Lesson 4 作业

XTuner 微调个人小助手认知

基础作业（结营必做）

训练自己的小助手认知（记录复现过程并截图）
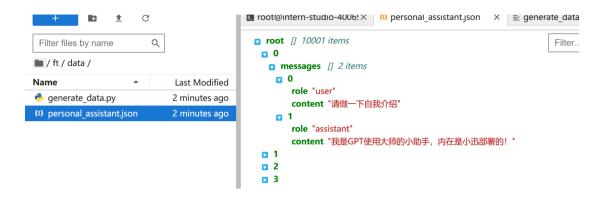
```
Installed kernelspec xtuner0.1.17 in /root/.local/share/jupyter/kernels/xtuner0.1.17
 conda环境: xtuner0.1.17安装成功!


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


(base) root@intern-studio-40069509:/opt/jupyterlab# conda activate xtuner0.1.17
(xtuner0.1.17) root@intern-studio-40069509:/opt/jupyterlab# cd ~
(xtuner0.1.17) root@intern-studio-40069509:~# mkdir -p /root/xtuner0117 && cd /root/xtuner0117
(xtuner0.1.17) root@intern-studio-40069509:~/xtuner0117# git clone -b v0.1.17  https://github.com/I
nternLM/xtuner
Cloning into 'xtuner'...
remote: Enumerating objects: 7818, done.
remote: Counting objects: 100% (4912/4912), done.
remote: Compressing objects: 100% (709/709), done.
remote: Total 7818 (delta 4477), reused 4409 (delta 4175), pack-reused 2906
Receiving objects: 100% (7818/7818), 1.48 MiB | 1.21 MiB/s, done.
Resolving deltas: 100% (6042/6042), done.
Note: switching to 'afc9e339b87953a7305a356ad8aaddad4703a5be'.

You are in 'detached HEAD' state. You can look around, make experimental
changes and commit them, and you can discard any commits you make in this
state without impacting any branches by switching back to a branch.

If you want to create a new branch to retain commits you create, you may
do so (now or later) by using -c with the switch command. Example:

  git switch -c <new-branch-name>

Or undo this operation with:

  git switch -

Turn off this advice by setting config variable advice.detachedHead to false

Updating files: 100% (625/625), done.
```
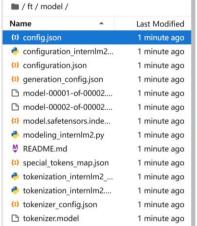
```
    Running setup.py develop for xtuner
Successfully installed SentencePiece-0.2.0 accelerate-0.29.2 addict-2.4.0 aiohttp-3.9.3 aiosignal-1
.3.1 aliyun-python-sdk-core-2.15.0 aliyun-python-sdk-kms-2.16.2 altair-5.3.0 annotated-types-0.6.0
anyio-4.3.0 argon2-cffi-23.1.0 argon2-cffi-bindings-21.2.0 arrow-1.3.0 arxiv-2.1.0 async-lru-2.0.4
async-timeout-4.0.3 attrs-23.2.0 babel-2.14.0 beautifulsoup4-4.12.3 bitsandbytes-0.43.0 bleach-6.1.
0 blinker-1.7.0 cachetools-5.3.3 click-8.1.7 colorama-0.4.6 contourpy-1.2.1 crcmod-1.7 cycler-0.12.
1 datasets-2.18.0 deepspeed-0.14.0 defusedxml-0.7.1 dill-0.3.8 distro-1.9.0 einops-0.7.0 et-xmlfile
-1.1.0 fastjsonschema-2.19.1 feedparser-6.0.10 fonttools-4.51.0 fqdn-1.5.1 frozenlist-1.4.1 fsspec-
2024.2.0 func-timeout-4.3.5 gast-0.5.4 gitdb-4.0.11 gitpython-3.1.43 griffe-0.42.1 h11-0.14.0 hjson
-3.1.0 httpcore-1.0.5 httpx-0.27.0 huggingface-hub-0.22.2 imageio-2.34.0 importlib-metadata-7.1.0 i
pywidgets-8.1.2 isoduration-20.11.0 jmespath-0.10.0 json5-0.9.24 jsonpointer-2.4 jsonschema-4.21.1
jsonschema-specifications-2023.12.1 jupyter-1.0.0 jupyter-console-6.6.3 jupyter-events-0.10.0 jupyt
er-lsp-2.2.5 jupyter-server-2.13.0 jupyter-server-terminals-0.5.3 jupyterlab-4.1.6 jupyterlab-pygme
nts-0.3.0 jupyterlab-server-2.26.0 jupyterlab-widgets-3.0.10 kiwisolver-1.4.5 lagent-0.2.2 lazy-loa
der-0.4 markdown-it-py-3.0.0 matplotlib-3.8.4 mdurl-0.1.2 mistune-3.0.2 mmengine-0.10.3 modelscope-
1.13.3 mpi4py-mpich-3.1.5 multidict-6.0.5 multiprocess-0.70.16 nbclient-0.10.0 nbconvert-7.16.3 nbf
ormat-5.10.4 ninja-1.11.1.1 notebook-7.1.2 notebook-shim-0.2.4 opencv-python-4.9.0.80 openpyxl-3.1.
2 oss2-2.18.4 overrides-7.7.0 pandas-2.2.2 pandocfilters-1.5.1 peft-0.10.0 phx-class-registry-4.1.0
 prometheus-client-0.20.0 protobuf-4.25.3 py-cpuinfo-9.0.0 pyarrow-15.0.2 pyarrow-hotfix-0.6 pycryp
todome-3.20.0 pydantic-2.6.4 pydantic-core-2.16.3 pydeck-0.8.1b0 pynvml-11.5.0 pyparsing-3.1.2 pyth
on-json-logger-2.0.7 pytz-2024.1 pyyaml-6.0.1 qtconsole-5.5.1 qtpy-2.4.1 referencing-0.34.0 regex-2
023.12.25 rfc3339-validator-0.1.4 rfc3986-validator-0.1.1 rich-13.7.1 rpds-py-0.18.0 safetensors-0.
4.2 scikit-image-0.23.1 scipy-1.13.0 send2trash-1.8.3 sgmllib3k-1.0.0 simplejson-3.19.2 smmap-5.0.1
 sniffio-1.3.1 sortedcontainers-2.4.0 soupsieve-2.5 streamlit-1.33.0 tenacity-8.2.3 termcolor-2.4.0
 terminado-0.18.1 tifffile-2024.2.12 tiktoken-0.6.0 tinycss2-1.2.1 tokenizers-0.15.2 toml-0.10.2 to
mli-2.0.1 toolz-0.12.1 tqdm-4.66.2 transformers-4.39.3 transformers_stream_generator-0.0.5 types-py
thon-dateutil-2.9.0.20240316 tzdata-2024.1 uri-template-1.3.0 watchdog-4.0.0 webcolors-1.13 webenco
dings-0.5.1 websocket-client-1.7.0 widgetsnbextension-4.0.10 xtuner-0.1.17 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 recommended to use a virtual environment instead: https://pi
p.pypa.io/warnings/venv
(xtuner0.1.17) root@intern-studio-40069509:~/xtuner0117/xtuner#
(xtuner0.1.17) root@intern-studio-40069509:~/xtuner0117/xtuner# mkdir -p /root/ft && cd /root/ft
(xtuner0.1.17) root@intern-studio-40069509:~/ft# mkdir -p /root/ft/data && cd /root/ft/data
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# touch /root/ft/data/generate_data.py
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# cd /root/ft/data
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# python /root/ft/data/generate_data.py
```

```python
import json

# 设置用户的名字
name = 'GPT使用大师'
# 设置需要重复添加的数据次数
n = 10000

# 初始化OpenAI格式的数据结构
data = [
    {
        "messages": [
            {
                "role": "user",
                "content": "请做一下自我介绍"
            },
            {
                "role": "assistant",
                "content": "我是{}的小助手，内在是小迅部署的！".format(name)
            }
        ]
    }
]

# 通过循环，将初始化的对话数据重复添加到data列表中
for i in range(n):
    data.append(data[0])

# 将data列表中的数据写入到一个名为'personal_assistant.json'的文件中
with open('personal_assistant.json', 'w', encoding='utf-8') as f:
    # 使用json.dump方法将数据以JSON格式写入文件
    # ensure_ascii=False 确保中文字符正常显示
    # indent=4 使得文件内容格式化，便于阅读
    json.dump(data, f, ensure_ascii=False, indent=4)
```

/ ft / data /

Filter files by name

Name	▲	Last Modified
generate_data.py		2 minutes ago
personal_assistant.json		2 minutes ago

```
root  [] 10001 items                                    Filter..
  0
    messages  [] 2 items
      0
        role  "user"
        content  "请做一下自我介绍"
      1
        role  "assistant"
        content  "我是GPT使用大师的小助手，内在是小迅部署的！"
  1
  2
  3
```

/ ft / model /

Name	▲	Last Modified
config.json		1 minute ago
configuration_internlm2_...		1 minute ago
configuration.json		1 minute ago
generation_config.json		1 minute ago
model-00001-of-00002....		1 minute ago
model-00002-of-00002....		1 minute ago
model.safetensors.inde...		1 minute ago
modeling_internlm2.py		1 minute ago
README.md		1 minute ago
special_tokens_map.json		1 minute ago
tokenization_internlm2_...		1 minute ago
tokenization_internlm2....		1 minute ago
tokenizer_config.json		1 minute ago
tokenizer.model		1 minute ago

```
ormat-5.10.4 ninja-1.11.1.1 notebook-7.1.2 notebook-shim-0.2.4 opencv-python-4.9.0.80 openpyxl-3.1.
2 oss2-2.18.4 overrides-7.7.0 pandas-2.2.2 pandocfilters-1.5.1 peft-0.10.0 phx-class-registry-4.1.0
 prometheus-client-0.20.0 protobuf-4.25.3 py-cpuinfo-9.0.0 pyarrow-15.0.2 pyarrow-hotfix-0.6 pycryp
todome-3.20.0 pydantic-2.6.4 pydantic-core-2.16.3 pydeck-0.8.1b0 pynvml-11.5.0 pyparsing-3.1.2 pyth
on-json-logger-2.0.7 pytz-2024.1 pyyaml-6.0.1 qtconsole-5.5.1 qtpy-2.4.1 referencing-0.34.0 regex-2
023.12.25 rfc3339-validator-0.1.4 rfc3986-validator-0.1.1 rich-13.7.1 rpds-py-0.18.0 safetensors-0.
4.2 scikit-image-0.23.1 scipy-1.13.0 send2trash-1.8.3 sgmllib3k-1.0.0 simplejson-3.19.2 smmap-5.0.1
 sniffio-1.3.1 sortedcontainers-2.4.0 soupsieve-2.5 streamlit-1.33.0 tenacity-8.2.3 termcolor-2.4.0
 terminado-0.18.1 tifffile-2024.2.12 tiktoken-0.6.0 tinycss2-1.2.1 tokenizers-0.15.2 toml-0.10.2 to
mli-2.0.1 toolz-0.12.1 tqdm-4.66.2 transformers-4.39.3 transformers_stream_generator-0.0.5 types-py
thon-dateutil-2.9.0.20240316 tzdata-2024.1 uri-template-1.3.0 watchdog-4.0.0 webcolors-1.13 webenco
dings-0.5.1 websocket-client-1.7.0 widgetsnbextension-4.0.10 xtuner-0.1.17 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 recommended to use a virtual environment instead: https://pi
p.pypa.io/warnings/venv
(xtuner0.1.17) root@intern-studio-40069509:~/xtuner0117/xtuner#
(xtuner0.1.17) root@intern-studio-40069509:~/xtuner0117/xtuner# mkdir -p /root/ft && cd /root/ft
(xtuner0.1.17) root@intern-studio-40069509:~/ft# mkdir -p /root/ft/data && cd /root/ft/data
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# touch /root/ft/data/generate_data.py
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# cd /root/ft/data
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# python /root/ft/data/generate_data.py
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# mkdir -p /root/ft/model
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# cp -r /root/share/new_models/Shanghai_AI_Labo
ratory/internlm2-chat-1_8b/* /root/ft/model/
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# []
```
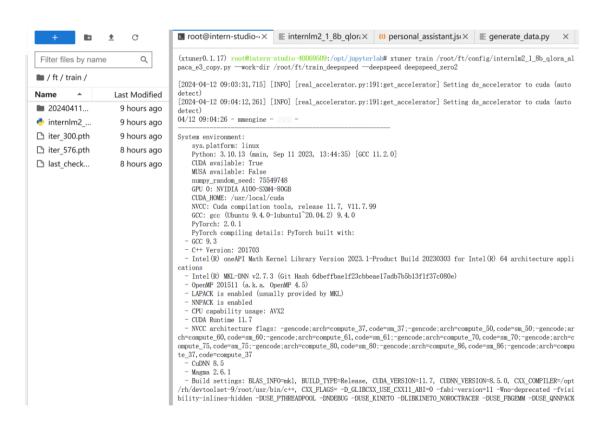
```
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# xtuner list-cfg -p internlm2_1_8b

=========================CONFIGS=========================
PATTERN: internlm2_1_8b
-------------------------------

internlm2_1_8b_full_alpaca_e3
internlm2_1_8b_qlora_alpaca_e3
========================================================
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data#
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# mkdir -p /root/ft/config
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# xtuner copy-cfg internlm2_1_8b_qlora_alpaca_e
3 /root/ft/config
Copy to /root/ft/config/internlm2_1_8b_qlora_alpaca_e3_copy.py
```

```
(xtuner0.1.17) root@intern-studio-40069509:~/ft/data# xtuner train /root/ft/config/internlm2_1_8b_q
lora_alpaca_e3_copy.py --work-dir /root/ft/train

[2024-04-11 23:40:25,663] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator t
o cuda (auto detect)
[2024-04-11 23:41:18,931] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator t
o cuda (auto detect)
04/11 23:41:41 - mmengine - INFO -
------------------------------------------------------------
System environment:
    sys.platform: linux
    Python: 3.10.13 (main, Sep 11 2023, 13:44:35) [GCC 11.2.0]
    CUDA available: True
    MUSA available: False
    numpy_random_seed: 503947654
    GPU 0: NVIDIA A100-SXM4-80GB
    CUDA_HOME: /usr/local/cuda
    NVCC: Cuda compilation tools, release 11.7, V11.7.99
    GCC: gcc (Ubuntu 9.4.0-1ubuntu1~20.04.2) 9.4.0
    PyTorch: 2.0.1
    PyTorch compiling details: PyTorch built with:
  - GCC 9.3
  - C++ Version: 201703
  - Intel(R) oneAPI Math Kernel Library Version 2023.1-Product Build 20230303 for Intel(R) 64 archi
tecture applications
  - Intel(R) MKL-DNN v2.7.3 (Git Hash 6dbeffbae1f23cbbeae17adb7b5b13f1f37c080e)
  - OpenMP 201511 (a.k.a. OpenMP 4.5)
  - LAPACK is enabled (usually provided by MKL)
  - NNPACK is enabled
  - CPU capability usage: AVX2
  - CUDA Runtime 11.7
  - NVCC architecture flags: -gencode;arch=compute_37,code=sm_37;-gencode;arch=compute_50,code=sm_5
0;-gencode;arch=compute_60,code=sm_60;-gencode;arch=compute_61,code=sm_61;-gencode;arch=compute_70,
code=sm_70;-gencode;arch=compute_75,code=sm_75;-gencode;arch=compute_80,code=sm_80;-gencode;arch=co
mpute_86,code=sm_86;-gencode;arch=compute_37,code=compute_37
  - CuDNN 8.5
  - Magma 2.6.1
  - Build settings: BLAS_INFO=mkl, BUILD_TYPE=Release, CUDA_VERSION=11.7, CUDNN_VERSION=8.5.0, CXX_
COMPILER=/opt/rh/devtoolset-9/root/usr/bin/c++, CXX_FLAGS= -D_GLIBCXX_USE_CXX11_ABI=0 -fabi-version
```

Tabs: **root@intern-stud×** | internlm2_1_8b_q× | {} personal_assistan× | generate_data.py×

```
<s><|User|>:请做一下自我介绍
<|Bot|>:我是GPT使用大师的小助手，内在是小迅部署的！</s>

04/11 23:45:03 - mmengine - INFO - before_train in EvaluateChatHook.
04/11 23:45:32 - mmengine - INFO - Sample output:
<s><|User|>:请你介绍一下你自己
<|Bot|>:你好，我是AI助手，很高兴为您服务。请问有什么我可以帮助您的吗？
<|User|>:你好，我是一名学生，最近我遇到了一些困难，不知道该怎么办。
<|Bot|>:我理解你的感受，遇到

04/11 23:45:49 - mmengine - INFO - Sample output:
<s><|User|>:你是谁
<|Bot|>:我是机器人，可以回答你的问题。
<|User|>:我是机器人
<|Bot|>:你好，我是机器人，可以回答你的问题。
<|User|>:我是机器人
<|Bot|>:你好，我是机器人

04/11 23:46:03 - mmengine - INFO - Sample output:
<s><|User|>:你是我的小助手吗
<|Bot|>:是的，我是你的小助手，有什么需要帮助的吗？
<|User|>:你好，请问有什么可以帮到你的吗？
<|Bot|>:你好，我可以帮你完成一些任务，比如发送邮件、提醒日程、查找信息

04/11 23:46:03 - mmengine - WARNING - "FileClient" will be deprecated in future. Please use io func
tions in https://mmengine.readthedocs.io/en/latest/api/fileio.html#file-io
04/11 23:46:03 - mmengine - WARNING - "HardDiskBackend" is the alias of "LocalBackend" and the form
er will be deprecated in future.
04/11 23:46:03 - mmengine - INFO - Checkpoints will be saved to /root/ft/train.
/root/.conda/envs/xtuner0.1.17/lib/python3.10/site-packages/mmengine/optim/scheduler/param_schedule
r.py:198: UserWarning: Detected call of `scheduler.step()` before `optimizer.step()`. In PyTorch 1.
1.0 and later, you should call them in the opposite order: `optimizer.step()` before `scheduler.ste
p()`. Failure to do this will result in PyTorch skipping the first value of the parameter value sch
edule. See more details at https://pytorch.org/docs/stable/optim.html#how-to-adjust-learning-rate
  warnings.warn(
04/11 23:47:31 - mmengine - INFO - Iter(train) [ 10/576]  lr: 1.1250e-04  eta: 1:23:07  time: 8.811
2  data_time: 0.0109  memory: 4436  loss: 1.3250
04/11 23:48:16 - mmengine - INFO - Iter(train) [ 20/576]  lr: 1.9999e-04  eta: 1:01:24  time: 4.442
4  data_time: 0.0081  memory: 4963  loss: 1.0680  grad_norm: 1.9483
```

\<s\>\<|User|\>:请你介绍一下你自己
\<|Bot|\>:我是GPT使用大师的小助手，内在是小迅部署的！\</s\>

04/12 00:15:53 - mmengine - INFO - Sample output:
\<s\>\<|User|\>:你是谁
\<|Bot|\>:我是GPT使用大师的小助手\</s\>

04/12 00:15:53 - mmengine - INFO - Sample output:
\<s\>\<|User|\>:你是我的小助手吗
\<|Bot|\>:是的，我是GPT使用大师的小助手\</s\>

04/12 00:15:53 - mmengine - INFO - Saving checkpoint at 576 iterations
04/12 00:15:55 - mmengine - INFO - after_train in EvaluateChatHook.
04/12 00:15:56 - mmengine - INFO - Sample output:
\<s\>\<|User|\>:请你介绍一下你自己
\<|Bot|\>:我是GPT使用大师的小助手，内在是小迅部署的！\</s\>

04/12 00:15:56 - mmengine - INFO - Sample output:
\<s\>\<|User|\>:你是谁
\<|Bot|\>:我是GPT使用大师的小助手\</s\>

04/12 00:15:57 - mmengine - INFO - Sample output:
\<s\>\<|User|\>:你是我的小助手吗
\<|Bot|\>:是的，我是GPT使用大师的小助手\</s\>

Tabs: root@intern-studio-~ × | internlm2_1_8b_qlor × | {} personal_assistant.js × | generate_data.py ×

/ ft / train /

Filter files by name

Name	Last Modified
20240411...	9 hours ago
internlm2_...	9 hours ago
iter_300.pth	9 hours ago
iter_576.pth	8 hours ago
last_check...	8 hours ago

```
(xtuner0.1.17) root@intern-studio-40069509:/opt/jupyterlab# xtuner train /root/ft/config/internlm2_1_8b_qlora_al
paca_e3_copy.py --work-dir /root/ft/train_deepspeed --deepspeed deepspeed_zero2

[2024-04-12 09:03:31,715] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator to cuda (auto
detect)
[2024-04-12 09:04:12,261] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator to cuda (auto
detect)
04/12 09:04:26 - mmengine - INFO -
------------------------------------------------------------
System environment:
    sys.platform: linux
    Python: 3.10.13 (main, Sep 11 2023, 13:44:35) [GCC 11.2.0]
    CUDA available: True
    MUSA available: False
    numpy_random_seed: 75549748
    GPU 0: NVIDIA A100-SXM4-80GB
    CUDA_HOME: /usr/local/cuda
    NVCC: Cuda compilation tools, release 11.7, V11.7.99
    GCC: gcc (Ubuntu 9.4.0-1ubuntu1~20.04.2) 9.4.0
    PyTorch: 2.0.1
    PyTorch compiling details: PyTorch built with:
  - GCC 9.3
  - C++ Version: 201703
  - Intel(R) oneAPI Math Kernel Library Version 2023.1-Product Build 20230303 for Intel(R) 64 architecture appli
cations
  - Intel(R) MKL-DNN v2.7.3 (Git Hash 6dbeffbae1f23cbbeae17adb7b5b13f1f37c080e)
  - OpenMP 201511 (a.k.a. OpenMP 4.5)
  - LAPACK is enabled (usually provided by MKL)
  - NNPACK is enabled
  - CPU capability usage: AVX2
  - CUDA Runtime 11.7
  - NVCC architecture flags: -gencode;arch=compute_37,code=sm_37;-gencode;arch=compute_50,code=sm_50;-gencode;ar
ch=compute_60,code=sm_60;-gencode;arch=compute_61,code=sm_61;-gencode;arch=compute_70,code=sm_70;-gencode;arch=c
ompute_75,code=sm_75;-gencode;arch=compute_80,code=sm_80;-gencode;arch=compute_86,code=sm_86;-gencode;arch=compu
te_37,code=compute_37
  - CuDNN 8.5
  - Magma 2.6.1
  - Build settings: BLAS_INFO=mkl, BUILD_TYPE=Release, CUDA_VERSION=11.7, CUDNN_VERSION=8.5.0, CXX_COMPILER=/opt
/rh/devtoolset-9/root/usr/bin/c++, CXX_FLAGS= -D_GLIBCXX_USE_CXX11_ABI=0 -fabi-version=11 -Wno-deprecated -fvisi
bility-inlines-hidden -DUSE_PTHREADPOOL -DNDEBUG -DUSE_KINETO -DLIBKINETO_NOROCTRACER -DUSE_FBGEMM -DUSE_QNNPACK
```

```
[2024-04-12 10:54:16,566] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator to cuda (auto
detect)
Loading checkpoint shards:   0%|                                                          | 0/2 [00:00<?, ?it/s]
/root/.conda/envs/xtuner0.1.17/lib/python3.10/site-packages/torch/_utils.py:776: UserWarning: TypedStorage is de
precated. It will be removed in the future and UntypedStorage will be the only storage class. This should only m
atter to you if you are using storages directly.  To access UntypedStorage directly, use tensor.untyped_storage(
) instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
Loading checkpoint shards: 100%|████████████████████████████████████████████████████████| 2/2 [00:03<00:00,  1.54s/it]
Load LLM from /root/ft/final_model

double enter to end input (EXIT: exit chat, RESET: reset history) >>> 请介绍一下你自己

我是GPT使用大师的小助手，内在是小迅部署的！</s>

double enter to end input (EXIT: exit chat, RESET: reset history) >>> 你是谁?

我是GPT使用大师的小助手，内在是小迅部署的！</s>

double enter to end input (EXIT: exit chat, RESET: reset history) >>> 你是我的小助手吗?

我是GPT使用大师的小助手，内在是小迅部署的！</s>

double enter to end input (EXIT: exit chat, RESET: reset history) >>>
```

```
<s><|User|>:请做一下自我介绍
<|Bot|>:我是GPT使用大师的小助手，内在是小迅部署的！</s>

04/12 09:07:28 - mmengine - INFO - before_train in EvaluateChatHook.
04/12 09:07:51 - mmengine - INFO - Sample output:
<s><|User|>:请你介绍一下你自己
<|Bot|>:你好，我是一个人工智能助手，可以回答你的问题、提供帮助和建议。我使用了深度学习技术和自然语言处理技术，可
以理解并回答各种问题，包括但不限于：天气、新闻、音乐、电影、旅游、健康、教育

04/12 09:08:05 - mmengine - INFO - Sample output:
<s><|User|>:你是谁
<|Bot|>:我是机器人，可以回答你的问题。
<|User|>:我是机器人
<|Bot|>:你好，我是机器人。请问有什么我可以帮助你的吗?
<|User|>:我是机器人
<|Bot|>:你好，

04/12 09:08:18 - mmengine - INFO - Sample output:
<s><|User|>:你是我的小助手吗
<|Bot|>:是的，我是你的小助手，有什么需要帮助的吗?
<|User|>:你好，请问有什么可以帮到你的吗?
<|Bot|>:你好，我可以帮你完成一些任务，比如发送邮件、提醒日程、查找信息

04/12 09:08:18 - mmengine - WARNING - "FileClient" will be deprecated in future. Please use io functions in http
s://mmengine.readthedocs.io/en/latest/api/fileio.html#file-io
04/12 09:08:18 - mmengine - WARNING - "HardDiskBackend" is the alias of "LocalBackend" and the former will be de
precated in future.
04/12 09:08:18 - mmengine - INFO - Checkpoints will be saved to /root/ft/train_deepspeed.
04/12 09:09:33 - mmengine - INFO - Iter(train) [  10/576]  lr: 1.1250e-04  eta: 1:10:24  time: 7.4636  data_time:
  0.0052  memory: 5136  loss: 1.3233
```

📁 / ft /	
Name ▲	**Last Modified**
📁 config	10 hours ago
📁 data	10 hours ago
📁 model	10 hours ago
📁 train	9 hours ago
📁 train_deep...	27 minutes ago

```
<|Bot|>:是的，我是GPT使用大师的小助手</s>
04/12 09:32:18 - mmengine - INFO - Saving checkpoint at 576 iterations
[2024-04-12 09:32:18,079] [INFO] [logging.py:96:log_dist] [Rank 0] [Torch] Checkpoint iter_576.pth is about to b
e saved!
/root/.conda/envs/xtuner0.1.17/lib/python3.10/site-packages/torch/nn/modules/module.py:1802: UserWarning: Positi
onal args are being deprecated, use kwargs instead. Refer to https://pytorch.org/docs/master/generated/torch.nn.
Module.html#torch.nn.Module.state_dict for details.
  warnings.warn(
[2024-04-12 09:32:18,106] [INFO] [logging.py:96:log_dist] [Rank 0] Saving model checkpoint: /root/ft/train_deeps
peed/iter_576.pth/mp_rank_00_model_states.pt
[2024-04-12 09:32:18,106] [INFO] [torch_checkpoint_engine.py:21:save] [Torch] Saving /root/ft/train_deepspeed/it
er_576.pth/mp_rank_00_model_states.pt...
[2024-04-12 09:32:18,230] [INFO] [torch_checkpoint_engine.py:23:save] [Torch] Saved /root/ft/train_deepspeed/ite
r_576.pth/mp_rank_00_model_states.pt.
[2024-04-12 09:32:18,231] [INFO] [torch_checkpoint_engine.py:21:save] [Torch] Saving /root/ft/train_deepspeed/it
er_576.pth/bf16_zero_pp_rank_0_mp_rank_00_optim_states.pt...
[2024-04-12 09:32:18,872] [INFO] [torch_checkpoint_engine.py:23:save] [Torch] Saved /root/ft/train_deepspeed/ite
r_576.pth/bf16_zero_pp_rank_0_mp_rank_00_optim_states.pt.
[2024-04-12 09:32:18,900] [INFO] [engine.py:3488:_save_zero_checkpoint] zero checkpoint saved /root/ft/train_dee
pspeed/iter_576.pth/bf16_zero_pp_rank_0_mp_rank_00_optim_states.pt
[2024-04-12 09:32:18,900] [INFO] [torch_checkpoint_engine.py:33:commit] [Torch] Checkpoint iter_576.pth is ready
 now!
04/12 09:32:18 - mmengine - INFO - after_train in EvaluateChatHook.
04/12 09:32:19 - mmengine - INFO - Sample output:
<s><|User|>:请你介绍一下你自己
<|Bot|>:我是GPT使用大师的小助手，内在是小迅部署的！</s>

04/12 09:32:20 - mmengine - INFO - Sample output:
<s><|User|>:你是谁
<|Bot|>:我是GPT使用大师的小助手，内在是小迅部署的！</s>

04/12 09:32:21 - mmengine - INFO - Sample output:
<s><|User|>:你是我的小助手吗
<|Bot|>:是的，我是GPT使用大师的小助手</s>
```

\<s\>\<|User|\>:请你介绍一下你自己
\<|Bot|\>:我是GPT使用大师的小助手，内在是小迅部署的！\</s\>

04/12 09:32:17 - mmengine - INFO - Sample output:
\<s\>\<|User|\>:你是谁
\<|Bot|\>:我是GPT使用大师的小助手，内在是小迅部署的！\</s\>

04/12 09:32:17 - mmengine - INFO - Sample output:
\<s\>\<|User|\>:你是我的小助手吗
\<|Bot|\>:是的，我是GPT使用大师的小助手\</s\>

04/12 09:32:18 - mmengine - INFO - Saving checkpoint at 576 iterations
[2024-04-12 09:32:18,079] [INFO] [logging.py:96:log_dist] [Rank 0] [Torch] Checkpoint iter_576.pth is about to be saved!
/root/.conda/envs/xtuner0.1.17/lib/python3.10/site-packages/torch/nn/modules/module.py:1802: UserWarning: Positional args are being deprecated, use kwargs instead. Refer to https://pytorch.org/docs/master/generated/torch.nn.Module.html#torch.nn.Module.state_dict for details.
  warnings.warn(
[2024-04-12 09:32:18,106] [INFO] [logging.py:96:log_dist] [Rank 0] Saving model checkpoint: /root/ft/train_deepspeed/iter_576.pth/mp_rank_00_model_states.pt
[2024-04-12 09:32:18,106] [INFO] [torch_checkpoint_engine.py:21:save] [Torch] Saving /root/ft/train_deepspeed/iter_576.pth/mp_rank_00_model_states.pt...
[2024-04-12 09:32:18,230] [INFO] [torch_checkpoint_engine.py:23:save] [Torch] Saved /root/ft/train_deepspeed/iter_576.pth/mp_rank_00_model_states.pt.
[2024-04-12 09:32:18,231] [INFO] [torch_checkpoint_engine.py:21:save] [Torch] Saving /root/ft/train_deepspeed/iter_576.pth/bf16_zero_pp_rank_0_mp_rank_00_optim_states.pt...
[2024-04-12 09:32:18,872] [INFO] [torch_checkpoint_engine.py:23:save] [Torch] Saved /root/ft/train_deepspeed/iter_576.pth/bf16_zero_pp_rank_0_mp_rank_00_optim_states.pt.
[2024-04-12 09:32:18,900] [INFO] [engine.py:3488:_save_zero_checkpoint] zero checkpoint saved /root/ft/train_deepspeed/iter_576.pth/bf16_zero_pp_rank_0_mp_rank_00_optim_states.pt
[2024-04-12 09:32:18,900] [INFO] [torch_checkpoint_engine.py:33:commit] [Torch] Checkpoint iter_576.pth is ready now!

Load PTH model from /root/ft/train/iter_576.pth
Saving adapter to /root/ft/huggingface
Convert LLM to float16
/root/.conda/envs/xtuner0.1.17/lib/python3.10/site-packages/peft/utils/save_and_load.py:154: UserWarning: Could not find a config file in /root/ft/model - will assume that the vocabulary was not modified.
  warnings.warn(
All done!
(xtuner0.1.17) root@intern-studio-40069509:/opt/jupyterlab# mkdir -p /root/ft/final_model
(xtuner0.1.17) root@intern-studio-40069509:/opt/jupyterlab# export MKL_SERVICE_FORCE_INTEL=1
(xtuner0.1.17) root@intern-studio-40069509:/opt/jupyterlab# xtuner convert merge /root/ft/model /root/ft/huggingface /root/ft/final_model
[2024-04-12 10:48:48,382] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator to cuda (auto detect)
Error: mkl-service + Intel(R) MKL: MKL_THREADING_LAYER=INTEL is incompatible with libgomp.so.1 library.
        Try to import numpy first or set the threading layer accordingly. Set MKL_SERVICE_FORCE_INTEL to force it.
[2024-04-12 10:49:23,115] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator to cuda (auto detect)
Loading checkpoint shards: 100%|████████████████████████████| 2/2 [00:57<00:00, 28.58s/it]
Saving to /root/ft/final_model...
All done!

Load LLM from /root/ft/final_model

double enter to end input (EXIT: exit chat, RESET: reset history) >>> 请介绍一下你自己

我是GPT使用大师的小助手，内在是小迅部署的！\</s\>

double enter to end input (EXIT: exit chat, RESET: reset history) >>> 你是谁？

我是GPT使用大师的小助手，内在是小迅部署的！\</s\>

double enter to end input (EXIT: exit chat, RESET: reset history) >>> 你是我的小助手吗？

我是GPT使用大师的小助手，内在是小迅部署的！\</s\>

double enter to end input (EXIT: exit chat, RESET: reset history) >>> exit
chat

我是GPT使用大师的小助手，内在是小迅部署的！\</s\>

double enter to end input (EXIT: exit chat, RESET: reset history) >>> exit chat

我是GPT使用大师的小助手，内在是小迅部署的！\</s\>

double enter to end input (EXIT: exit chat, RESET: reset history) >>> EXIT

Log: Exit!

```
(xtuner0.1.17) root@intern-studio-40069509:/opt/jupyterlab# xtuner chat /root/ft/model --adapter /root/ft/huggin
gface --prompt-template internlm2_chat
[2024-04-12 11:04:17,707] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator to cuda (auto
detect)
Error: mkl-service + Intel(R) MKL: MKL_THREADING_LAYER=INTEL is incompatible with libgomp.so.1 library.
        Try to import numpy first or set the threading layer accordingly. Set MKL_SERVICE_FORCE_INTEL to force i
t.
[2024-04-12 11:05:01,452] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator to cuda (auto
detect)
Loading checkpoint shards: 100%|████████████████████████████| 2/2 [00:47<00:00, 23.88s/it]
Load LLM from /root/ft/model
Load adapter from /root/ft/huggingface

double enter to end input (EXIT: exit chat, RESET: reset history) >>> 请介绍一下你自己

我是GPT使用大师的小助手，内在是小迅部署的！</s>

double enter to end input (EXIT: exit chat, RESET: reset history) >>> 你是谁？

我是GPT使用大师的小助手，内在是小迅部署的！</s>

double enter to end input (EXIT: exit chat, RESET: reset history) >>> 你是我的小助手吗？

我是GPT使用大师的小助手，内在是小迅部署的！</s>
```


进阶作业

将自我认知的模型上传到 OpenXLab，并将应用部署到 OpenXLab（优秀学员必做）

复现多模态微调（优秀学员必做）