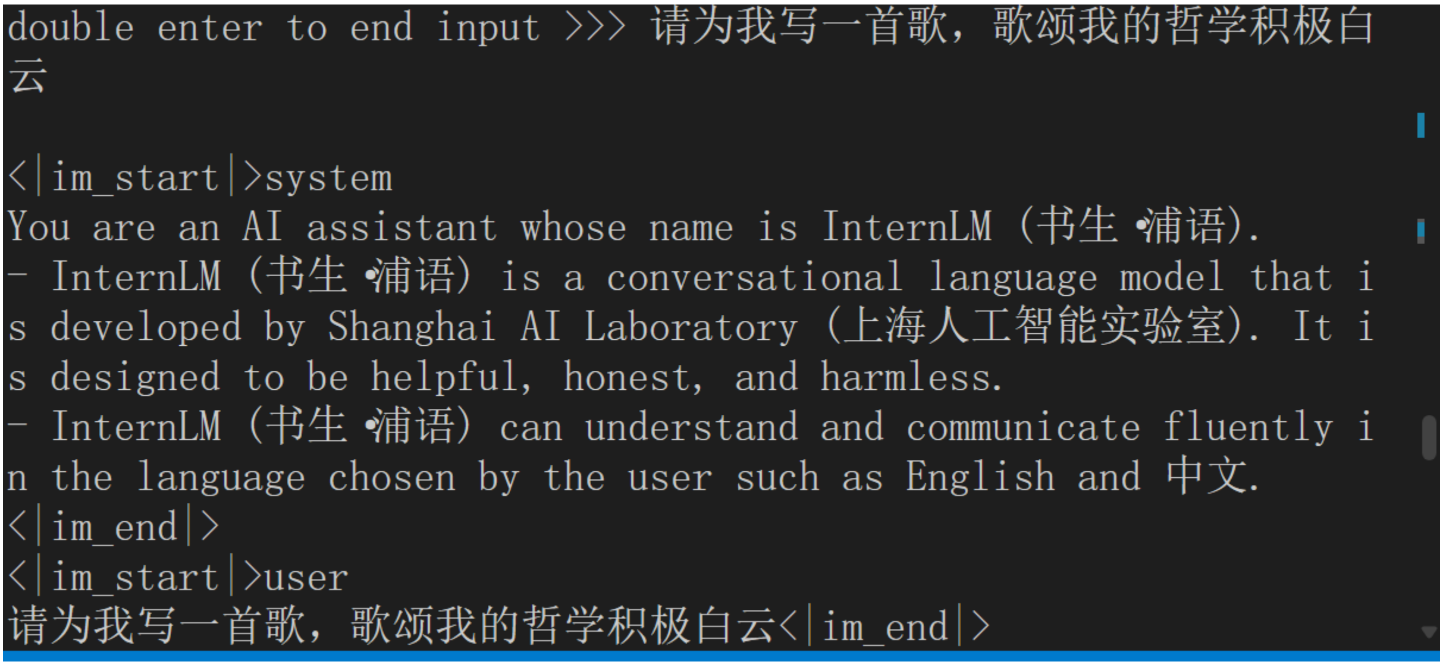
Requirement already satisfied: einops in ./.conda/lib/python3.11/site-packages (0.8.0)

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.pypa.io/warnings/venv

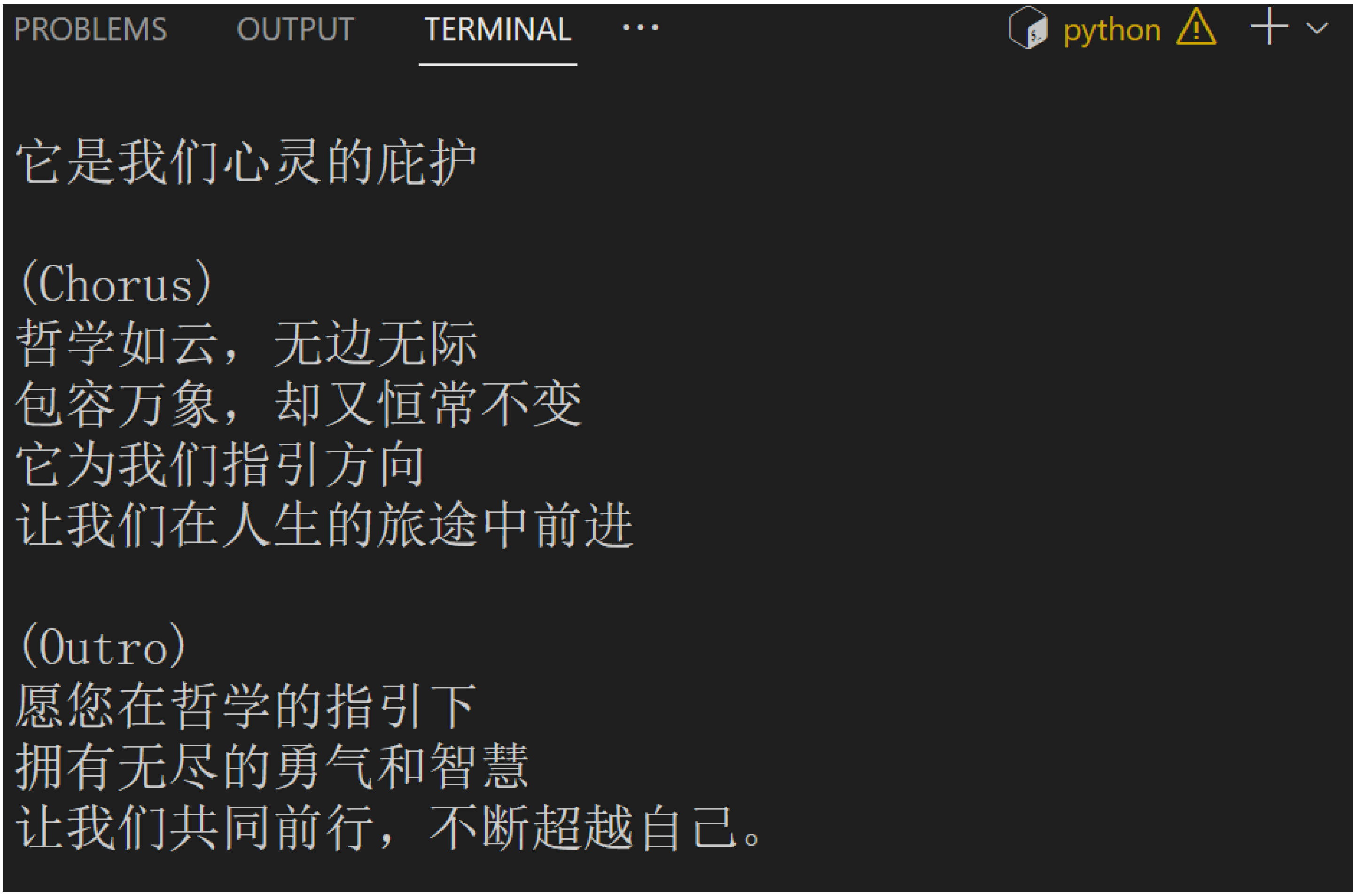


```python
# Set `torch_dtype=torch.float16` to load model in float16, otherwise it will be loaded as float32 and cause O
model = AutoModelForCausalLM.from_pretrained("/root/internlm2-chat-1_8b", torch_dtype=torch.float16, trust_rem
model = model.eval()

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)
```

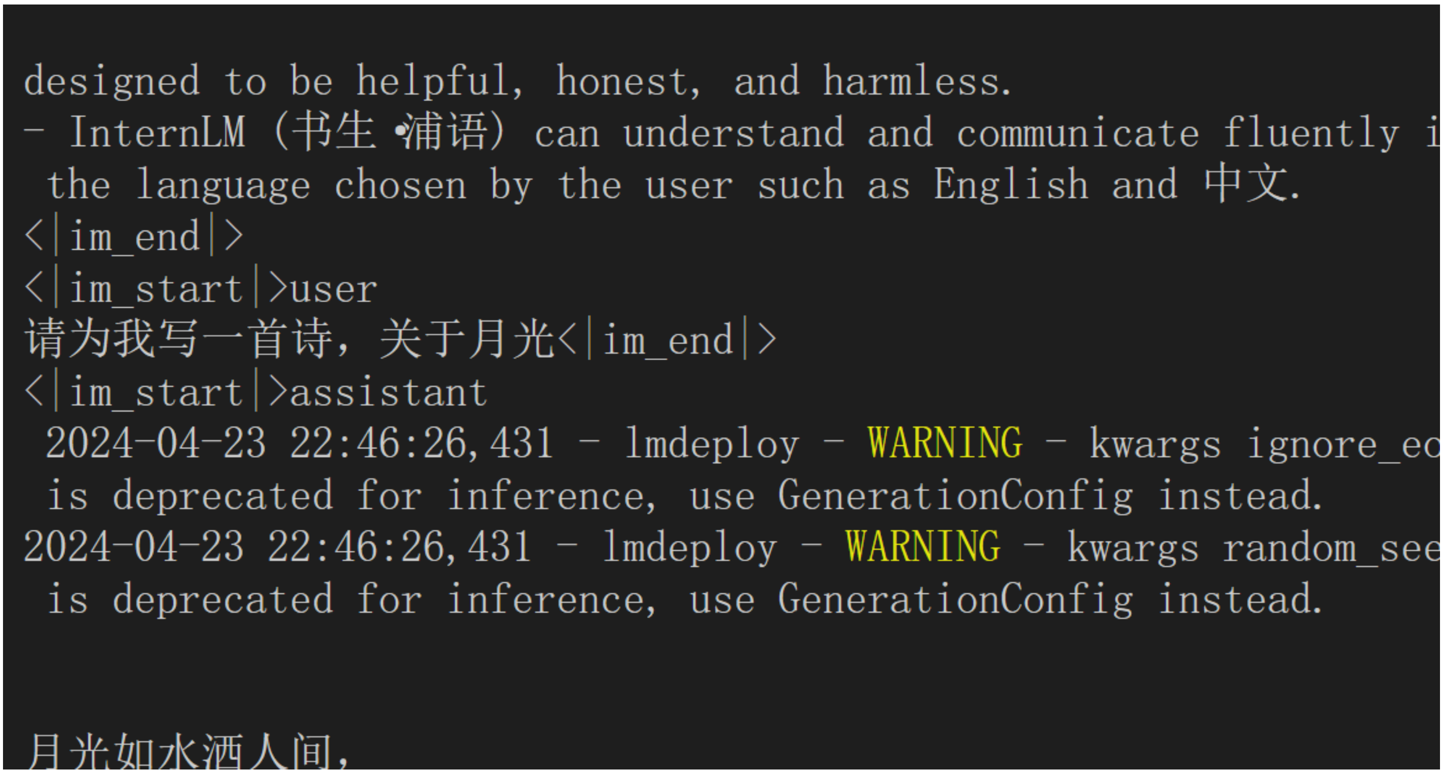
```
[INPUT] hello
[OUTPUT] 你好，有什么我可以帮助你的吗？
[INPUT] please provide three suggestions about time management
[OUTPUT] 当涉及到时间管理时，以下是一些建议：

1. 制定计划：制定一个计划，将你的任务和目标分解成更小的部分，并为每个部分设置截止日期。这将使你更容易跟踪你的进度并保持专注。

2. 利用时间表：使用时间表或日程表来跟踪你的任务和活动。这将使你更容易看到哪些任务需要优先处理，以及哪些任务可以推迟。

3. 避免拖延：拖延是时间管理的一个常见问题。尝试将任务分解成更小的部分，并设定截止日期，以避免拖延。此外，避免分心和分散注意力，集中精力完成一项任务，然后再开始下一项任务。
○ (lmdeploy) root@intern-studio-40017286:~#
```



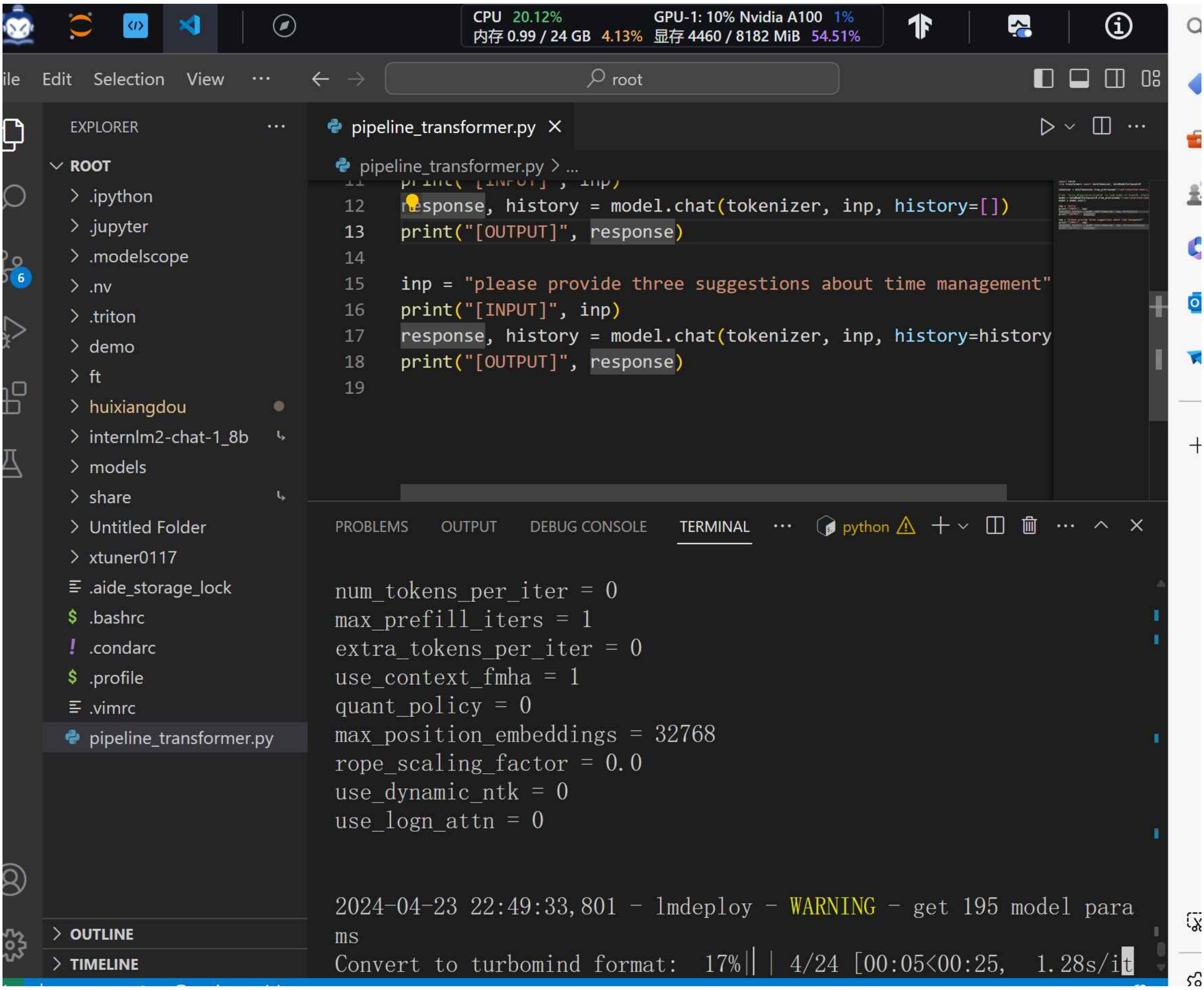
```
double enter to end input >>> 请为我写一首歌，歌颂我的哲学积极白
云

<|im_start|>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 i
s developed by Shanghai AI Laboratory (上海人工智能实验室). It i
s designed to be helpful, honest, and harmless.
- InternLM (书生·浦语) can understand and communicate fluently i
n the language chosen by the user such as English and 中文.
<|im_end|>
<|im_start|>user
请为我写一首歌，歌颂我的哲学积极白云<|im_end|>
```



```
它是我们心灵的庇护

(Chorus)
哲学如云，无边无际
包容万象，却又恒常不变
它为我们指引方向
让我们在人生的旅途中前进

(Outro)
愿您在哲学的指引下
拥有无尽的勇气和智慧
让我们共同前行，不断超越自己。
```

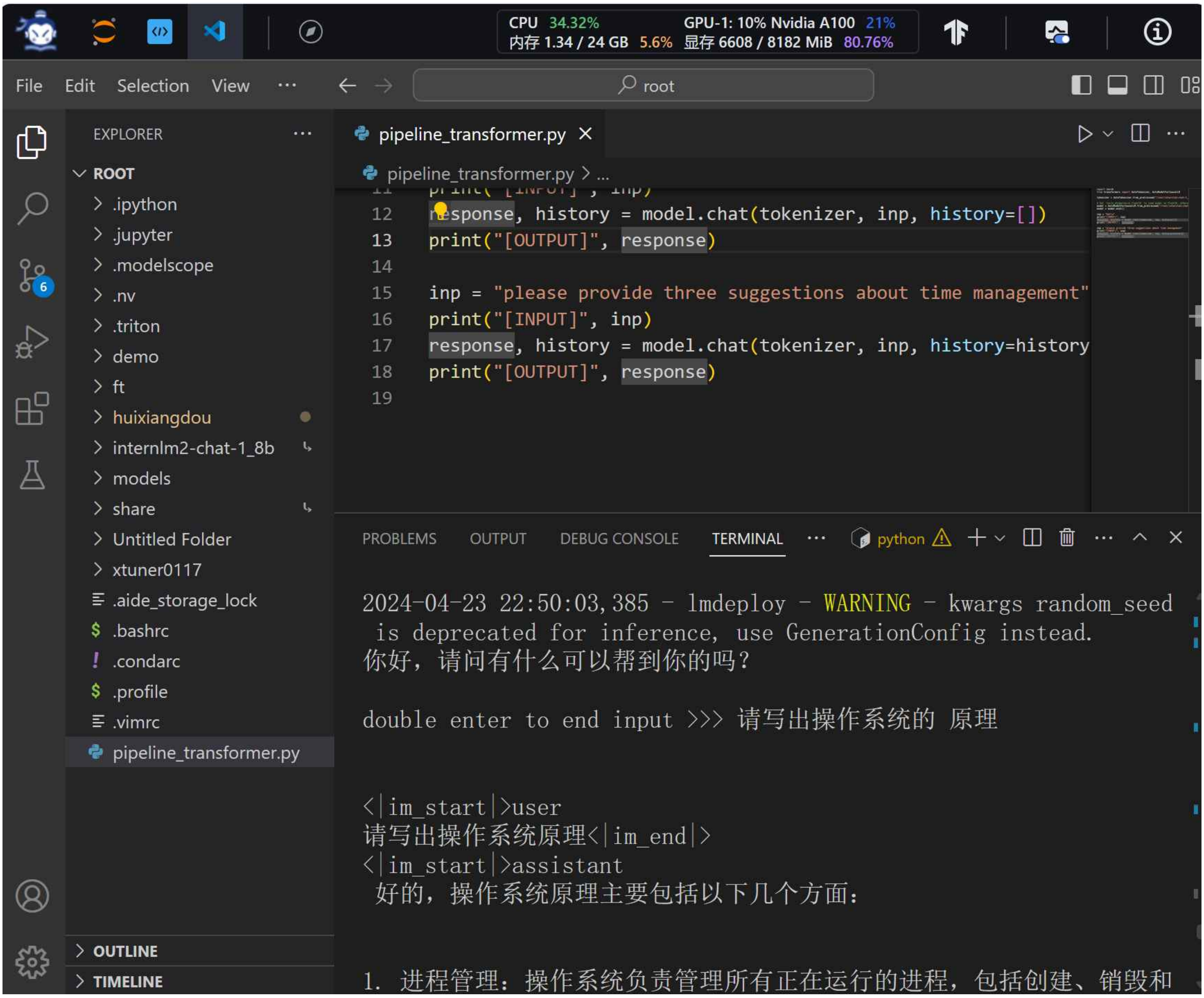


```
designed to be helpful, honest, and harmless.
- InternLM (书生·浦语) can understand and communicate fluently i
 the language chosen by the user such as English and 中文.
<|im_end|>
<|im_start|>user
请为我写一首诗，关于月光<|im_end|>
<|im_start|>assistant
 2024-04-23 22:46:26,431 - lmdeploy - WARNING - kwargs ignore_eo
 is deprecated for inference, use GenerationConfig instead.
2024-04-23 22:46:26,431 - lmdeploy - WARNING - kwargs random_see
 is deprecated for inference, use GenerationConfig instead.


月光如水洒人间，
```



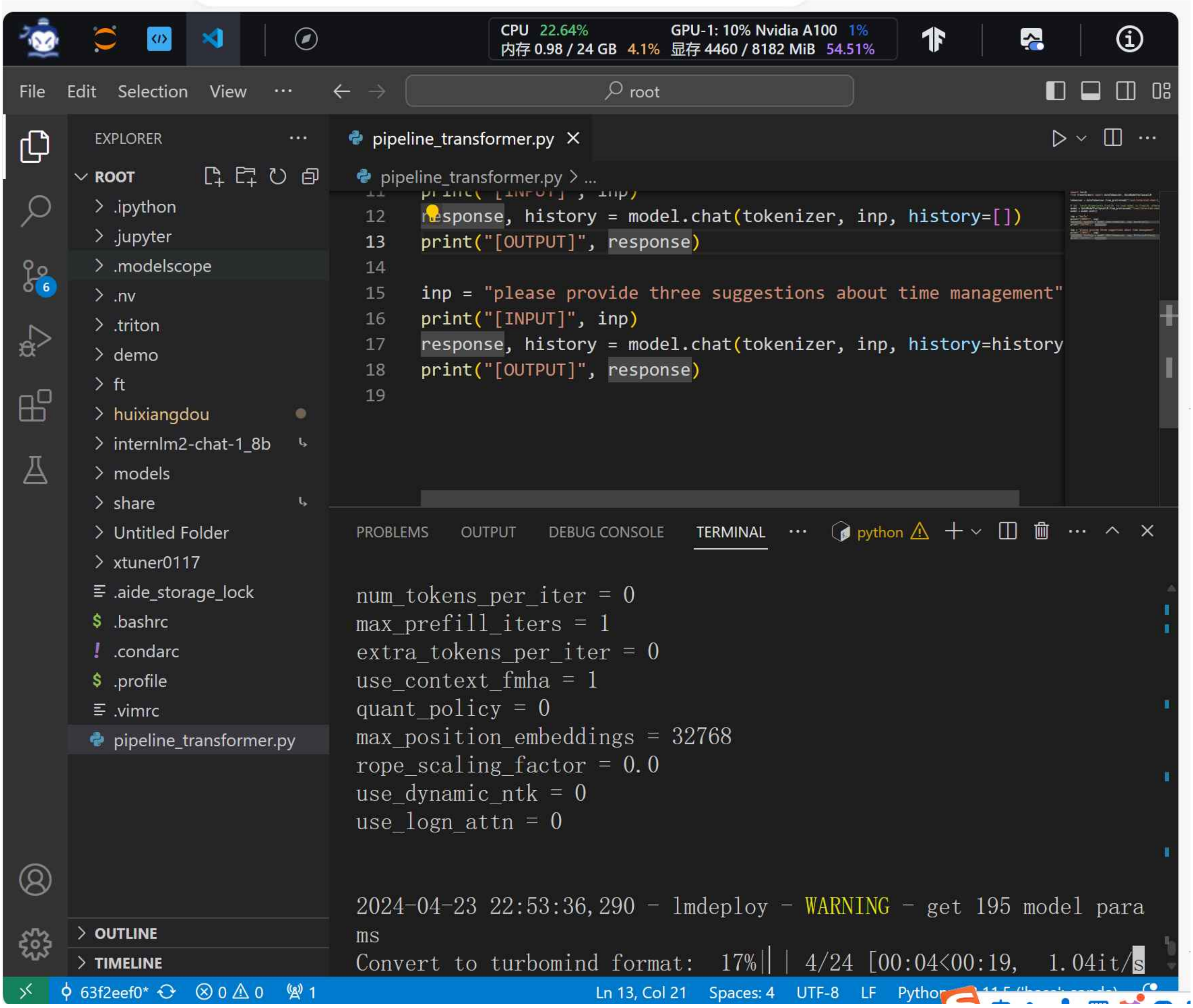
```python
print("[INPUT]", inp)
response, history = model.chat(tokenizer, inp, history=[])
print("[OUTPUT]", response)

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

```
num_tokens_per_iter = 0
max_prefill_iters = 1
extra_tokens_per_iter = 0
use_context_fmha = 1
quant_policy = 0
max_position_embeddings = 32768
rope_scaling_factor = 0.0
use_dynamic_ntk = 0
use_logn_attn = 0


2024-04-23 22:49:33,801 - lmdeploy - WARNING - get 195 model para
ms
Convert to turbomind format:  17%|█     | 4/24 [00:05<00:25,  1.28s/it
```



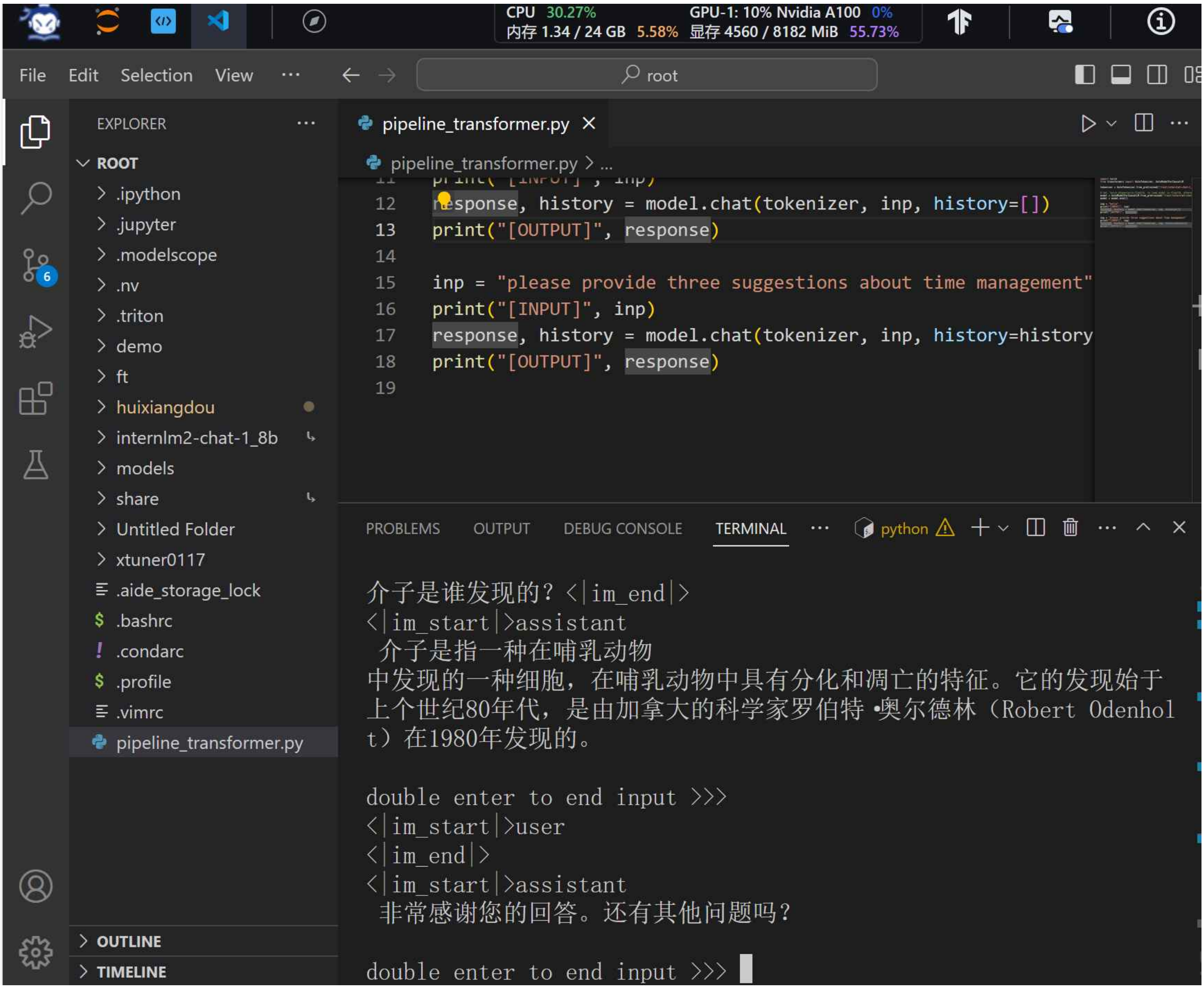
```python
print("[INPUT]", inp)
response, history = model.chat(tokenizer, inp, history=[])
print("[OUTPUT]", response)

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

```
2024-04-23 22:50:03,385 - lmdeploy - WARNING - kwargs random_seed
 is deprecated for inference, use GenerationConfig instead.
你好，请问有什么可以帮到你的吗？

double enter to end input >>> 请写出操作系统的 原理


<|im_start|>user
请写出操作系统原理<|im_end|>
<|im_start|>assistant
 好的，操作系统原理主要包括以下几个方面：


1. 进程管理：操作系统负责管理所有正在运行的进程，包括创建、销毁和
```



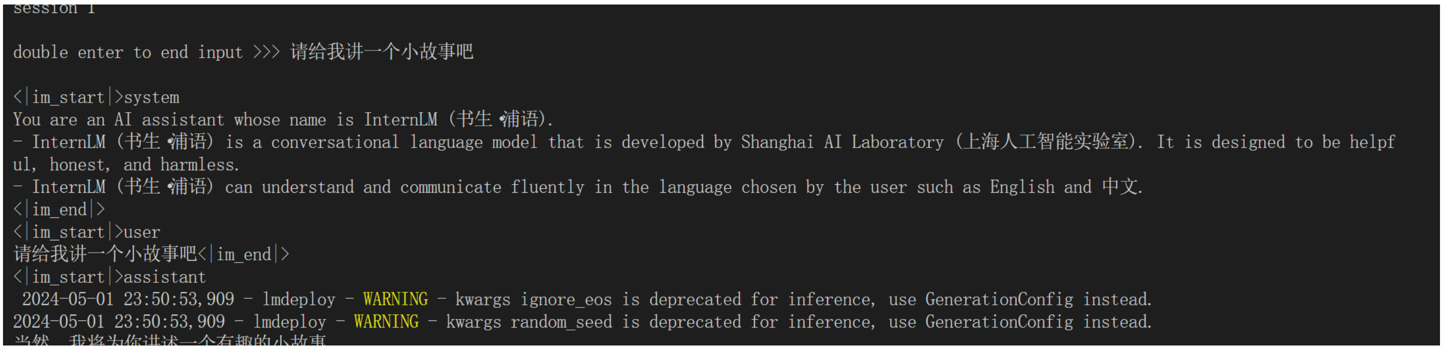
```python
print("[INPUT]", inp)
response, history = model.chat(tokenizer, inp, history=[])
print("[OUTPUT]", response)

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

```
num_tokens_per_iter = 0
max_prefill_iters = 1
extra_tokens_per_iter = 0
use_context_fmha = 1
quant_policy = 0
max_position_embeddings = 32768
rope_scaling_factor = 0.0
use_dynamic_ntk = 0
use_logn_attn = 0


2024-04-23 22:53:36,290 - lmdeploy - WARNING - get 195 model para
ms
Convert to turbomind format:  17%|█ | 4/24 [00:04<00:19,  1.04it/s
```



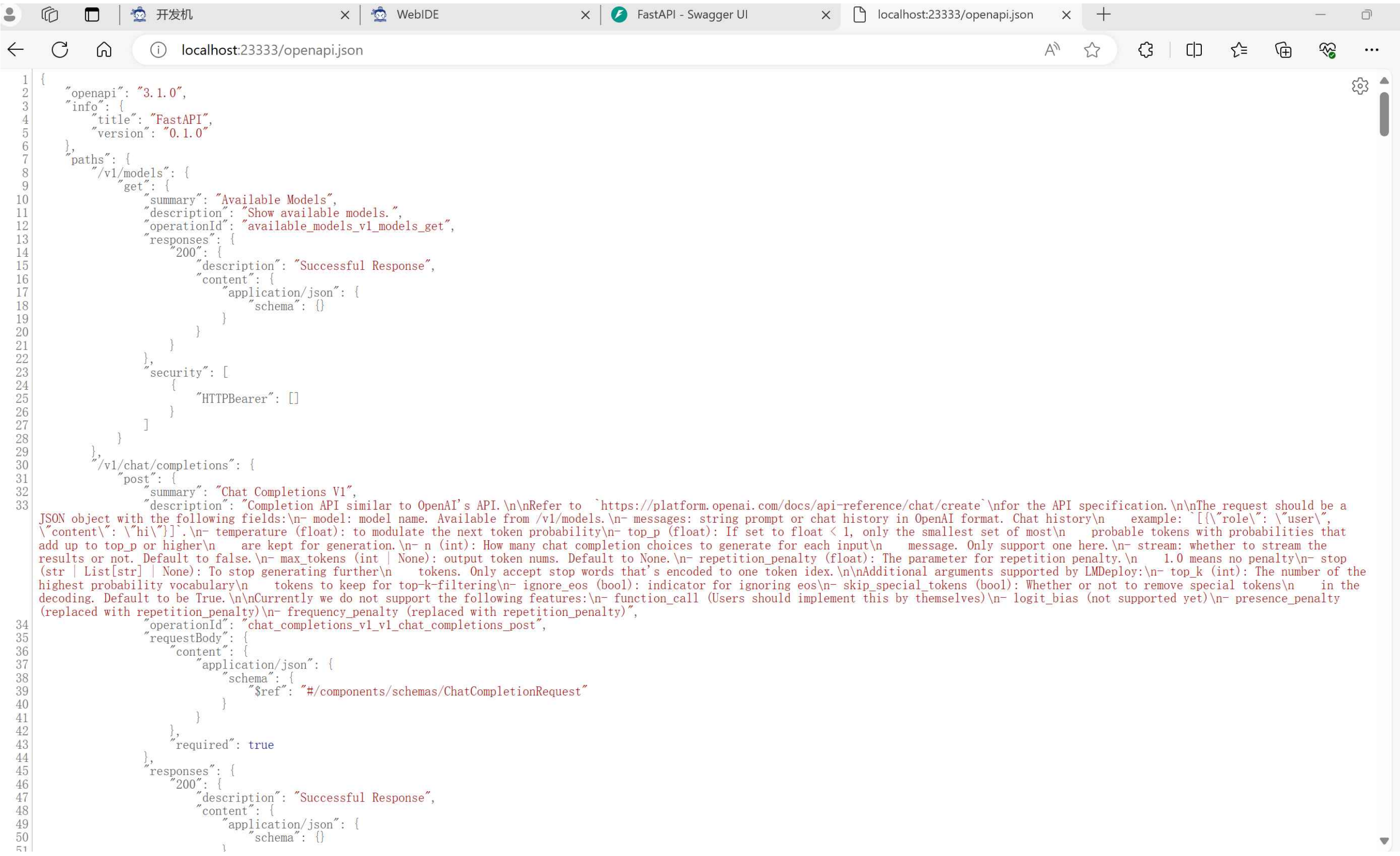
```python
print("[INPUT]", inp)
response, history = model.chat(tokenizer, inp, history=[])
print("[OUTPUT]", response)

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

```

```
介子是谁发现的？<|im_end|>
<|im_start|>assistant
 介子是指一种在哺乳动物
中发现的一种细胞，在哺乳动物中具有分化和凋亡的特征。它的发现始于
上个世纪80年代，是由加拿大的科学家罗伯特·奥尔德林（Robert Odenhol
t）在1980年发现的。

double enter to end input >>>
<|im_start|>user
<|im_end|>
<|im_start|>assistant
 非常感谢您的回答。还有其他问题吗？

double enter to end input >>>
```



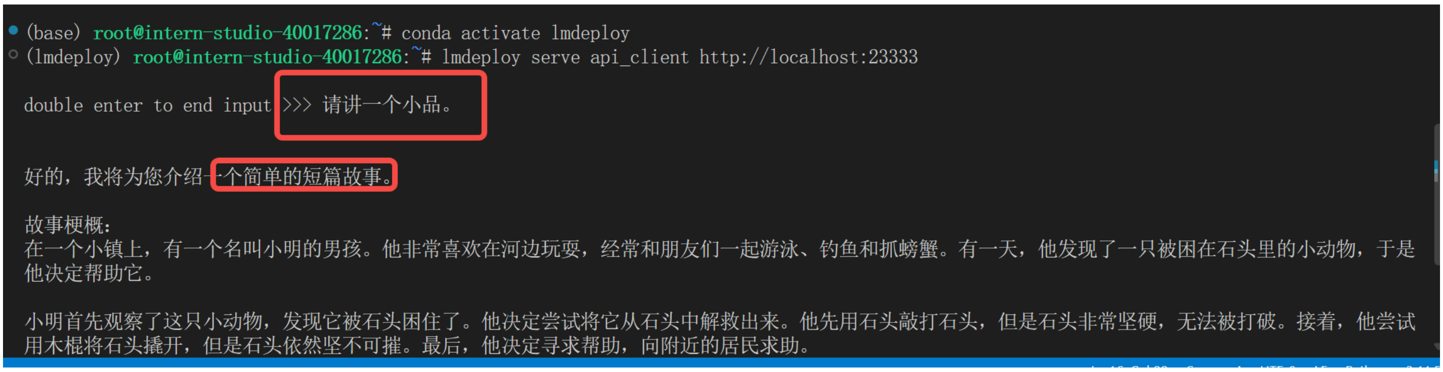
```
session 1

double enter to end input >>> 请给我讲一个小故事吧

<|im_start|>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 helpf
ul, 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
请给我讲一个小故事吧<|im_end|>
<|im_start|>assistant
 2024-05-01 23:50:53,909 - lmdeploy - WARNING - kwargs ignore_eos is deprecated for inference, use GenerationConfig instead.
2024-05-01 23:50:53,909 - lmdeploy - WARNING - kwargs random_seed is deprecated for inference, use GenerationConfig instead.
当然，我将为你讲述一个有趣的小故事
```

前端有问题，通过：http://localhost:23333/openapi.json 可以获得如下反馈。



```
{
   "openapi": "3.1.0",
   "info": {
       "title": "FastAPI",
       "version": "0.1.0"
   },
   "paths": {
       "/v1/models": {
           "get": {
               "summary": "Available Models",
               "description": "Show available models.",
               "operationId": "available_models_v1_models_get",
               "responses": {
                   "200": {
                       "description": "Successful Response",
                       "content": {
                           "application/json": {
                               "schema": {}
                           }
                       }
                   }
               },
               "security": [
                   {
                       "HTTPBearer": []
                   }
               ]
           }
       },
       "/v1/chat/completions": {
           "post": {
               "summary": "Chat Completions V1",
               "description": "Completion API similar to OpenAI's API.\n\nRefer to  `https://platform.openai.com/docs/api-reference/chat/create`\nfor the API specification.\n\nThe request should be a JSON object with the following fields:\n- model: model name. Available from /v1/models.\n- messages: string prompt or chat history in OpenAI format. Chat history\n    example: `[{\"role\": \"user\", \"content\": \"hi\"}]`.\n- temperature (float): to modulate the next token probability\n- top_p (float): If set to float < 1, only the smallest set of most\n    probable tokens with probabilities that add up to top_p or higher\n    are kept for generation.\n- n (int): How many chat completion choices to generate for each input\n    message. Only support one here.\n- stream: whether to stream the results or not. Default to false.\n- max_tokens (int | None): output token nums. Default to None.\n- repetition_penalty (float): The parameter for repetition penalty.\n    1.0 means no penalty\n- stop (str | List[str] | None): To stop generating further\n    tokens. Only accept stop words that's encoded to one token idex.\n\nAdditional arguments supported by LMDeploy:\n- top_k (int): The number of the highest probability vocabulary\n    tokens to keep for top-k-filtering\n- ignore_eos (bool): indicator for ignoring eos\n- skip_special_tokens (bool): Whether or not to remove special tokens\n    in the decoding. Default to be True.\n\nCurrently we do not support the following features:\n- function_call (Users should implement this by themselves)\n- logit_bias (not supported yet)\n- presence_penalty (replaced with repetition_penalty)\n- frequency_penalty (replaced with repetition_penalty)",
               "operationId": "chat_completions_v1_v1_chat_completions_post",
               "requestBody": {
                   "content": {
                       "application/json": {
                           "schema": {
                               "$ref": "#/components/schemas/ChatCompletionRequest"
                           }
                       }
                   },
                   "required": true
               },
               "responses": {
                   "200": {
                       "description": "Successful Response",
                       "content": {
                           "application/json": {
                               "schema": {}
                           }
```

要求模型讲一个小品，但是模型还是说讲一个故事。



```
(base) root@intern-studio-40017286:~# conda activate lmdeploy
(lmdeploy) root@intern-studio-40017286:~# lmdeploy serve api_client http://localhost:23333

double enter to end input >>> 请讲一个小品。

好的，我将为您介绍一个简单的短篇故事。

故事梗概：
在一个小镇上，有一个名叫小明的男孩。他非常喜欢在河边玩耍，经常和朋友们一起游泳、钓鱼和抓螃蟹。有一天，他发现了一只被困在石头里的小动物，于是
他决定帮助它。

小明首先观察了这只小动物，发现它被石头困住了。他决定尝试将它从石头中解救出来。他先用石头敲打石头，但是石头非常坚硬，无法被打破。接着，他尝试
用木棍将石头撬开，但是石头依然坚不可摧。最后，他决定寻求帮助，向附近的居民求助。
```