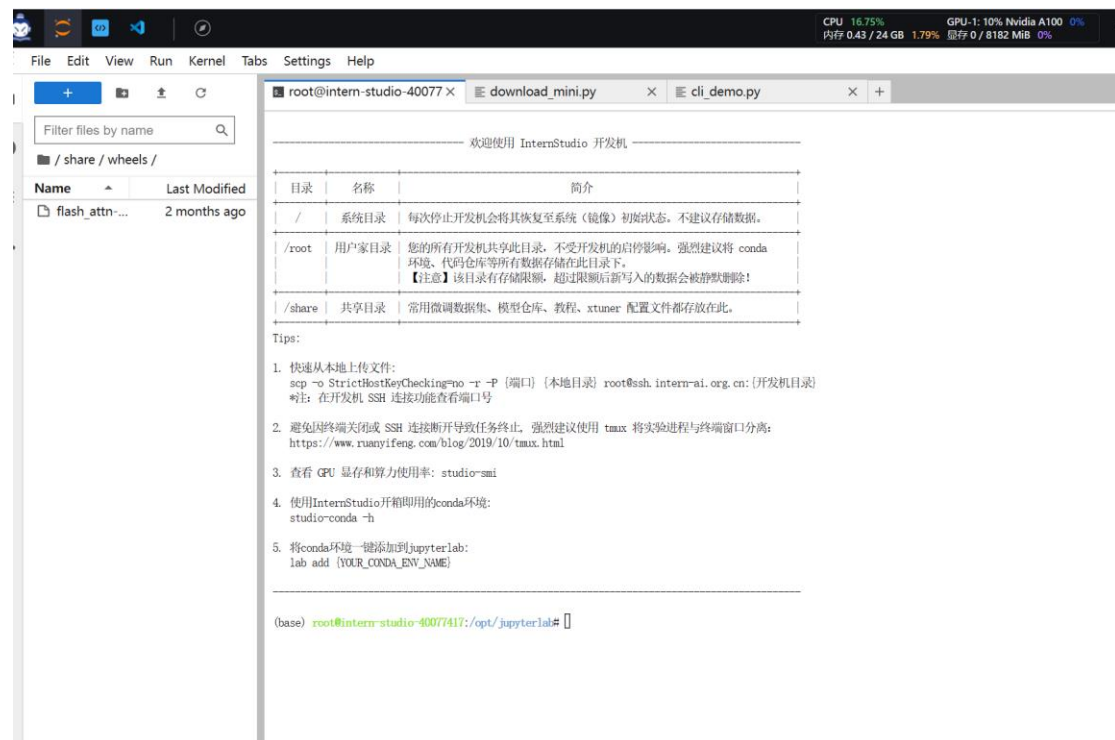


选择 Jupyterlab 进入 conda 环境



配置环境:

```
studio-conda -o internlm-base -t demo
# 与 studio-conda 等效的配置方案
# conda create -n demo python==3.10 -y
# conda activate demo
# conda install pytorch==2.0.1 torchvision==0.15.2 torchaudio==2.0.2
```



```
pytorch-cuda=11.7 -c pytorch -c nvidia
```

```
lab add {YOUR_CONDA_ENV_NAME}
```

```
(base) root@intern-studio-40077417:/opt/jupyterlab# mkdir -p /root/model/Shanghai_AI_Laboratory
(base) root@intern-studio-40077417:/opt/jupyterlab# cp -r /root/share/temp/model_repos/internlm-chat-7b /root/model/Shanghai_AI_Laboratory
(base) root@intern-studio-40077417:/opt/jupyterlab#
```

创建 demo 文件:

```
mkdir -p /root/demo
touch /root/demo/cli_demo.py
touch /root/demo/download_mini.py
cd /root/demo
```

	Last Modified
 download...	yesterday
 cli_demo.py	yesterday

```
运行 cli_demo:
```

```
import torch
```

```

from transformers import AutoTokenizer, AutoModelForCausalLM

model_name_or_path = "/root/models/Shanghai_AI_Laboratory/internlm2-
chat-1_8b"

tokenizer = AutoTokenizer.from_pretrained(model_name_or_path,
trust_remote_code=True, device_map='cuda:0')
model = AutoModelForCausalLM.from_pretrained(model_name_or_path,
trust_remote_code=True, torch_dtype=torch.bfloat16,
device_map='cuda:0')
model = model.eval()

system_prompt = """"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, 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.
""""

messages = [(system_prompt, '')]

print("=====Welcome to InternLM chatbot, type 'exit' to
exit.=====")

while True:
    input_text = input("\nUser >>> ")
    input_text = input_text.replace(' ', '')
    if input_text == "exit":
        break

    length = 0
    for response, _ in model.stream_chat(tokenizer, input_text,
messages):
        if response is not None:
            print(response[length:], flush=True, end="")
            length = len(response)

```

```
root@intern-studio-40077 x download_mini.py x cli_demo.py x +
1 import torch
2 from transformers import AutoTokenizer, AutoModelForCausalLM
3
4
5 model_name_or_path = "/root/models/Shanghai_AI_Laboratory/internlm2-chat-1_8b"
6
7 tokenizer = AutoTokenizer.from_pretrained(model_name_or_path, trust_remote_code=True, device_map='cuda:0')
8 model = AutoModelForCausalLM.from_pretrained(model_name_or_path, trust_remote_code=True, torch_dtype=torch.bfloat16, device_map='cuda:0')
9 model = model.eval()
10
11 system_prompt = """You are an AI assistant whose name is InternLM (书生·浦语).
12 - 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
13 helpful, honest, and harmless.
14 - InternLM (书生·浦语) can understand and communicate fluently in the language chosen by the user such as English and 中文.
15 """
16 messages = [(system_prompt, '')]
17
18 print("=====Welcome to InternLM chatbot, type 'exit' to exit.=====")
19
20 while True:
21     input_text = input("\nUser >>> ")
22     input_text = input_text.replace(' ', '')
23     if input_text == "exit":
24         break
25
26     length = 0
27     for response, _ in model.stream_chat(tokenizer, input_text, messages):
28         if response is not None:
29             print(response[length:], flush=True, end="")
30             length = len(response)
31
```

最终显示界面：

```
File Edit View Run Kernel Tabs Settings Help
Filter files by name
Name Last Modified
flash_attn... 2 months ago

shldir: cannot create directory /root/models: File exists
2024-04-01 20:09:46.206 - modelscope - INFO - Use user-specified model revision: v1.1.0
(demo) root@intern-studio-40077: /demo# python /root/demo/download_mini.py
2024-04-01 20:17:02.188 - modelscope - INFO - PyTorch version 2.0.1 Found.
2024-04-01 20:17:02.189 - modelscope - INFO - Loading ast index from /root/.cache/modelscope/ast_indexer
2024-04-01 20:17:12.572 - modelscope - INFO - Loading done! Current index file version is 1.9.5, with md5 5393c307841f22e324481815bd82 and a total number of 945 compon
ents indexed
2024-04-01 20:17:13.077 - modelscope - INFO - Use user-specified model revision: v1.1.0
Downloading: 100% | 850/850 [00:00:00.00, 5.81MB/s]
Downloading: 100% | 48.0/48.0 [00:00:00.00, 4682B/s]
Downloading: 100% | 6.88k/6.88k [00:00:00.00, 441kB/s]
Downloading: 100% | 132/132 [00:00:00.00, 1.18MB/s]
Downloading: 100% | 1.85G/1.85G [00:40:00.00, 49.2MB/s]
Downloading: 100% | 1.67G/1.67G [01:19:00.00, 22.6MB/s]
Downloading: 100% | 13.4k/13.4k [00:00:00.00, 3.36MB/s]
Downloading: 100% | 58.6k/58.6k [00:00:00.00, 6.95MB/s]
Downloading: 100% | 10.4k/10.4k [00:00:00.00, 69.2MB/s]
Downloading: 100% | 713/713 [00:00:00.00, 7.37MB/s]
Downloading: 100% | 8.60k/8.60k [00:00:00.00, 58.1MB/s]
Downloading: 100% | 7.63k/7.63k [00:00:00.00, 62.0MB/s]
Downloading: 100% | 1.41M/1.41M [00:00:00.00, 10.4MB/s]
Downloading: 100% | 2.45k/2.45k [00:00:00.00, 28.2MB/s]
(demo) root@intern-studio-40077: /demo# conda activate demo
(demo) root@intern-studio-40077: /demo# python /root/demo/cli_demo.py
(demo) root@intern-studio-40077: /demo# conda activate demo
(demo) root@intern-studio-40077: /demo# python /root/demo/cli_demo.py
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.
Loading checkpoint shards: 100% | 2/2 [00:33:00.00, 16.67s/it]
=====Welcome to InternLM chatbot, type 'exit' to exit.=====

User >>> 请创作一个 300 字的小故事
从前，有一个名叫小明的男孩，他非常喜欢探险和冒险。一天，他听说了一座神秘的山脉，据说那里有神秘的宝藏。小明非常兴奋，决定踏上这个充满挑战的旅程。
他穿过了茂密的森林，爬过了陡峭的山峰，穿过了一片神秘的沼泽。一路上，他遇到了许多困难和挑战，但他始终坚定地朝着目标前进。
终于，他到达了山脉的尽头，发现了一座巨大的洞穴。小明兴奋地走了进去，发现里面有很多宝藏和珍宝。但是，他突然意识到，这个洞穴里还有一个可怕的怪物，它随时可能攻击他。
小明非常害怕，但他没有退缩。他决定与怪物展开一场激烈的战斗，最终成功地打败了它。他找到了宝藏，带着它们回到了家乡。
小明回到家乡后，他的朋友们都问他为什么如此勇敢，他告诉他们，他知道，只有勇敢面对困难，才能克服它们。他明白了，勇气和决心是成功的关键。
User >>>
```