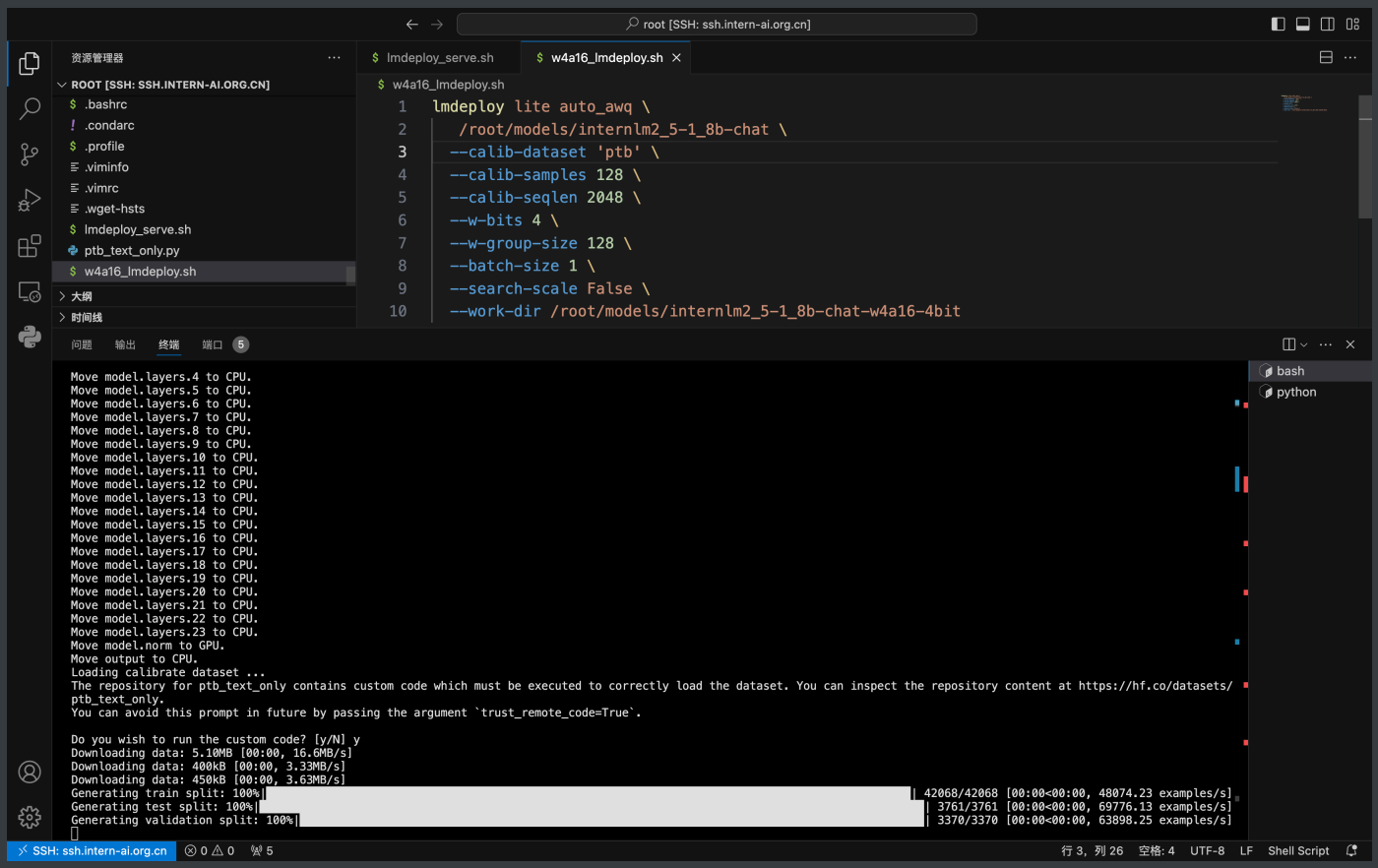
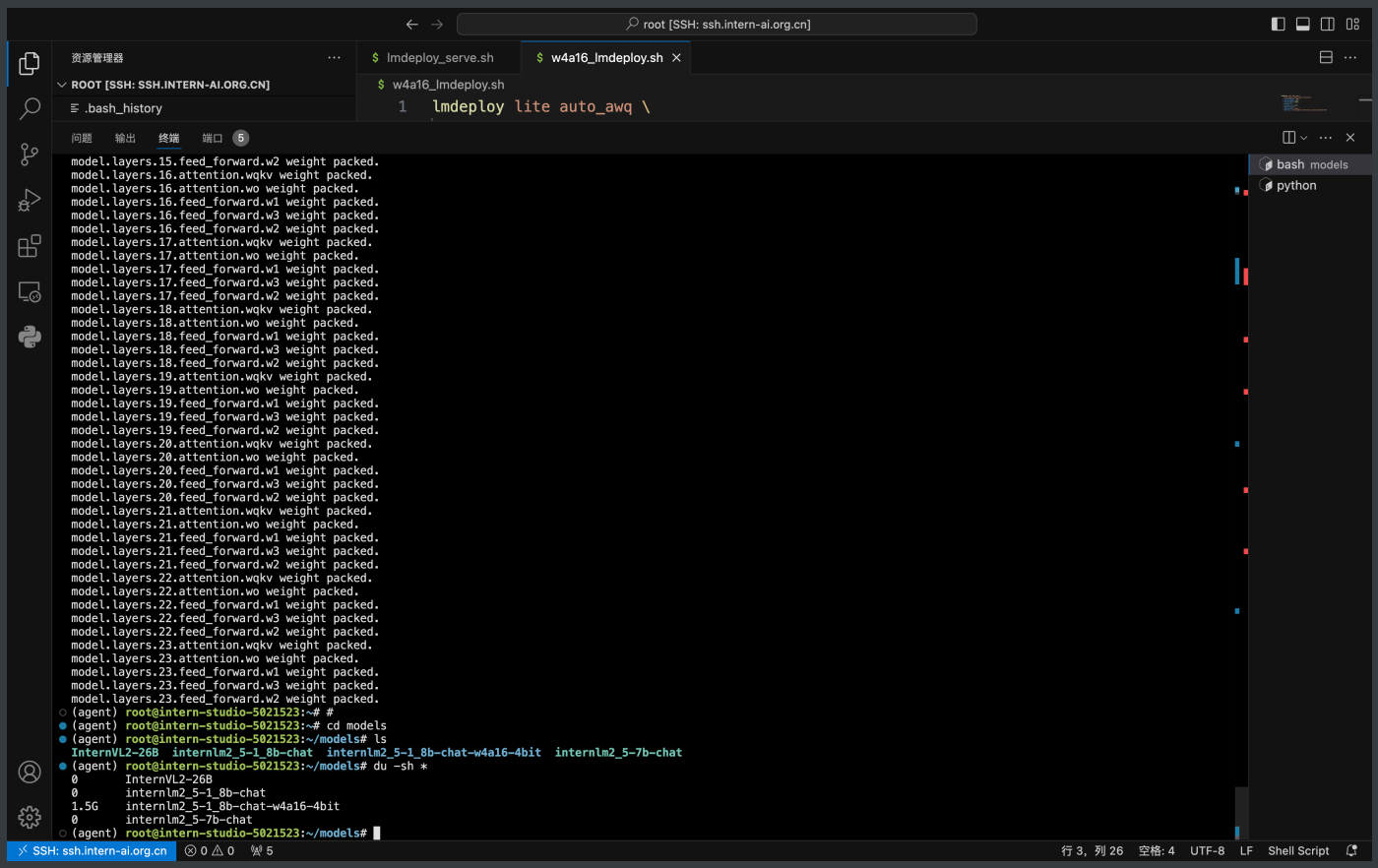
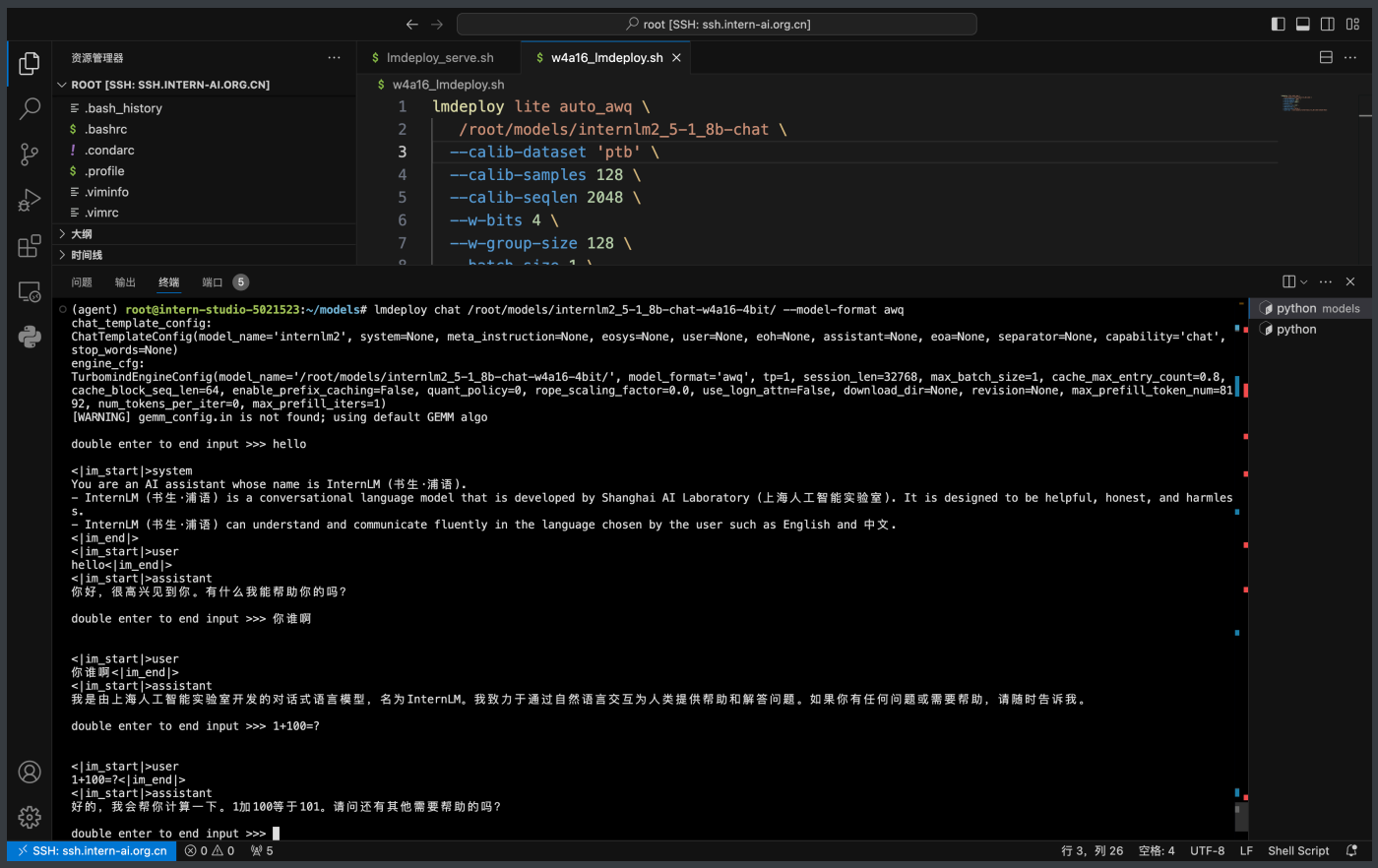
量化internlm2_5-1_8b-chat



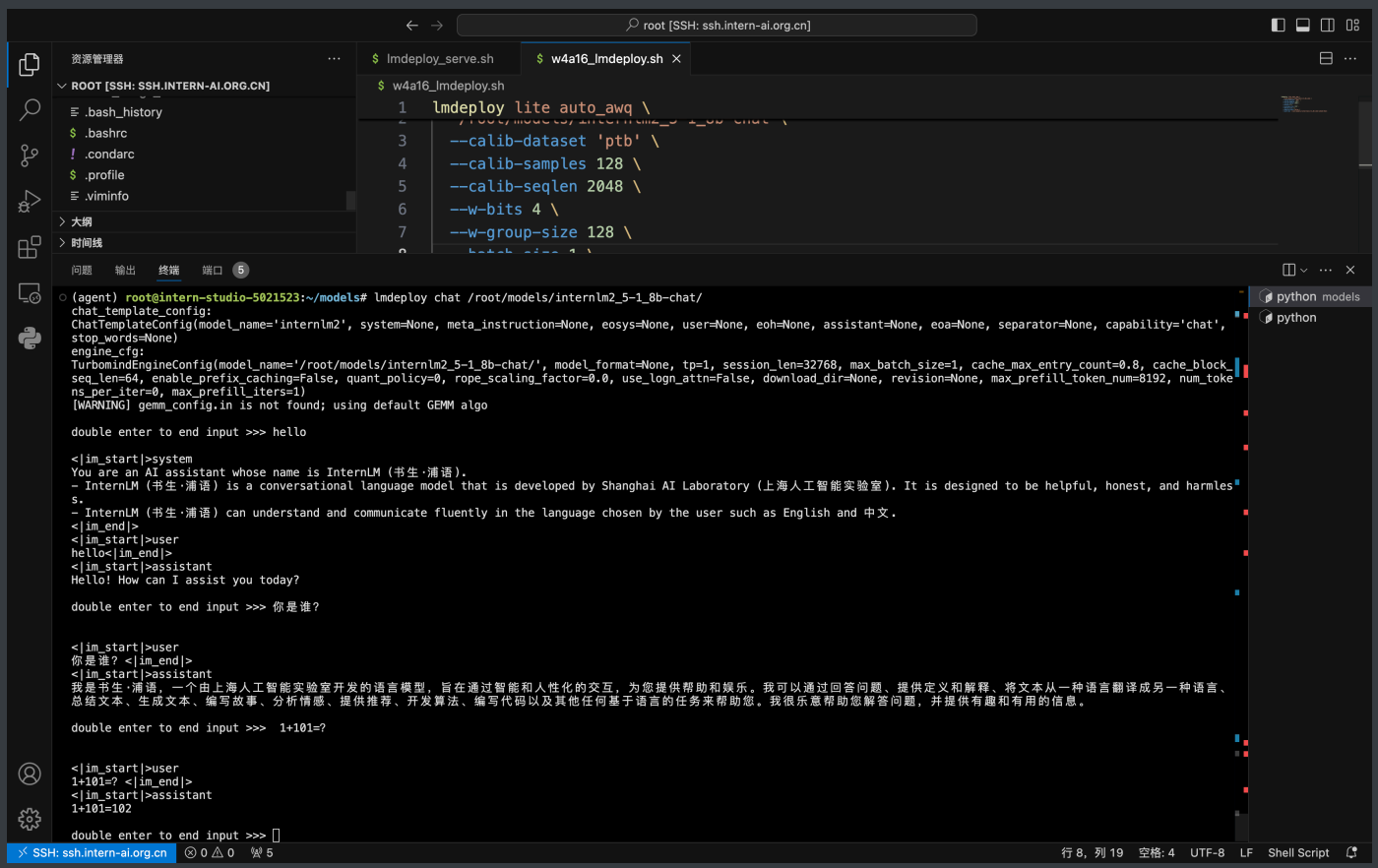


量化模型API对话

量化后的模型

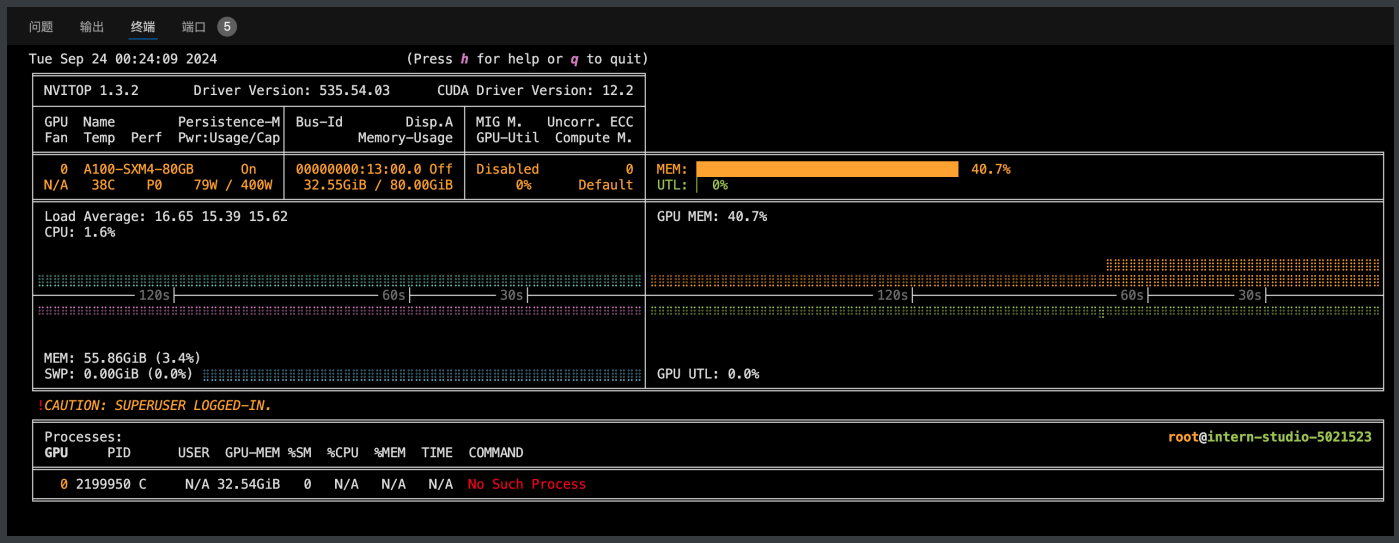

root [SSH: ssh.intern-ai.org.cn]
资源管理器
ROOT [SSH: SSH.INTERN-AI.ORG.CN]
.bash_history
.bashrc
.condarc
.profile
.viminfo
.vimrc
大纲
时间线
lmdeploy_serve.sh
w4a16_lmdeploy.sh
w4a16_lmdeploy.sh
1 lmdeploy lite auto_awq \
2 /root/models/internlm2_5-1_8b-chat \
3 --calib-dataset 'ptb' \
4 --calib-samples 128 \
5 --calib-seqlen 2048 \
6 --w-bits 4 \
7 --w-group-size 128 \
问题 输出 终端 端口 5
(agent) root@intern-studio-5021523:~/models# lmdeploy chat /root/models/internlm2_5-1_8b-chat-w4a16-4bit/ --model-format awq
chat_template_config:
ChatTemplateConfig(model_name='internlm2', system=None, meta_instruction=None, eosys=None, user=None, eoh=None, assistant=None, eoa=None, separator=None, capability='chat', stop_words=None)
engine_cfg:
TurbomindEngineConfig(model_name='/root/models/internlm2_5-1_8b-chat-w4a16-4bit/', model_format='awq', tp=1, session_len=32768, max_batch_size=1, cache_max_entry_count=0.8, cache_block_seq_len=64, enable_prefix_caching=False, quant_policy=0, rope_scaling_factor=0.0, use_logn_attn=False, download_dir=None, revision=None, max_prefill_token_num=8192, num_tokens_per_iter=0, max_prefill_iters=1)
[WARNING] gemm_config.in is not found; using default GEMM algo
double enter to end input >>> hello
<|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 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
hello<|im_end|>
<|im_start|>assistant
你好，很高兴见到你。有什么我能帮助你的吗？
double enter to end input >>> 你谁啊
<|im_start|>user
你谁啊<|im_end|>
<|im_start|>assistant
我是由上海人工智能实验室开发的对话式语言模型，名为InternLM。我致力于通过自然语言交互为人类提供帮助和解答问题。如果你有任何问题或需要帮助，请随时告诉我。
double enter to end input >>> 1+100=?
<|im_start|>user
1+100=?<|im_end|>
<|im_start|>assistant
好的，我会帮你计算一下。1加100等于101。请问还有其他需要帮助的吗？
double enter to end input >>>
python models
python
SSH: ssh.intern-ai.org.cn
行 3，列 26 空格: 4 UTF-8 LF Shell Script

量化前的模型

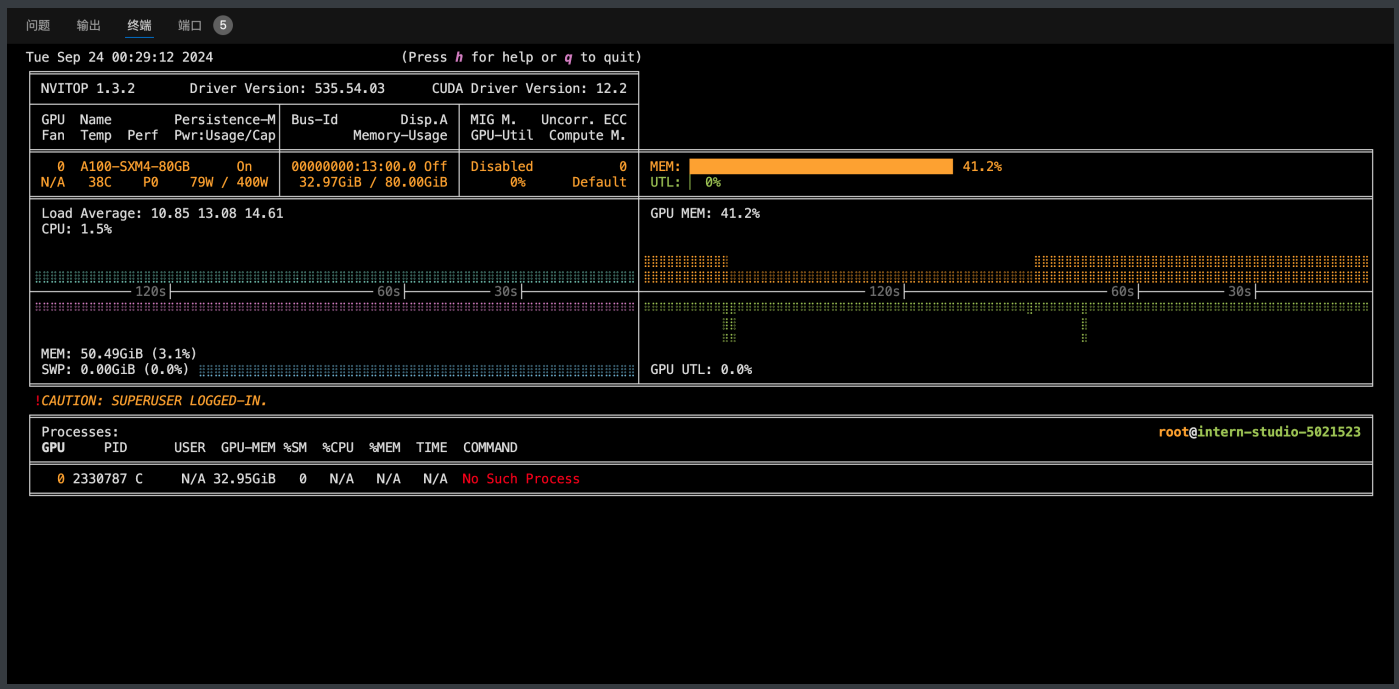

root [SSH: ssh.intern-ai.org.cn]
资源管理器
ROOT [SSH: SSH.INTERN-AI.ORG.CN]
.bash_history
.bashrc
.condarc
.profile
.viminfo
大纲
时间线
lmdeploy_serve.sh
w4a16_lmdeploy.sh
w4a16_lmdeploy.sh
1 lmdeploy lite auto_awq \
3 --calib-dataset 'ptb' \
4 --calib-samples 128 \
5 --calib-seqlen 2048 \
6 --w-bits 4 \
7 --w-group-size 128 \
问题 输出 终端 端口 5
(agent) root@intern-studio-5021523:~/models# lmdeploy chat /root/models/internlm2_5-1_8b-chat/
chat_template_config:
ChatTemplateConfig(model_name='internlm2', system=None, meta_instruction=None, eosys=None, user=None, eoh=None, assistant=None, eoa=None, separator=None, capability='chat', stop_words=None)
engine_cfg:
TurbomindEngineConfig(model_name='/root/models/internlm2_5-1_8b-chat/', model_format=None, tp=1, session_len=32768, max_batch_size=1, cache_max_entry_count=0.8, cache_block_seq_len=64, enable_prefix_caching=False, quant_policy=0, rope_scaling_factor=0.0, use_logn_attn=False, download_dir=None, revision=None, max_prefill_token_num=8192, num_tokens_per_iter=0, max_prefill_iters=1)
[WARNING] gemm_config.in is not found; using default GEMM algo
double enter to end input >>> hello
<|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 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
hello<|im_end|>
<|im_start|>assistant
Hello! How can I assist you today?
double enter to end input >>> 你是谁?
<|im_start|>user
你是谁? <|im_end|>
<|im_start|>assistant
我是书生·浦语，一个由上海人工智能实验室开发的语言模型，旨在通过智能和人性化的交互，为您提供帮助和娱乐。我可以通过回答问题、提供定义和解释、将文本从一种语言翻译成另一种语言、总结文本、生成文本、编写故事、分析情感、提供推荐、开发算法、编写代码以及其他任何基于语言的任务来帮助您。我很乐意帮助您解答问题，并提供有趣和有用的信息。
double enter to end input >>> 1+101=?
<|im_start|>user
1+101=? <|im_end|>
<|im_start|>assistant
1+101=102
double enter to end input >>>
python models
python
SSH: ssh.intern-ai.org.cn
行 8，列 19 空格: 4 UTF-8 LF Shell Script

50%A100运行现存占用

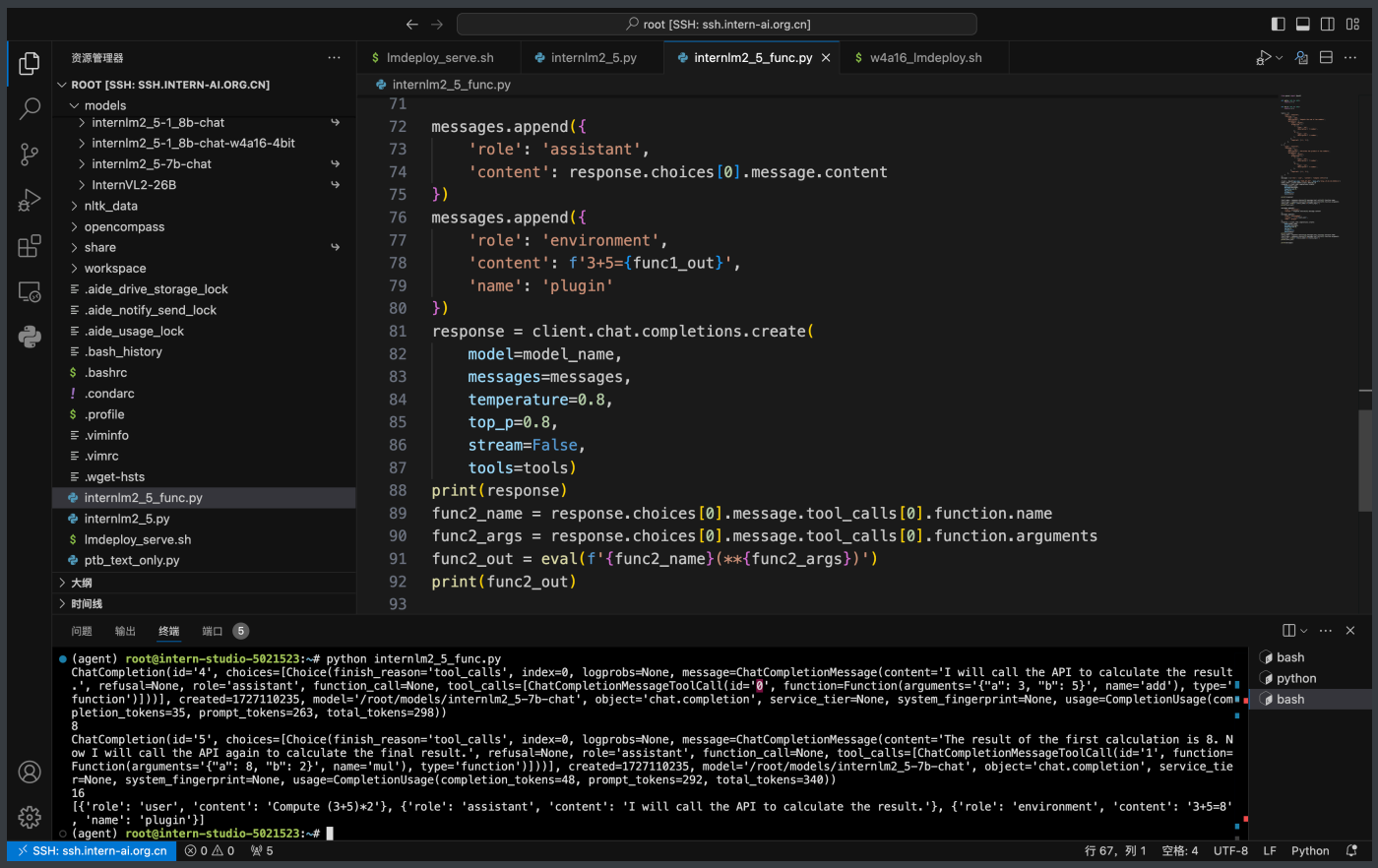
量化后的模型



量化前的模型



Function call


root [SSH: ssh.intern-ai.org.cn]
资源管理器
ROOT [SSH: SSH.INTERN-AI.ORG.CN]
models
internlm2_5-1_8b-chat
internlm2_5-1_8b-chat-w4a16-4bit
internlm2_5-7b-chat
InternVL2-26B
nltk_data
opencompass
share
workspace
.aide_drive_storage_lock
.aide_notify_send_lock
.aide_usage_lock
.bash_history
.bashrc
.condarc
.profile
.viminfo
.vimrc
.wget-hsts
internlm2_5_func.py
internlm2_5.py
lmdeploy_serve.sh
ptb_text_only.py
大纲
时间线
lmdeploy_serve.sh
internlm2_5.py
internlm2_5_func.py
w4a16_lmdeploy.sh
internlm2_5_func.py
71
messages.append({
'role': 'assistant',
'content': response.choices[0].message.content
})
messages.append({
'role': 'environment',
'content': f'3+5={func1_out}',
'name': 'plugin'
})
response = client.chat.completions.create(
model=model_name,
messages=messages,
temperature=0.8,
top_p=0.8,
stream=False,
tools=tools)
print(response)
func2_name = response.choices[0].message.tool_calls[0].function.name
func2_args = response.choices[0].message.tool_calls[0].function.arguments
func2_out = eval(f'{func2_name}(**{func2_args})')
print(func2_out)
93
问题
输出
终端
端口
5
(agent) root@intern-studio-5021523:~# python internlm2_5_func.py
ChatCompletion(id='4', choices=[Choice(finish_reason='tool_calls', index=0, logprobs=None, message=ChatCompletionMessage(content='I will call the API to calculate the result.', refusal=None, role='assistant', function_call=None, tool_calls=[ChatCompletionMessageToolCall(id='0', function=Function(arguments='{"a": 3, "b": 5}', name='add'), type='function')]))], created=1727110235, model='/root/models/internlm2_5-7b-chat', object='chat.completion', service_tier=None, system_fingerprint=None, usage=CompletionUsage(completion_tokens=35, prompt_tokens=263, total_tokens=298))
8
ChatCompletion(id='5', choices=[Choice(finish_reason='tool_calls', index=0, logprobs=None, message=ChatCompletionMessage(content='The result of the first calculation is 8. Now I will call the API again to calculate the final result.', refusal=None, role='assistant', function_call=None, tool_calls=[ChatCompletionMessageToolCall(id='1', function=Function(arguments='{"a": 8, "b": 2}', name='mul'), type='function')]))], created=1727110235, model='/root/models/internlm2_5-7b-chat', object='chat.completion', service_tier=None, system_fingerprint=None, usage=CompletionUsage(completion_tokens=48, prompt_tokens=292, total_tokens=340))
16
[{'role': 'user', 'content': 'Compute (3+5)*2'}, {'role': 'assistant', 'content': 'I will call the API to calculate the result.'}, {'role': 'environment', 'content': '3+5=8', 'name': 'plugin'}]
(agent) root@intern-studio-5021523:~#
bash
python
bash
SSH: ssh.intern-ai.org.cn
0 0
5
行 67，列 1
空格: 4
UTF-8
LF
Python