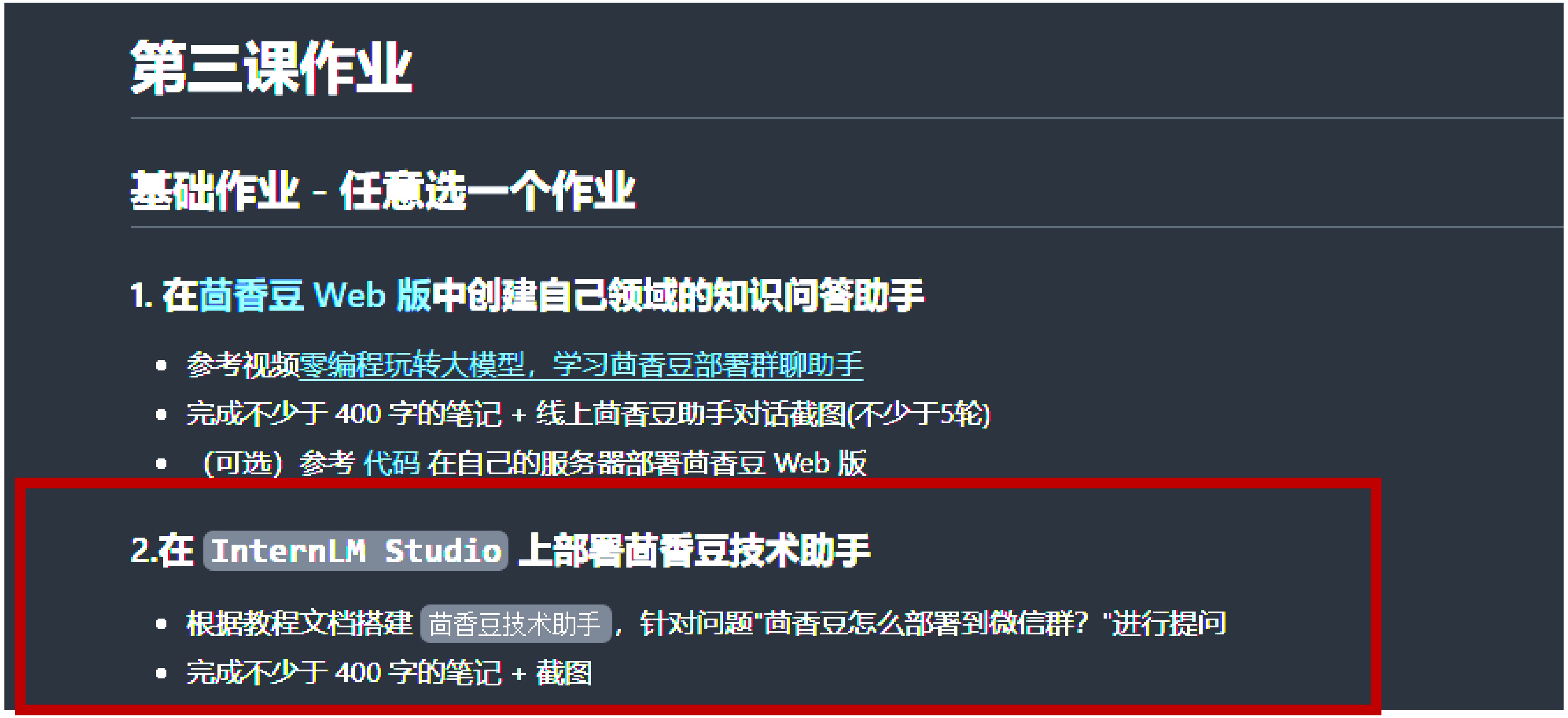


第三课作业

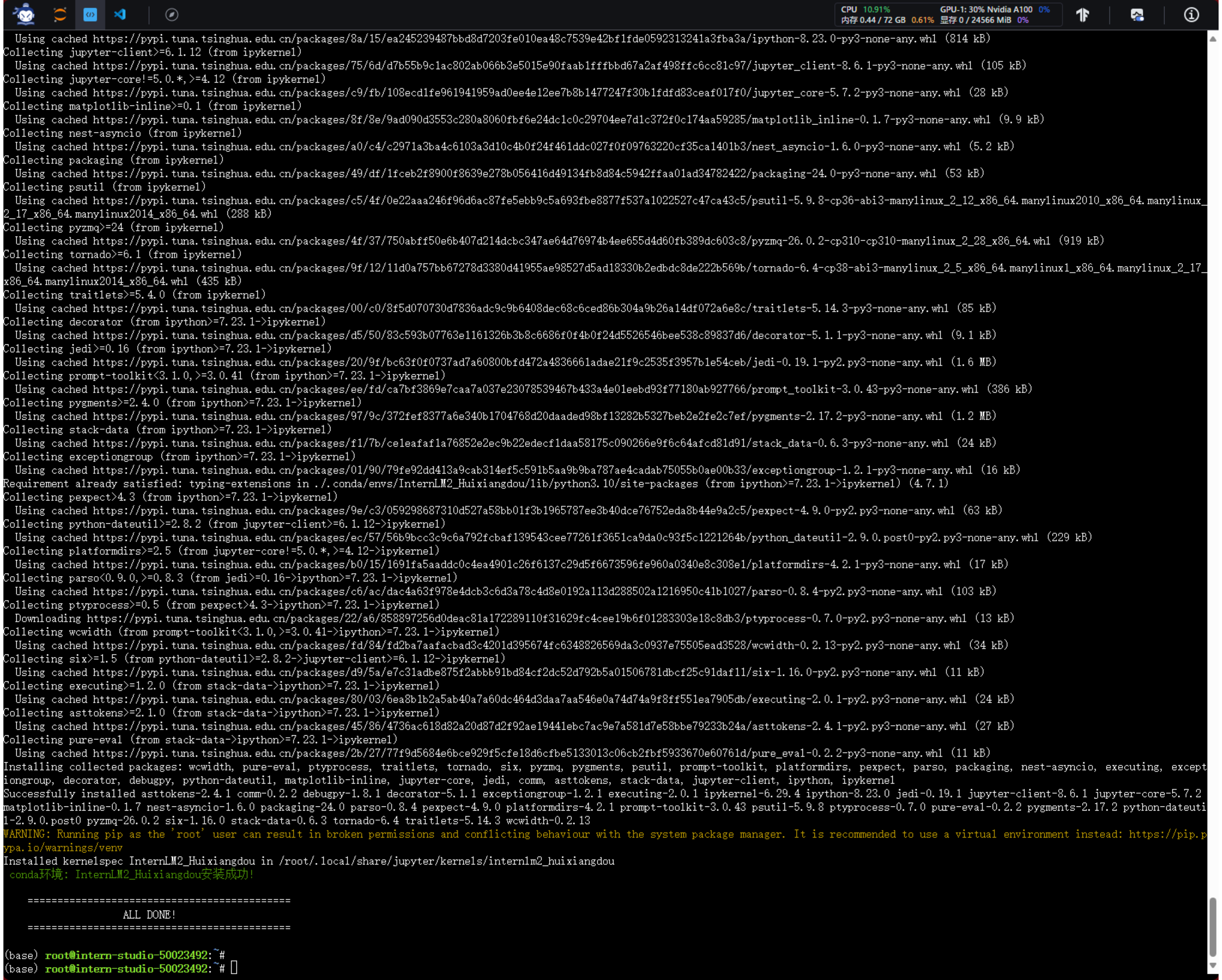
基础作业 - 任意选一个作业

1. 在茴香豆 Web 版中创建自己领域的知识问答助手

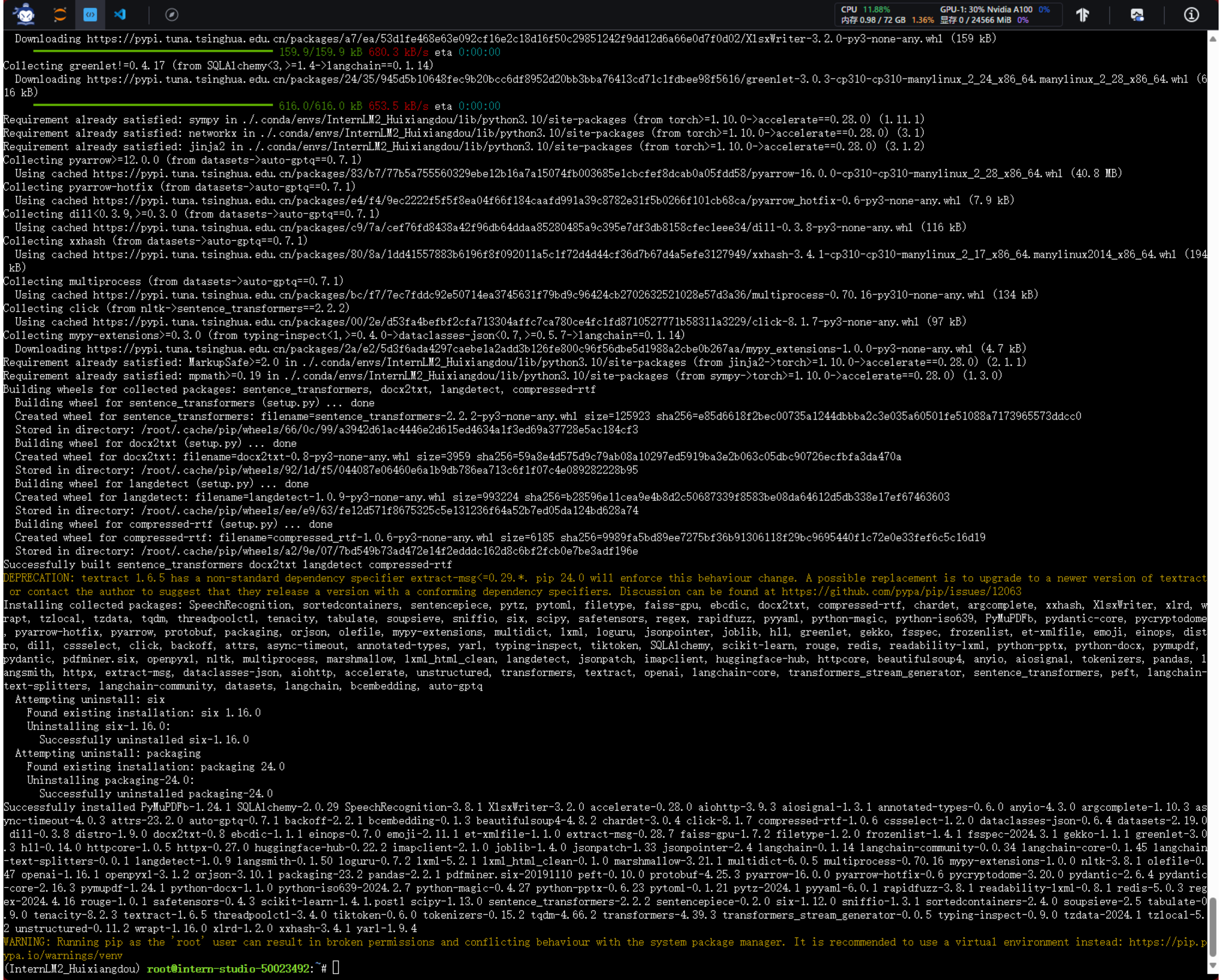
- 参考视频零编程玩转大模型，学习茴香豆部署群聊助手
- 完成不少于 400 字的笔记 + 线上茴香豆助手对话截图(不少于5轮)
- （可选）参考 代码 在自己的服务器部署茴香豆 Web 版

2.在 `InternLM Studio` 上部署茴香豆技术助手

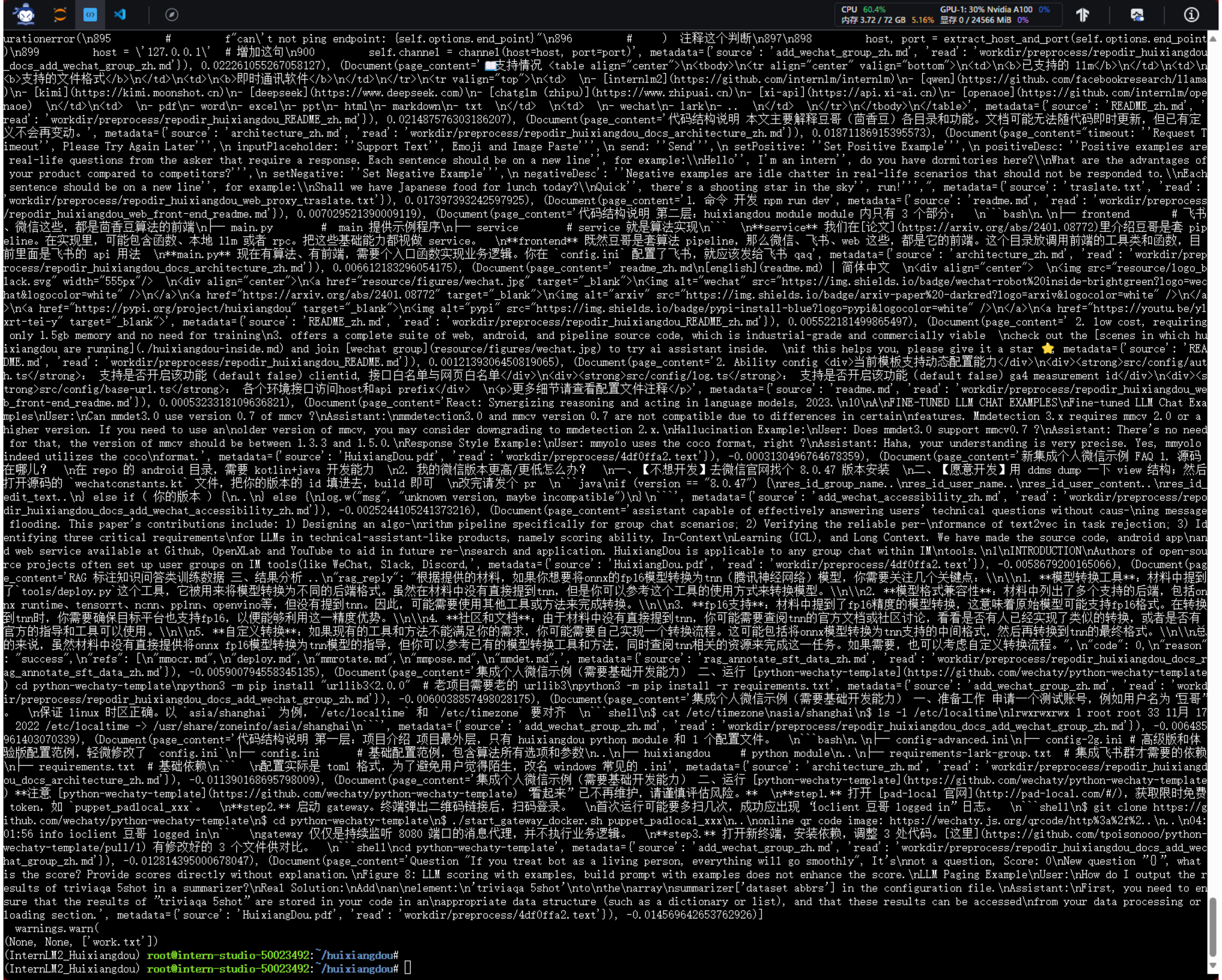
- 根据教程文档搭建 `茴香豆技术助手`，针对问题"茴香豆怎么部署到微信群？"进行提问
- 完成不少于 400 字的笔记 + 截图



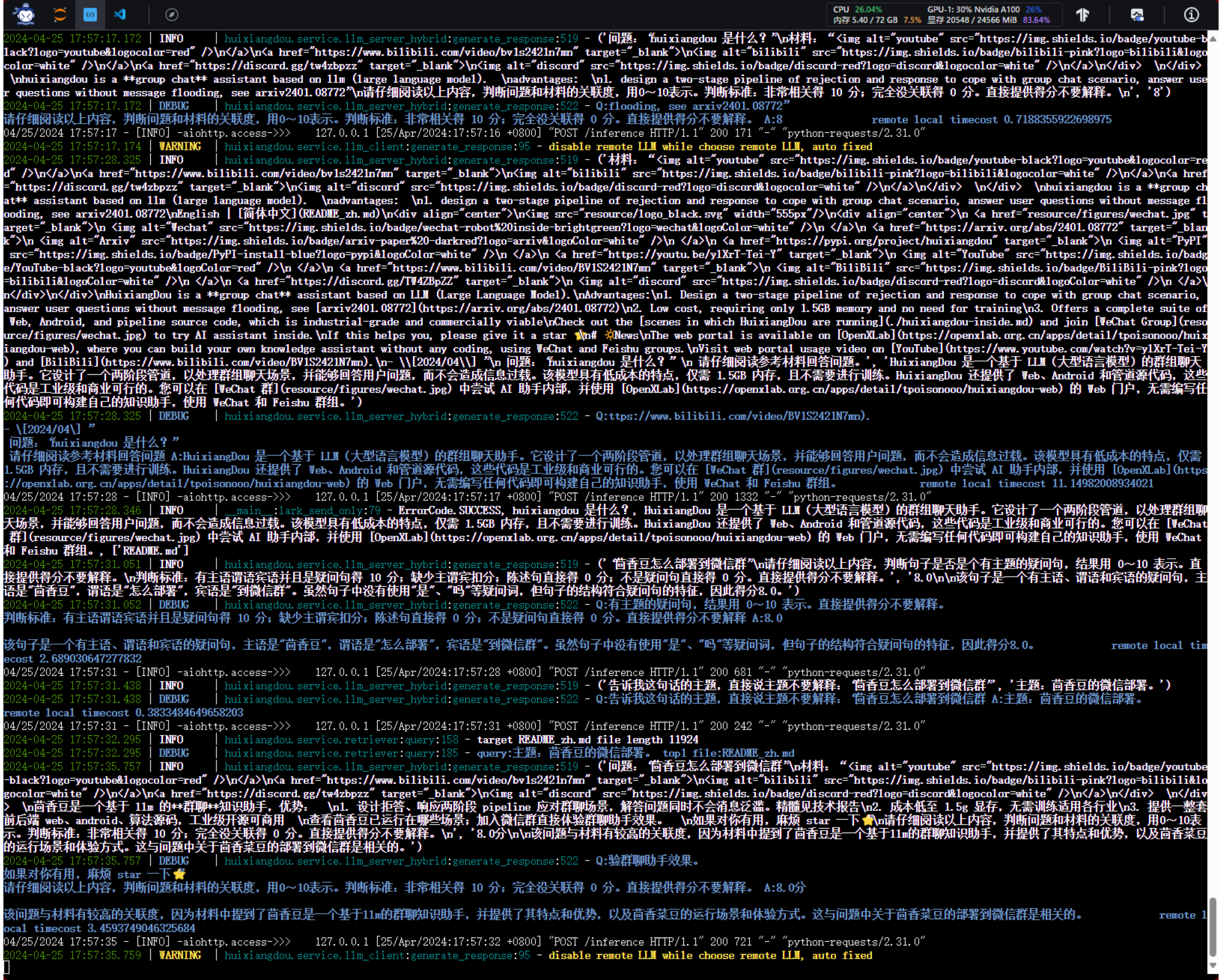
```
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/8a/15/ea245239487bbd8d7203fe010ea48c7539e42bf1fde0592313241a3fba3a/ipython-8.23.0-py3-none-any.whl (814 kB)
Collecting jupyter-client>=6.1.12 (from ipykernel)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/75/6d/d7b55b9c1ac802ab066b3e5015e90faab1fffbbd67a2af498ffc6cc81c97/jupyter_client-8.6.1-py3-none-any.whl (105 kB)
Collecting jupyter-core!=5.0.*,>=4.12 (from ipykernel)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/c9/fb/108ecd1fe961941959ad0ee4e12ee7b8b1477247f30b1fdfd83ceaf017f0/jupyter_core-5.7.2-py3-none-any.whl (28 kB)
Collecting matplotlib-inline>=0.1 (from ipykernel)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/8f/8e/9ad090d3553c280a8060fbf6e24dc1c0c29704ee7d1c372f0c174aa59285/matplotlib_inline-0.1.7-py3-none-any.whl (9.9 kB)
Collecting nest-asyncio (from ipykernel)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/a0/c4/c2971a3ba4c6103a3d10c4b0f24f461ddc027f0f09763220cf35ca1401b3/nest_asyncio-1.6.0-py3-none-any.whl (5.2 kB)
Collecting packaging (from ipykernel)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/49/df/1fceb2f8900f8639e278b056416d49134fb8d84c5942ffaa01ad34782422/packaging-24.0-py3-none-any.whl (53 kB)
Collecting psutil (from ipykernel)
  Using cached 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)
Collecting pyzmq>=24 (from ipykernel)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/4f/37/750abff50e6b407d214dcbc347ae64d76974b4ee655d4d60fb389dc603c8/pyzmq-26.0.2-cp310-cp310-manylinux_2_28_x86_64.whl (919 kB)
Collecting tornado>=6.1 (from ipykernel)
  Using cached 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_
x86_64.manylinux2014_x86_64.whl (435 kB)
Collecting traitlets>=5.4.0 (from ipykernel)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/00/c0/8f5d070730d7836adc9c9b6408dec68c6ced86b304a9b26a14df072a6e8c/traitlets-5.14.3-py3-none-any.whl (85 kB)
Collecting decorator (from ipython>=7.23.1->ipykernel)
  Using cached 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)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/20/9f/bc63f0f0737ad7a60800bfd472a4836661adae21f9c2535f3957b1e54ceb/jedi-0.19.1-py2.py3-none-any.whl (1.6 MB)
Collecting prompt-toolkit<3.1.0,>=3.0.41 (from ipython>=7.23.1->ipykernel)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/ee/fd/ca7bf3869e7caa7a037e23078539467b433a4e01eebd93f77180ab927766/prompt_toolkit-3.0.43-py3-none-any.whl (386 kB)
Collecting pygments>=2.4.0 (from ipython>=7.23.1->ipykernel)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/97/9c/372fef8377a6e340b1704768d20daaded98bf13282b5327beb2e2fe2c7ef/pygments-2.17.2-py3-none-any.whl (1.2 MB)
Collecting stack-data (from ipython>=7.23.1->ipykernel)
  Using cached 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)
  Using cached 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/InternLM2_Huixiangdou/lib/python3.10/site-packages (from ipython>=7.23.1->ipykernel) (4.7.1)
Collecting pexpect>4.3 (from ipython>=7.23.1->ipykernel)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/9e/c3/059298687310d527a58bb01f3b1965787ee3b40dce76752eda8b44e9a2c5/pexpect-4.9.0-py2.py3-none-any.whl (63 kB)
Collecting python-dateutil>=2.8.2 (from jupyter-client>=6.1.12->ipykernel)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/ec/57/56b9bcc3c9c6a792fcbaf139543cee77261f3651ca9da0c93f5c1221264b/python_dateutil-2.9.0.post0-py2.py3-none-any.whl (229 kB)
Collecting platformdirs>=2.5 (from jupyter-core!=5.0.*,>=4.12->ipykernel)
  Using cached 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)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/c6/ac/dac4a63f978e4dcb3c6d3a78c4d8e0192a113d288502a1216950c41b1027/parso-0.8.4-py2.py3-none-any.whl (103 kB)
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)
  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, 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 InternLM2_Huixiangdou in /root/.local/share/jupyter/kernels/internlm2_huixiangdou
 conda环境: InternLM2_Huixiangdou安装成功!

    ============================================

                    ALL DONE!

    ============================================

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



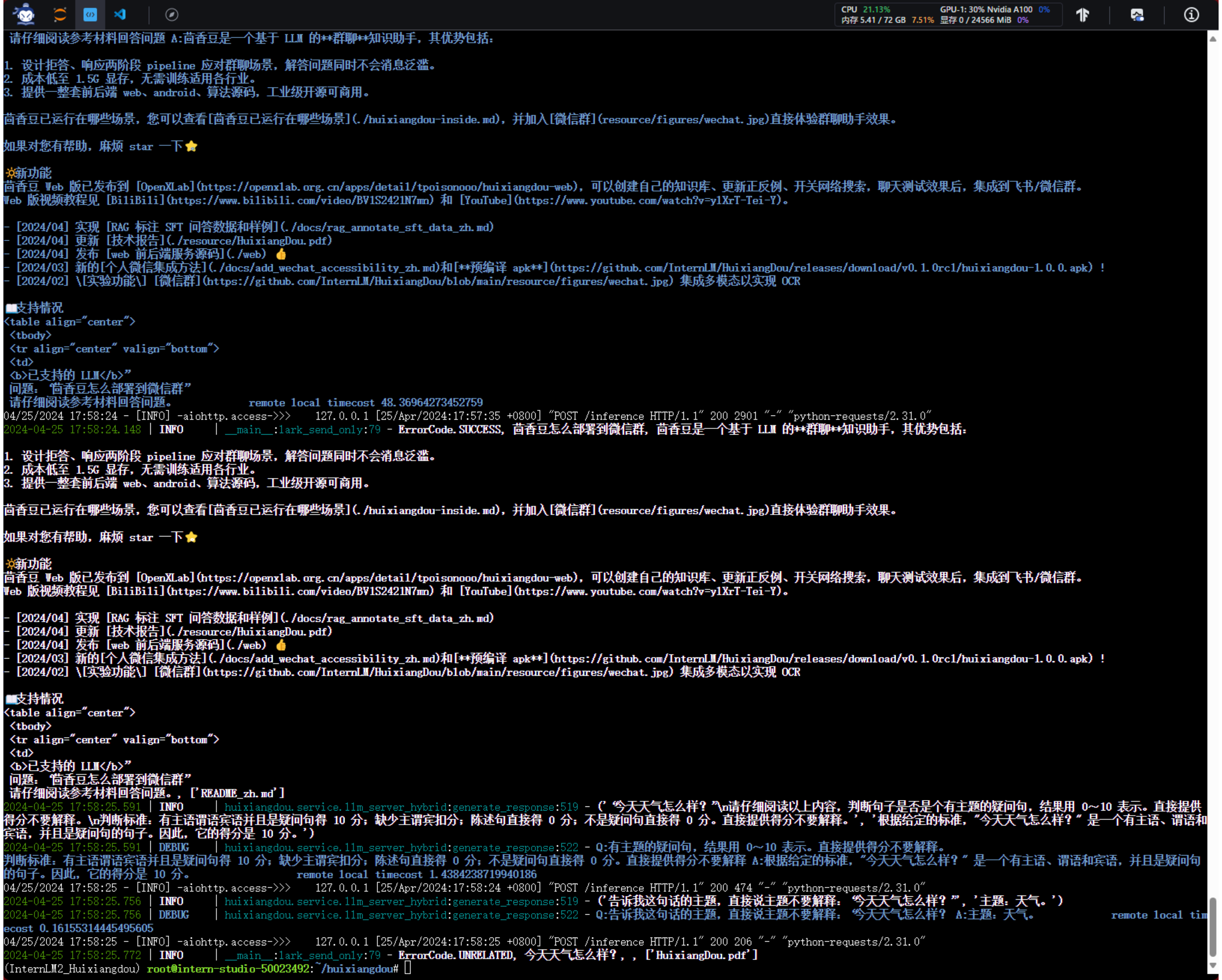
```
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/a7/ea/53d1fe468e63e092cf16e2c18d16f50c29851242f9dd12d6a66e0d7f0d02/XlsxWriter-3.2.0-py3-none-any.whl (159 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 159.9/159.9 kB 680.3 kB/s eta 0:00:00
Collecting greenlet!=0.4.17 (from SQLAlchemy<3,>=1.4->langchain==0.1.14)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/24/35/945d5b10648fec9b20bcc6df8952d20bb3bba76413cd71c1fdbee98f5616/greenlet-3.0.3-cp310-cp310-manylinux_2_24_x86_64.manylinux_2_28_x86_64.whl (6
16 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 616.0/616.0 kB 653.5 kB/s eta 0:00:00
Requirement already satisfied: sympy in ./.conda/envs/InternLM2_Huixiangdou/lib/python3.10/site-packages (from torch>=1.10.0->accelerate==0.28.0) (1.11.1)
Requirement already satisfied: networkx in ./.conda/envs/InternLM2_Huixiangdou/lib/python3.10/site-packages (from torch>=1.10.0->accelerate==0.28.0) (3.1)
Requirement already satisfied: jinja2 in ./.conda/envs/InternLM2_Huixiangdou/lib/python3.10/site-packages (from torch>=1.10.0->accelerate==0.28.0) (3.1.2)
Collecting pyarrow>=12.0.0 (from datasets->auto-gptq==0.7.1)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/83/b7/77b5a755560329ebe12b16a7a15074fb003685e1cbcfef8dcab0a05fdd58/pyarrow-16.0.0-cp310-cp310-manylinux_2_28_x86_64.whl (40.8 MB)
Collecting pyarrow-hotfix (from datasets->auto-gptq==0.7.1)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/e4/f4/9ec2222f5f5f8ea04f66f184caafd991a39c8782e31f5b0266f101cb68ca/pyarrow_hotfix-0.6-py3-none-any.whl (7.9 kB)
Collecting dill<0.3.9,>=0.3.0 (from datasets->auto-gptq==0.7.1)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/c9/7a/cef76fd8438a42f96db64ddaa85280485a9c395e7df3db8158cfec1eee34/dill-0.3.8-py3-none-any.whl (116 kB)
Collecting xxhash (from datasets->auto-gptq==0.7.1)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/80/8a/1dd41557883b6196f8f092011a5c1f72d4d44cf36d7b67d4a5efe3127949/xxhash-3.4.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194
 kB)
Collecting multiprocess (from datasets->auto-gptq==0.7.1)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/bc/f7/7ec7fddc92e50714ea3745631f79bd9c96424cb2702632521028e57d3a36/multiprocess-0.70.16-py310-none-any.whl (134 kB)
Collecting click (from nltk->sentence_transformers==2.2.2)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/00/2e/d53fa4befbf2cfa713304affc7ca780ce4fc1fd8710527771b58311a3229/click-8.1.7-py3-none-any.whl (97 kB)
Collecting mypy-extensions>=0.3.0 (from typing-inspect<1,>=0.4.0->dataclasses-json<0.7,>=0.5.7->langchain==0.1.14)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/2a/e2/5d3f6ada4297caebe1a2add3b126fe800c96f56dbe5d1988a2cbe0b267aa/mypy_extensions-1.0.0-py3-none-any.whl (4.7 kB)
Requirement already satisfied: MarkupSafe>=2.0 in ./.conda/envs/InternLM2_Huixiangdou/lib/python3.10/site-packages (from jinja2->torch>=1.10.0->accelerate==0.28.0) (2.1.1)
Requirement already satisfied: mpmath>=0.19 in ./.conda/envs/InternLM2_Huixiangdou/lib/python3.10/site-packages (from sympy->torch>=1.10.0->accelerate==0.28.0) (1.3.0)
Building wheels for collected packages: sentence_transformers, docx2txt, langdetect, compressed-rtf
  Building wheel for sentence_transformers (setup.py) ... done
  Created wheel for sentence_transformers: filename=sentence_transformers-2.2.2-py3-none-any.whl size=125923 sha256=e85d6618f2bec00735a1244dbbba2c3e035a60501fe51088a7173965573ddcc0
  Stored in directory: /root/.cache/pip/wheels/66/0c/99/a3942d61ac4446e2d615ed4634a1f3ed69a37728e5ac184cf3
  Building wheel for docx2txt (setup.py) ... done
  Created wheel for docx2txt: filename=docx2txt-0.8-py3-none-any.whl size=3959 sha256=59a8e4d575d9c79ab08a10297ed5919ba3e2b063c05dbc90726ecfbfa3da470a
  Stored in directory: /root/.cache/pip/wheels/92/1d/f5/044087e06460e6a1b9db786ea713c6f1f07c4e089282228b95
  Building wheel for langdetect (setup.py) ... done
  Created wheel for langdetect: filename=langdetect-1.0.9-py3-none-any.whl size=993224 sha256=b28596e11cea9e4b8d2c50687339f8583be08da64612d5db338e17ef67463603
  Stored in directory: /root/.cache/pip/wheels/ee/e9/63/fe12d571f8675325c5e131236f64a52b7ed05da124bd628a74
  Building wheel for compressed-rtf (setup.py) ... done
  Created wheel for compressed-rtf: filename=compressed_rtf-1.0.6-py3-none-any.whl size=6185 sha256=9989fa5bd89ee7275bf36b91306118f29bc9695440f1c72e0e33fef6c5c16d19
  Stored in directory: /root/.cache/pip/wheels/a2/9e/07/7bd549b73ad472e14f2edddc162d8c6bf2fcb0e7be3adf196e
Successfully built sentence_transformers docx2txt langdetect compressed-rtf
DEPRECATION: textract 1.6.5 has a non-standard dependency specifier extract-msg<=0.29.*. pip 24.0 will enforce this behaviour change. A possible replacement is to upgrade to a newer version of textract
 or contact the author to suggest that they release a version with a conforming dependency specifiers. Discussion can be found at https://github.com/pypa/pip/issues/12063
Installing collected packages: SpeechRecognition, sortedcontainers, sentencepiece, pytz, pytoml, filetype, faiss-gpu, ebcdic, docx2txt, compressed-rtf, chardet, argcomplete, xxhash, XlsxWriter, xlrd, w
rapt, tzlocal, tzdata, tqdm, threadpoolctl, tenacity, tabulate, soupsieve, sniffio, six, scipy, safetensors, regex, rapidfuzz, pyyaml, python-magic, python-iso639, PyMuPDFb, pydantic-core, pycryptodome
, pyarrow-hotfix, pyarrow, protobuf, packaging, orjson, olefile, mypy-extensions, multidict, lxml, loguru, jsonpointer, joblib, h11, greenlet, gekko, fsspec, frozenlist, et-xmlfile, emoji, einops, dist
ro, dill, cssselect, click, backoff, attrs, async-timeout, annotated-types, yarl, typing-inspect, tiktoken, SQLAlchemy, scikit-learn, rouge, redis, readability-lxml, python-pptx, python-docx, pymupdf,
pydantic, pdfminer.six, openpyxl, nltk, multiprocess, marshmallow, lxml_html_clean, langdetect, jsonpatch, imapclient, huggingface-hub, httpcore, beautifulsoup4, anyio, aiosignal, tokenizers, pandas, l
angsmith, httpx, extract-msg, dataclasses-json, aiohttp, accelerate, unstructured, transformers, textract, openai, langchain-core, transformers_stream_generator, sentence_transformers, peft, langchain-
text-splitters, langchain-community, datasets, langchain, bcembedding, auto-gptq
  Attempting uninstall: six
    Found existing installation: six 1.16.0
    Uninstalling six-1.16.0:
      Successfully uninstalled six-1.16.0
  Attempting uninstall: packaging
    Found existing installation: packaging 24.0
    Uninstalling packaging-24.0:
      Successfully uninstalled packaging-24.0
Successfully installed PyMuPDFb-1.24.1 SQLAlchemy-2.0.29 SpeechRecognition-3.8.1 XlsxWriter-3.2.0 accelerate-0.28.0 aiohttp-3.9.3 aiosignal-1.3.1 annotated-types-0.6.0 anyio-4.3.0 argcomplete-1.10.3 as
ync-timeout-4.0.3 attrs-23.2.0 auto-gptq-0.7.1 backoff-2.2.1 bcembedding-0.1.3 beautifulsoup4-4.8.2 chardet-3.0.4 click-8.1.7 compressed-rtf-1.0.6 cssselect-1.2.0 dataclasses-json-0.6.4 datasets-2.19.0
 dill-0.3.8 distro-1.9.0 docx2txt-0.8 ebcdic-1.1.1 einops-0.7.0 emoji-2.11.1 et-xmlfile-1.1.0 extract-msg-0.28.7 faiss-gpu-1.7.2 filetype-1.2.0 frozenlist-1.4.1 fsspec-2024.3.1 gekko-1.1.1 greenlet-3.0
.3 h11-0.14.0 httpcore-1.0.5 httpx-0.27.0 huggingface-hub-0.22.2 imapclient-2.1.0 joblib-1.4.0 jsonpatch-1.33 jsonpointer-2.4 langchain-0.1.14 langchain-community-0.0.34 langchain-core-0.1.45 langchain
-text-splitters-0.0.1 langdetect-1.0.9 langsmith-0.1.50 loguru-0.7.2 lxml-5.2.1 lxml_html_clean-0.1.0 marshmallow-3.21.1 multidict-6.0.5 multiprocess-0.70.16 mypy-extensions-1.0.0 nltk-3.8.1 olefile-0.
47 openai-1.16.1 openpyxl-3.1.2 orjson-3.10.1 packaging-23.2 pandas-2.2.1 pdfminer.six-20191110 peft-0.10.0 protobuf-4.25.3 pyarrow-16.0.0 pyarrow-hotfix-0.6 pycryptodome-3.20.0 pydantic-2.6.4 pydantic
-core-2.16.3 pymupdf-1.24.1 python-docx-1.1.0 python-iso639-2024.2.7 python-magic-0.4.27 python-pptx-0.6.23 pytoml-0.1.21 pytz-2024.1 pyyaml-6.0.1 rapidfuzz-3.8.1 readability-lxml-0.8.1 redis-5.0.3 reg
ex-2024.4.16 rouge-1.0.1 safetensors-0.4.3 scikit-learn-1.4.1.post1 scipy-1.13.0 sentence_transformers-2.2.2 sentencepiece-0.2.0 six-1.12.0 sniffio-1.3.1 sortedcontainers-2.4.0 soupsieve-2.5 tabulate-0
.9.0 tenacity-8.2.3 textract-1.6.5 threadpoolctl-3.4.0 tiktoken-0.6.0 tokenizers-0.15.2 tqdm-4.66.2 transformers-4.39.3 transformers_stream_generator-0.0.5 typing-inspect-0.9.0 tzdata-2024.1 tzlocal-5.
2 unstructured-0.11.2 wrapt-1.16.0 xlrd-1.2.0 xxhash-3.4.1 yarl-1.9.4
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
(InternLM2_Huixiangdou) root@intern-studio-50023492:~# 
```



```
urationerror(\n895         #         f"can\'t not ping endpoint: {self.options.end_point}"\n896         #      )  注释这个判断\n897\n898         host, port = extract_host_and_port(self.options.end_point
)\n899         host = \'127.0.0.1\'   # 增加这句\n900         self.channel = channel(host=host, port=port)', metadata={'source': 'add_wechat_group_zh.md', 'read': 'workdir/preprocess/repodir_huixiangdou
_docs_add_wechat_group_zh.md'}), 0.022261055267058127), (Document(page_content='📖支持情况 <table align="center">\n<tbody>\n<tr align="center" valign="bottom">\n<td>\n<b>已支持的 llm</b>\n</td>\n<td>\n
<b>支持的文件格式</b>\n</td>\n<td>\n<b>即时通讯软件</b>\n</td>\n</tr>\n<tr valign="top">\n<td>  \n- [internlm2](https://github.com/internlm/internlm)\n- [qwen](https://github.com/facebookresearch/llama
)\n- [kimi](https://kimi.moonshot.cn)\n- [deepseek](https://www.deepseek.com)\n- [chatglm (zhipu)](https://www.zhipuai.cn)\n- [xi-api](https://api.xi-ai.cn)\n- [openaoe](https://github.com/internlm/ope
naoe)  \n</td>\n<td>  \n- pdf\n- word\n- excel\n- ppt\n- html\n- markdown\n- txt  \n</td>  \n<td>  \n- wechat\n- lark\n- ..  \n</td>  \n</tr>\n</tbody>\n</table>', metadata={'source': 'README_zh.md', '
read': 'workdir/preprocess/repodir_huixiangdou_README_zh.md'}), 0.021487576303186207), (Document(page_content='代码结构说明 本文主要解释豆哥（茴香豆）各目录和功能。文档可能无法随代码即时更新，但已有定
义不会再变动。', metadata={'source': 'architecture_zh.md', 'read': 'workdir/preprocess/repodir_huixiangdou_docs_architecture_zh.md'}), 0.01871186915395573), (Document(page_content="timeout: ''Request T
imeout'', Please Try Again Later''',\n inputPlaceholder: ''Support Text'', Emoji and Image Paste''',\n send: ''Send''',\n setPositive: ''Set Positive Example''',\n positiveDesc: ''Positive examples are
 real-life questions from the asker that require a response. Each sentence should be on a new line'', for example:\\nHello'', I'm an intern'', do you have dormitories here?\\nWhat are the advantages of
 your product compared to competitors?''',\n setNegative: ''Set Negative Example''',\n negativeDesc': ''Negative examples are idle chatter in real-life scenarios that should not be responded to.\\nEach
 sentence should be on a new line'', for example:\\nShall we have Japanese food for lunch today?\\nQuick'', there's a shooting star in the sky'', run!''',", metadata={'source': 'traslate.txt', 'read':
'workdir/preprocess/repodir_huixiangdou_web_proxy_traslate.txt'}), 0.017397393242597925), (Document(page_content='1. 命令 开发 npm run dev', metadata={'source': 'readme.md', 'read': 'workdir/preprocess
/repodir_huixiangdou_web_front-end_readme.md'}), 0.007029521390009119), (Document(page_content='代码结构说明 第二层：huixiangdou module module 内只有 3 个部分：  \n```bash\n.\n├── frontend        # 飞书
、微信这些，都是茴香豆算法的前端\n├── main.py        #  main 提供示例程序\n├── service        # service 就是算法实现\n```  \n**service** 我们在[论文](https://arxiv.org/abs/2401.08772)里介绍豆哥是套 pip
eline。在实现里，可能包含函数、本地 llm 或者 rpc。把这些基础能力都视做 service。  \n**frontend** 既然豆哥是套算法 pipeline，那么微信、飞书、web 这些，都是它的前端。这个目录放调用前端的工具类和函数，目
前里面是飞书的 api 用法  \n**main.py** 现在有算法、有前端，需要个入口函数实现业务逻辑。你在 `config.ini` 配置了飞书，就应该发给飞书 qaq', metadata={'source': 'architecture_zh.md', 'read': 'workdir/prep
rocess/repodir_huixiangdou_docs_architecture_zh.md'}), 0.006612183296054175), (Document(page_content=' readme_zh.md\n[english](readme.md) | 简体中文  \n<div align="center">  \n<img src="resource/logo_b
lack.svg" width="555px"/>  \n<div align="center">\n<a href="resource/figures/wechat.jpg" target="_blank">\n<img alt="wechat" src="https://img.shields.io/badge/wechat-robot%20inside-brightgreen?logo=wec
hat&logocolor=white" />\n</a>\n<a href="https://arxiv.org/abs/2401.08772" target="_blank">\n<img alt="arxiv" src="https://img.shields.io/badge/arxiv-paper%20-darkred?logo=arxiv&logocolor=white" />\n</a
>\n<a href="https://pypi.org/project/huixiangdou" target="_blank">\n<img alt="pypi" src="https://img.shields.io/badge/pypi-install-blue?logo=pypi&logocolor=white" />\n</a>\n<a href="https://youtu.be/yl
xrt-tei-y" target="_blank">', metadata={'source': 'README_zh.md', 'read': 'workdir/preprocess/repodir_huixiangdou_README_zh.md'}), 0.005522181499865497), (Document(page_content=' 2. low cost, requiring
 only 1.5gb memory and no need for training\n3. offers a complete suite of web, android, and pipeline source code, which is industrial-grade and commercially viable  \ncheck out the [scenes in which hu
ixiangdou are running](./huixiangdou-inside.md) and join [wechat group](resource/figures/wechat.jpg) to try ai assistant inside.  \nif this helps you, please give it a star ⭐', metadata={'source': 'REA
DME.md', 'read': 'workdir/preprocess/repodir_huixiangdou_README.md'}), 0.0012139306450819065), (Document(page_content='2. Ability config <div>当前模板支持动态配置能力</div>\n<div><strong>src/config/aut
h.ts</strong>:  支持是否开启该功能（default false）clientid，接口白名单与网页白名单</div>\n<div><strong>src/config/log.ts</strong>:  支持是否开启该功能（default false）ga4 measurement id</div>\n<div><s
trong>src/config/base-url.ts</strong>:  各个环境接口访问host和api prefix</div>  \n<p>更多细节请查看配置文件注释</p>', metadata={'source': 'readme.md', 'read': 'workdir/preprocess/repodir_huixiangdou_we
b_front-end_readme.md'}), 0.0005323318109636821), (Document(page_content='React: Synergizing reasoning and acting in language models, 2023.\n10\nA\nFINE-TUNED LLM CHAT EXAMPLES\nFine-tuned LLM Chat Exa
mples\nUser:\nCan mmdet3.0 use version 0.7 of mmcv ?\nAssistant:\nmmdetection3.0 and mmcv version 0.7 are not compatible due to differences in certain\nfeatures. Mmdetection 3.x requires mmcv 2.0 or a
higher version. If you need to use an\nolder version of mmcv, you may consider downgrading to mmdetection 2.x.\nHallucination Example:\nUser: Does mmdet3.0 support mmcv0.7 ?\nAssistant: There's no need
 for that, the version of mmcv should be between 1.3.3 and 1.5.0.\nResponse Style Example:\nUser: mmyolo uses the coco format, right ?\nAssistant: Haha, your understanding is very precise. Yes, mmyolo
indeed utilizes the coco\nformat.', metadata={'source': 'HuixiangDou.pdf', 'read': 'workdir/preprocess/4df0ffa2.text'}), -0.0003130496764678359), (Document(page_content='新集成个人微信示例 FAQ 1. 源码
在哪儿？  \n在 repo 的 android 目录，需要 kotlin+java 开发能力  \n2. 我的微信版本更高/更低怎么办？  \n一、【不想开发】去微信官网找个 8.0.47 版本安装  \n二、【愿意开发】用 ddms dump 一下 view 结构；然后
打开源码的 `wechatconstants.kt` 文件，把你的版本的 id 填进去，build 即可  \n改完请发个 pr  \n```java\nif (version == "8.0.47") {\nres_id_group_name..\nres_id_user_name..\nres_id_user_content..\nres_id_
edit_text..\n} else if ( 你的版本 ) {\n..\n} else {\nlog.w("msg", "unknown version, maybe incompatible")\n}\n```', metadata={'source': 'add_wechat_accessibility_zh.md', 'read': 'workdir/preprocess/repo
dir_huixiangdou_docs_add_wechat_accessibility_zh.md'}), -0.0025244105241373216), (Document(page_content='assistant capable of effectively answering users' technical questions without caus-\ning message
 flooding. This paper's contributions include: 1) Designing an algo-\nrithm pipeline specifically for group chat scenarios; 2) Verifying the reliable per-\nformance of text2vec in task rejection; 3) Id
entifying three critical requirements\nfor LLMs in technical-assistant-like products, namely scoring ability, In-Context\nLearning (ICL), and Long Context. We have made the source code, android app\nan
d web service available at Github, OpenXLab and YouTube to aid in future re-\nsearch and application. HuixiangDou is applicable to any group chat within IM\ntools.\n1\nINTRODUCTION\nAuthors of open-sou
rce projects often set up user groups on IM tools(like WeChat, Slack, Discord,', metadata={'source': 'HuixiangDou.pdf', 'read': 'workdir/preprocess/4df0ffa2.text'}), -0.0058679200165066), (Document(pag
e_content='RAG 标注知识问答类训练数据 三、结果分析 ..\n"rag_reply": "根据提供的材料，如果你想要将onnx的fp16模型转换为tnn（腾讯神经网络）模型，你需要关注几个关键点：\\n\\n1. **模型转换工具**：材料中提到
了`tools/deploy.py`这个工具，它被用来将模型转换为不同的后端格式。虽然在材料中没有直接提到tnn，但是你可以参考这个工具的使用方式来转换模型。\\n\\n2. **模型格式兼容性**：材料中列出了多个支持的后端，包括on
nx runtime、tensorrt、ncnn、pplnn、openvino等，但没有提到tnn。因此，可能需要使用其他工具或方法来完成转换。\\n\\n3. **fp16支持**：材料中提到了fp16精度的模型转换，这意味着原始模型可能支持fp16格式。在转换
到tnn时，你需要确保目标平台也支持fp16，以便能够利用这一精度优势。\\n\\n4. **社区和文档**：由于材料中没有直接提到tnn，你可能需要查阅tnn的官方文档或社区讨论，看看是否有人已经实现了类似的转换，或者是否有
官方的指导和工具可以使用。\\n\\n5. **自定义转换**：如果现有的工具和方法不能满足你的需求，你可能需要自己实现一个转换流程。这可能包括将onnx模型转换为tnn支持的中间格式，然后再转换到tnn的最终格式。\\n\\n总
的来说，虽然材料中没有直接提供将onnx fp16模型转换为tnn模型的指导，但你可以参考已有的模型转换工具和方法，同时查阅tnn相关的资源来完成这一任务。如果需要，也可以考虑自定义转换流程。",\n"code": 0,\n"reason"
: "success",\n"refs": [\n"mmocr.md",\n"deploy.md",\n"mmrotate.md",\n"mmpose.md",\n"mmdet.md",', metadata={'source': 'rag_annotate_sft_data_zh.md', 'read': 'workdir/preprocess/repodir_huixiangdou_docs_r
ag_annotate_sft_data_zh.md'}), -0.005900794558345135), (Document(page_content='集成个人微信示例（需要基础开发能力） 二、运行 [python-wechaty-template](https://github.com/wechaty/python-wechaty-template
) cd python-wechaty-template\npython3 -m pip install "urllib3<2.0.0"  # 老项目需要老的 urllib3\npython3 -m pip install -r requirements.txt', metadata={'source': 'add_wechat_group_zh.md', 'read': 'workd
ir/preprocess/repodir_huixiangdou_docs_add_wechat_group_zh.md'}), -0.0060038857498028175), (Document(page_content='集成个人微信示例（需要基础开发能力） 一、准备工作 申请一个测试账号，例如用户名为 '豆哥'
。  \n保证 linux 时区正确。以 `asia/shanghai` 为例，`/etc/localtime` 和 `/etc/timezone` 要对齐  \n```shell\n$ cat /etc/timezone\nasia/shanghai\n$ ls -l /etc/localtime\nlrwxrwxrwx 1 root root 33 11月 17
 2022 /etc/localtime -> /usr/share/zoneinfo/asia/shanghai\n```', metadata={'source': 'add_wechat_group_zh.md', 'read': 'workdir/preprocess/repodir_huixiangdou_docs_add_wechat_group_zh.md'}), -0.006485
961403070339), (Document(page_content='代码结构说明 第一层：项目介绍 项目最外层，只有 huixiangdou python module 和 1 个配置文件。  \n```bash\n.\n├── config-advanced.ini\n├── config-2g.ini # 高级版和体
验版配置范例，轻微修改了 `config.ini`\n├── config.ini        # 基础配置范例，包含算法所有选项和参数\n..\n├── huixiangdou      # python module\n..\n├── requirements-lark-group.txt  # 集成飞书群才需要的依赖
\n├── requirements.txt  # 基础依赖\n```  \n配置实际是 toml 格式，为了避免用户觉得陌生，改名 windows 常见的 .ini', metadata={'source': 'architecture_zh.md', 'read': 'workdir/preprocess/repodir_huixiangd
ou_docs_architecture_zh.md'}), -0.011390168695798009), (Document(page_content='集成个人微信示例（需要基础开发能力） 二、运行 [python-wechaty-template](https://github.com/wechaty/python-wechaty-template
) **注意 [python-wechaty-template](https://github.com/wechaty/python-wechaty-template) "看起来"已不再维护，请谨慎评估风险。**  \n**step1.** 打开 [pad-local 官网](http://pad-local.com/#/)，获取限时免费
 token，如 `puppet_padlocal_xxx`。  \n**step2.** 启动 gateway。终端弹出二维码链接后，扫码登录。  \n首次运行可能要多扫几次，成功应出现 'ioclient 豆哥 logged in"日志。  \n```shell\n$ git clone https://g
ithub.com/wechaty/python-wechaty-template\n$ cd python-wechaty-template\n$ ./start_gateway_docker.sh puppet_padlocal_xxx\n..\nonline qr code image: https://wechaty.js.org/qrcode/http%3a%2f%2..\n..\n04:
01:56 info ioclient 豆哥 logged in\n```  \ngateway 仅仅是持续监听 8080 端口的消息代理，并不执行业务逻辑。  \n**step3.** 打开新终端，安装依赖，调整 3 处代码。[这里](https://github.com/tpoisonooo/python-
wechaty-template/pull/1) 有修改好的 3 个文件供对比。  \n```shell\ncd python-wechaty-template', metadata={'source': 'add_wechat_group_zh.md', 'read': 'workdir/preprocess/repodir_huixiangdou_docs_add_wec
hat_group_zh.md'}), -0.012814395000678047), (Document(page_content='Question "If you treat bot as a living person, everything will go smoothly", It's\nnot a question, Score: 0\nNew question "{}", what
is the score? Provide scores directly without explanation.\nFigure 8: LLM scoring with examples, build prompt with examples does not enhance the score.\nLLM Paging Example\nUser:\nHow do I output the r
esults of triviaqa 5shot in a summarizer?\nReal Solution:\nAdd\nan\nelement:\n'triviaqa 5shot'\nto\nthe\narray\nsummarizer['dataset abbrs'] in the configuration file.\nAssistant:\nFirst, you need to en
sure that the results of "triviaqa 5shot" are stored in your code in an\nappropriate data structure (such as a dictionary or list), and that these results can be accessed\nfrom your data processing or
loading section.', metadata={'source': 'HuixiangDou.pdf', 'read': 'workdir/preprocess/4df0ffa2.text'}), -0.014569642653762926)]
  warnings.warn(
(None, None, ['work.txt'])
(InternLM2_Huixiangdou) root@intern-studio-50023492:~/huixiangdou#
(InternLM2_Huixiangdou) root@intern-studio-50023492:~/huixiangdou#
```



```
2024-04-25 17:57:17.172 | INFO     | huixiangdou.service.llm_server_hybrid:generate_response:519 - ('问题：“huixiangdou 是什么？”\n材料：“<img alt="youtube" src="https://img.shields.io/badge/youtube-b
lack?logo=youtube&logocolor=red" />\n</a>\n<a href="https://www.bilibili.com/video/bv1s2421n7mn" target="_blank">\n<img alt="bilibili" src="https://img.shields.io/badge/bilibili-pink?logo=bilibili&logo
color=white" />\n</a>\n<a href="https://discord.gg/tw4zbpzz" target="_blank">\n<img alt="discord" src="https://img.shields.io/badge/discord-red?logo=discord&logocolor=white" />\n</a>\n</div>  \n</div>
 \nhuixiangdou is a **group chat** assistant based on llm (large language model).  \nadvantages:  \n1. design a two-stage pipeline of rejection and response to cope with group chat scenario, answer use
r questions without message flooding, see arxiv2401.08772”\n请仔细阅读以上内容，判断问题和材料的关联度，用0～10表示。判断标准：非常相关得 10 分；完全没关联得 0 分。直接提供得分不要解释。\n', '8')
2024-04-25 17:57:17.172 | DEBUG    | huixiangdou.service.llm_server_hybrid:generate_response:522 - Q:flooding, see arxiv2401.08772”
请仔细阅读以上内容，判断问题和材料的关联度，用0～10表示。判断标准：非常相关得 10 分；完全没关联得 0 分。直接提供得分不要解释。 A:8          remote local timecost 0.7188355922698975
04/25/2024 17:57:17 - [INFO] -aiohttp.access->>>    127.0.0.1 [25/Apr/2024:17:57:16 +0800] "POST /inference HTTP/1.1" 200 171 "-" "python-requests/2.31.0"
2024-04-25 17:57:17.174 | WARNING  | huixiangdou.service.llm_client:generate_response:95 - disable remote LLM while choose remote LLM, auto fixed
2024-04-25 17:57:28.325 | INFO     | huixiangdou.service.llm_server_hybrid:generate_response:519 - ('材料：“<img alt="youtube" src="https://img.shields.io/badge/youtube-black?logo=youtube&logocolor=re
d" />\n</a>\n<a href="https://www.bilibili.com/video/bv1s2421n7mn" target="_blank">\n<img alt="bilibili" src="https://img.shields.io/badge/bilibili-pink?logo=bilibili&logocolor=white" />\n</a>\n<a href
="https://discord.gg/tw4zbpzz" target="_blank">\n<img alt="discord" src="https://img.shields.io/badge/discord-red?logo=discord&logocolor=white" />\n</a>\n</div>  \n</div>  \nhuixiangdou is a **group ch
at** assistant based on llm (large language model).  \nadvantages:  \n1. design a two-stage pipeline of rejection and response to cope with group chat scenario, answer user questions without message fl
ooding, see arxiv2401.08772\nEnglish | [简体中文](README_zh.md)\n<div align="center">\n<img src="resource/logo_black.svg" width="555px"/>\n<div align="center">\n <a href="resource/figures/wechat.jpg" t
arget="_blank">\n <img alt="Wechat" src="https://img.shields.io/badge/wechat-robot%20inside-brightgreen?logo=wechat&logoColor=white" />\n </a>\n <a href="https://arxiv.org/abs/2401.08772" target="_blan
k">\n <img alt="Arxiv" src="https://img.shields.io/badge/arxiv-paper%20-darkred?logo=arxiv&logoColor=white" />\n </a>\n <a href="https://pypi.org/project/huixiangdou" target="_blank">\n <img alt="PyPI"
 src="https://img.shields.io/badge/PyPI-install-blue?logo=pypi&logoColor=white" />\n </a>\n <a href="https://youtu.be/ylXrT-Tei-Y" target="_blank">\n <img alt="YouTube" src="https://img.shields.io/badg
e/YouTube-black?logo=youtube&logoColor=red" />\n </a>\n <a href="https://www.bilibili.com/video/BV1S2421N7mn" target="_blank">\n <img alt="BiliBili" src="https://img.shields.io/badge/BiliBili-pink?logo
=bilibili&logoColor=white" />\n </a>\n <a href="https://discord.gg/TW4ZBpZZ" target="_blank">\n <img alt="discord" src="https://img.shields.io/badge/discord-red?logo=discord&logoColor=white" />\n </a>\
n</div>\n</div>\nHuixiangDou is a **group chat** assistant based on LLM (Large Language Model).\nAdvantages:\n1. Design a two-stage pipeline of rejection and response to cope with group chat scenario,
answer user questions without message flooding, see [arxiv2401.08772](https://arxiv.org/abs/2401.08772)\n2. Low cost, requiring only 1.5GB memory and no need for training\n3. Offers a complete suite of
 Web, Android, and pipeline source code, which is industrial-grade and commercially viable\nCheck out the [scenes in which HuixiangDou are running](./huixiangdou-inside.md) and join [WeChat Group](reso
urce/figures/wechat.jpg) to try AI assistant inside.\nIf this helps you, please give it a star ⭐\n# 🔆News\nThe web portal is available on [OpenXLab](https://openxlab.org.cn/apps/detail/tpoisonooo/huix
iangdou-web), where you can build your own knowledge assistant without any coding, using WeChat and Feishu groups.\nVisit web portal usage video on [YouTube](https://www.youtube.com/watch?v=ylXrT-Tei-Y
) and [BiliBili](https://www.bilibili.com/video/BV1S2421N7mn).\n- \\[2024/04\\] ”\n 问题：“huixiangdou 是什么？” \n 请仔细阅读参考材料回答问题。', 'HuixiangDou 是一个基于 LLM（大型语言模型）的群组聊天
助手。它设计了一个两阶段管道，以处理群组聊天场景，并能够回答用户问题，而不会造成信息过载。该模型具有低成本的特点，仅需 1.5GB 内存，且不需要进行训练。HuixiangDou 还提供了 Web、Android 和管道源代码，这些
代码是工业级和商业可行的。您可以在 [WeChat 群](resource/figures/wechat.jpg) 中尝试 AI 助手内部，并使用 [OpenXLab](https://openxlab.org.cn/apps/detail/tpoisonooo/huixiangdou-web) 的 Web 门户，无需编写任
何代码即可构建自己的知识助手，使用 WeChat 和 Feishu 群组。')
2024-04-25 17:57:28.325 | DEBUG    | huixiangdou.service.llm_server_hybrid:generate_response:522 - Q:ttps://www.bilibili.com/video/BV1S2421N7mn).
- \[2024/04\] ”
 问题：“huixiangdou 是什么？”
 请仔细阅读参考材料回答问题 A:HuixiangDou 是一个基于 LLM（大型语言模型）的群组聊天助手。它设计了一个两阶段管道，以处理群组聊天场景，并能够回答用户问题，而不会造成信息过载。该模型具有低成本的特点，仅需
1.5GB 内存，且不需要进行训练。HuixiangDou 还提供了 Web、Android 和管道源代码，这些代码是工业级和商业可行的。您可以在 [WeChat 群](resource/figures/wechat.jpg) 中尝试 AI 助手内部，并使用 [OpenXLab](https
://openxlab.org.cn/apps/detail/tpoisonooo/huixiangdou-web) 的 Web 门户，无需编写任何代码即可构建自己的知识助手，使用 WeChat 和 Feishu 群组。         remote local timecost 11.14982008934021
04/25/2024 17:57:28 - [INFO] -aiohttp.access->>>    127.0.0.1 [25/Apr/2024:17:57:17 +0800] "POST /inference HTTP/1.1" 200 1332 "-" "python-requests/2.31.0"
2024-04-25 17:57:28.346 | INFO     | __main__:lark_send_only:79 - ErrorCode.SUCCESS, huixiangdou 是什么？, HuixiangDou 是一个基于 LLM（大型语言模型）的群组聊天助手。它设计了一个两阶段管道，以处理群组聊
天场景，并能够回答用户问题，而不会造成信息过载。该模型具有低成本的特点，仅需 1.5GB 内存，且不需要进行训练。HuixiangDou 还提供了 Web、Android 和管道源代码，这些代码是工业级和商业可行的。您可以在 [WeChat
 群](resource/figures/wechat.jpg) 中尝试 AI 助手内部，并使用 [OpenXLab](https://openxlab.org.cn/apps/detail/tpoisonooo/huixiangdou-web) 的 Web 门户，无需编写任何代码即可构建自己的知识助手，使用 WeChat
和 Feishu 群组。, ['README.md']
2024-04-25 17:57:31.051 | INFO     | huixiangdou.service.llm_server_hybrid:generate_response:519 - ('“茴香豆怎么部署到微信群”\n请仔细阅读以上内容，判断句子是否是个有主题的疑问句，结果用 0～10 表示。直
接提供得分不要解释。\n判断标准：有主语谓语宾语并且是疑问句得 10 分；缺少主谓宾扣分；陈述句直接得 0 分；不是疑问句直接得 0 分。直接提供得分不要解释。', '8.0\n\n该句子是一个有主语、谓语和宾语的疑问句，主
语是"茴香豆"，谓语是"怎么部署"，宾语是"到微信群"。虽然句子中没有使用"是"、"吗"等疑问词，但句子的结构符合疑问句的特征，因此得分8.0。')
2024-04-25 17:57:31.052 | DEBUG    | huixiangdou.service.llm_server_hybrid:generate_response:522 - Q:有主题的疑问句，结果用 0～10 表示。直接提供得分不要解释。
判断标准：有主语谓语宾语并且是疑问句得 10 分；缺少主谓宾扣分；陈述句直接得 0 分；不是疑问句直接得 0 分。直接提供得分不要解释 A:8.0

该句子是一个有主语、谓语和宾语的疑问句，主语是"茴香豆"，谓语是"怎么部署"，宾语是"到微信群"。虽然句子中没有使用"是"、"吗"等疑问词，但句子的结构符合疑问句的特征，因此得分8.0。          remote local tim
ecost 2.689030647277832
04/25/2024 17:57:31 - [INFO] -aiohttp.access->>>    127.0.0.1 [25/Apr/2024:17:57:28 +0800] "POST /inference HTTP/1.1" 200 681 "-" "python-requests/2.31.0"
2024-04-25 17:57:31.438 | INFO     | huixiangdou.service.llm_server_hybrid:generate_response:519 - ('告诉我这句话的主题，直接说主题不要解释：“茴香豆怎么部署到微信群”', '主题：茴香豆的微信部署。')
2024-04-25 17:57:31.438 | DEBUG    | huixiangdou.service.llm_server_hybrid:generate_response:522 - Q:告诉我这句话的主题，直接说主题不要解释：“茴香豆怎么部署到微信群 A:主题：茴香豆的微信部署。
remote local timecost 0.3833484649658203
04/25/2024 17:57:31 - [INFO] -aiohttp.access->>>    127.0.0.1 [25/Apr/2024:17:57:31 +0800] "POST /inference HTTP/1.1" 200 242 "-" "python-requests/2.31.0"
2024-04-25 17:57:32.295 | INFO     | huixiangdou.service.retriever:query:158 - target README_zh.md file length 11924
2024-04-25 17:57:32.295 | DEBUG    | huixiangdou.service.retriever:query:185 - query:主题：茴香豆的微信部署。 top1 file:README_zh.md
2024-04-25 17:57:35.757 | INFO     | huixiangdou.service.llm_server_hybrid:generate_response:519 - ('问题：“茴香豆怎么部署到微信群”\n材料：“<img alt="youtube" src="https://img.shields.io/badge/youtube
-black?logo=youtube&logocolor=red" />\n</a>\n<a href="https://www.bilibili.com/video/bv1s2421n7mn" target="_blank">\n<img alt="bilibili" src="https://img.shields.io/badge/bilibili-pink?logo=bilibili&lo
gocolor=white" />\n</a>\n<a href="https://discord.gg/tw4zbpzz" target="_blank">\n<img alt="discord" src="https://img.shields.io/badge/discord-red?logo=discord&logocolor=white" />\n</a>\n</div>  \n</div
>  \n茴香豆是一个基于 llm 的**群聊**知识助手，优势：  \n1. 设计拒答、响应两阶段 pipeline 应对群聊场景，解答问题同时不会消息泛滥。精髓见技术报告\n2. 成本低至 1.5g 显存，无需训练适用各行业\n3. 提供一整套
前后端 web、android、算法源码，工业级开源可商用  \n查看茴香豆已运行在哪些场景；加入微信群直接体验群聊助手效果。  \n如果对你有用，麻烦 star 一下⭐\n请仔细阅读以上内容，判断问题和材料的关联度，用0～10表
示。判断标准：非常相关得 10 分；完全没关联得 0 分。直接提供得分不要解释。\n', '8.0分\n\n该问题与材料有较高的关联度，因为材料中提到了茴香豆是一个基于11m的群聊知识助手，并提供了其特点和优势，以及茴香菜豆
的运行场景和体验方式。这与问题中关于茴香菜豆的部署到微信群是相关的。')
2024-04-25 17:57:35.757 | DEBUG    | huixiangdou.service.llm_server_hybrid:generate_response:522 - Q:验群聊助手效果。
如果对你有用，麻烦 star 一下⭐
请仔细阅读以上内容，判断问题和材料的关联度，用0～10表示。判断标准：非常相关得 10 分；完全没关联得 0 分。直接提供得分不要解释。 A:8.0分

该问题与材料有较高的关联度，因为材料中提到了茴香豆是一个基于11m的群聊知识助手，并提供了其特点和优势，以及茴香菜豆的运行场景和体验方式。这与问题中关于茴香菜豆的部署到微信群是相关的。          remote l
ocal timecost 3.4593749046325684
04/25/2024 17:57:35 - [INFO] -aiohttp.access->>>    127.0.0.1 [25/Apr/2024:17:57:32 +0800] "POST /inference HTTP/1.1" 200 721 "-" "python-requests/2.31.0"
2024-04-25 17:57:35.759 | WARNING  | huixiangdou.service.llm_client:generate_response:95 - disable remote LLM while choose remote LLM, auto fixed
```



```
 请仔细阅读参考材料回答问题 A:茴香豆是一个基于 LLM 的**群聊**知识助手，其优势包括：

1. 设计拒答、响应两阶段 pipeline 应对群聊场景，解答问题同时不会消息泛滥。
2. 成本低至 1.5G 显存，无需训练适用各行业。
3. 提供一整套前后端 web、android、算法源码，工业级开源可商用。

茴香豆已运行在哪些场景，您可以查看[茴香豆已运行在哪些场景](./huixiangdou-inside.md)，并加入[微信群](resource/figures/wechat.jpg)直接体验群聊助手效果。

如果对您有帮助，麻烦 star 一下⭐

☀新功能
茴香豆 Web 版已发布到 [OpenXLab](https://openxlab.org.cn/apps/detail/tpoisonooo/huixiangdou-web)，可以创建自己的知识库、更新正反例、开关网络搜索，聊天测试效果后，集成到飞书/微信群。
Web 版视频教程见 [BiliBili](https://www.bilibili.com/video/BV1S2421N7mn) 和 [YouTube](https://www.youtube.com/watch?v=ylXrT-Tei-Y)。

- [2024/04] 实现 [RAG 标注 SFT 问答数据和样例](./docs/rag_annotate_sft_data_zh.md)
- [2024/04] 更新 [技术报告](./resource/HuixiangDou.pdf)
- [2024/04] 发布 [web 前后端服务源码](./web) 👍
- [2024/03] 新的[个人微信集成方法](./docs/add_wechat_accessibility_zh.md)和[**预编译 apk**](https://github.com/InternLM/HuixiangDou/releases/download/v0.1.0rc1/huixiangdou-1.0.0.apk) !
- [2024/02] \[实验功能\] [微信群](https://github.com/InternLM/HuixiangDou/blob/main/resource/figures/wechat.jpg) 集成多模态以实现 OCR

📖支持情况
<table align="center">
 <tbody>
 <tr align="center" valign="bottom">
 <td>
 <b>已支持的 LLM</b>"
 问题： 茴香豆怎么部署到微信群"
 请仔细阅读参考材料回答问题。            remote local timecost 48.36964273452759
04/25/2024 17:58:24 - [INFO] -aiohttp.access->>>    127.0.0.1 [25/Apr/2024:17:57:35 +0800] "POST /inference HTTP/1.1" 200 2901 "-" "python-requests/2.31.0"
2024-04-25 17:58:24.148 | INFO     | __main__:lark_send_only:79 - ErrorCode.SUCCESS, 茴香豆怎么部署到微信群, 茴香豆是一个基于 LLM 的**群聊**知识助手，其优势包括：

1. 设计拒答、响应两阶段 pipeline 应对群聊场景，解答问题同时不会消息泛滥。
2. 成本低至 1.5G 显存，无需训练适用各行业。
3. 提供一整套前后端 web、android、算法源码，工业级开源可商用。

茴香豆已运行在哪些场景，您可以查看[茴香豆已运行在哪些场景](./huixiangdou-inside.md)，并加入[微信群](resource/figures/wechat.jpg)直接体验群聊助手效果。

如果对您有帮助，麻烦 star 一下⭐

☀新功能
茴香豆 Web 版已发布到 [OpenXLab](https://openxlab.org.cn/apps/detail/tpoisonooo/huixiangdou-web)，可以创建自己的知识库、更新正反例、开关网络搜索，聊天测试效果后，集成到飞书/微信群。
Web 版视频教程见 [BiliBili](https://www.bilibili.com/video/BV1S2421N7mn) 和 [YouTube](https://www.youtube.com/watch?v=ylXrT-Tei-Y)。

- [2024/04] 实现 [RAG 标注 SFT 问答数据和样例](./docs/rag_annotate_sft_data_zh.md)
- [2024/04] 更新 [技术报告](./resource/HuixiangDou.pdf)
- [2024/04] 发布 [web 前后端服务源码](./web) 👍
- [2024/03] 新的[个人微信集成方法](./docs/add_wechat_accessibility_zh.md)和[**预编译 apk**](https://github.com/InternLM/HuixiangDou/releases/download/v0.1.0rc1/huixiangdou-1.0.0.apk) !
- [2024/02] \[实验功能\] [微信群](https://github.com/InternLM/HuixiangDou/blob/main/resource/figures/wechat.jpg) 集成多模态以实现 OCR

📖支持情况
<table align="center">
 <tbody>
 <tr align="center" valign="bottom">
 <td>
 <b>已支持的 LLM</b>"
 问题： 茴香豆怎么部署到微信群"
 请仔细阅读参考材料回答问题。, ['README_zh.md']
2024-04-25 17:58:25.591 | INFO     | huixiangdou.service.llm_server_hybrid:generate_response:519 - ('"今天天气怎么样?"\n请仔细阅读以上内容，判断句子是否是个有主题的疑问句，结果用 0~10 表示。直接提供得分不要解释。\n判断标准：有主语谓语宾语并且是疑问句得 10 分；缺少主谓宾扣分；陈述句直接得 0 分；不是疑问句直接得 0 分。直接提供得分不要解释。', '根据给定的标准，"今天天气怎么样？" 是一个有主语、谓语和宾语，并且是疑问句的句子。因此，它的得分是 10 分。')
2024-04-25 17:58:25.591 | DEBUG    | huixiangdou.service.llm_server_hybrid:generate_response:522 - Q:有主题的疑问句，结果用 0~10 表示。直接提供得分不要解释。
判断标准：有主语谓语宾语并且是疑问句得 10 分；缺少主谓宾扣分；陈述句直接得 0 分；不是疑问句直接得 0 分。直接提供得分不要解释 A:根据给定的标准，"今天天气怎么样？" 是一个有主语、谓语和宾语，并且是疑问句的句子。因此，它的得分是 10 分。        remote local timecost 1.4384238719940186
04/25/2024 17:58:25 - [INFO] -aiohttp.access->>>    127.0.0.1 [25/Apr/2024:17:58:24 +0800] "POST /inference HTTP/1.1" 200 474 "-" "python-requests/2.31.0"
2024-04-25 17:58:25.756 | INFO     | huixiangdou.service.llm_server_hybrid:generate_response:519 - ('告诉我这句话的主题，直接说主题不要解释： "今天天气怎么样？"', '主题：天气。')
2024-04-25 17:58:25.756 | DEBUG    | huixiangdou.service.llm_server_hybrid:generate_response:522 - Q:告诉我这句话的主题，直接说主题不要解释： "今天天气怎么样？ A:主题：天气。        remote local timecost 0.161553144495605
04/25/2024 17:58:25 - [INFO] -aiohttp.access->>>    127.0.0.1 [25/Apr/2024:17:58:25 +0800] "POST /inference HTTP/1.1" 200 206 "-" "python-requests/2.31.0"
2024-04-25 17:58:25.772 | INFO     | __main__:lark_send_only:79 - ErrorCode.UNRELATED, 今天天气怎么样？, , ['HuixiangDou.pdf']
(InternLM2_Huixiangdou) root@intern-studio-50023492:~/huixiangdou#
```

Tutorial/opencompass/opencompass_tutorial.md at main · InternLM/Tutorial (github.com)

内容总结如下：

RAG技术概述：RAG（检索增强生成）技术通过检索与用户输入相关的信息片段，结合外部知识库生成更准确、更丰富的回答。这种技术可以解决大型语言模型（LLMs）在处理知识密集型任务时可能遇到的诸多挑战，例如生成幻觉、处理过时的信息以及缺乏透明和可追溯的推理过程。RAG技术通过使基础模型能够进行非参数知识更新，实现了对新领域知识的快速掌握，无需额外训练即可适应新的信息环境。

RAG的效果比对：通过具体的使用实例——茴香豆应用，RAG技术显示了其在未经增训的情况下通过外部知识增强对新信息的快速适应和回答质量的提高。茴香豆应用的问答效果对比表明，传统模型如InternLM2-Chat-7B在没有接入RAG技术时，很难处理未被训练到的新兴话题。

环境配置：详细描述了如何在Intern Studio服务器上部署“茴香豆”应用，从创建开发机、配置系统镜像，到选择合适的硬件资源。此外，还包括了如何在创建的开发机中设置和激活所需的虚拟环境，确保所有开发和运行操作都在适当的环境下进行。

下载及安装依赖：介绍了如何准备环境，包括从Intern Studio的共享文件中获取必需的模型文件以避免外部下载和登录问题，并详细列出了安装的Python库和依赖，这些都是运行“茴香豆”所必需的。

使用茴香豆搭建RAG助手：
-配置文件调整：讲解了如何通过修改config.ini文件，来指定模型路径，确保向量数据库和重排序模型正确加载。
-知识库创建：步骤包括从茴香豆语料库中提取特征，建立向量数据库，并区分接受和拒绝的问题，以优化检索过程，并通过精确匹配来提高回答的相关性和质量。

运行茴香豆知识助手：最后阶段包括设置和测试茴香豆应用，验证RAG技术的实际效果。通过预定义的问题集测试应用程序的响应，展示了基于知识增强的答案生成能力，从而证实了茴香豆技术助理在面对具体问题时的实用性和效率。