第三次课

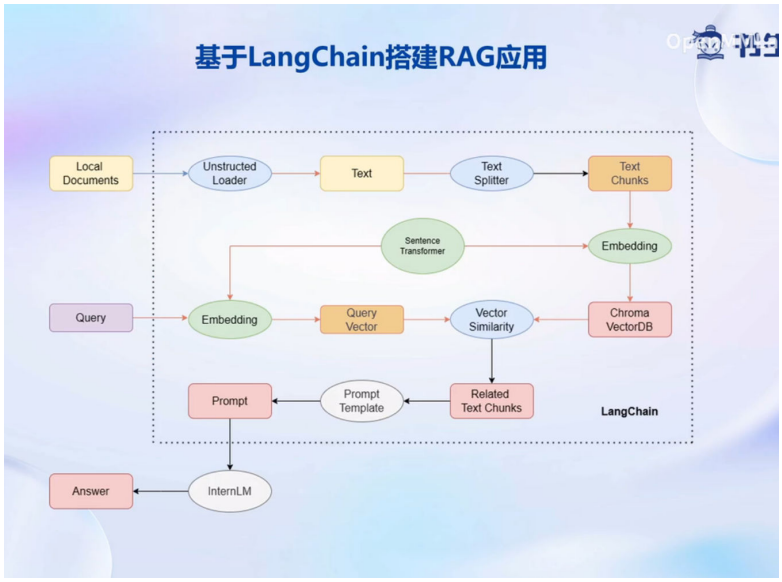
2024年1月18日 17:20

两种不同解决策略：

RAG:外挂知识库，把检索的文章和用户提示词一起交给大模型，占用上下文长度大

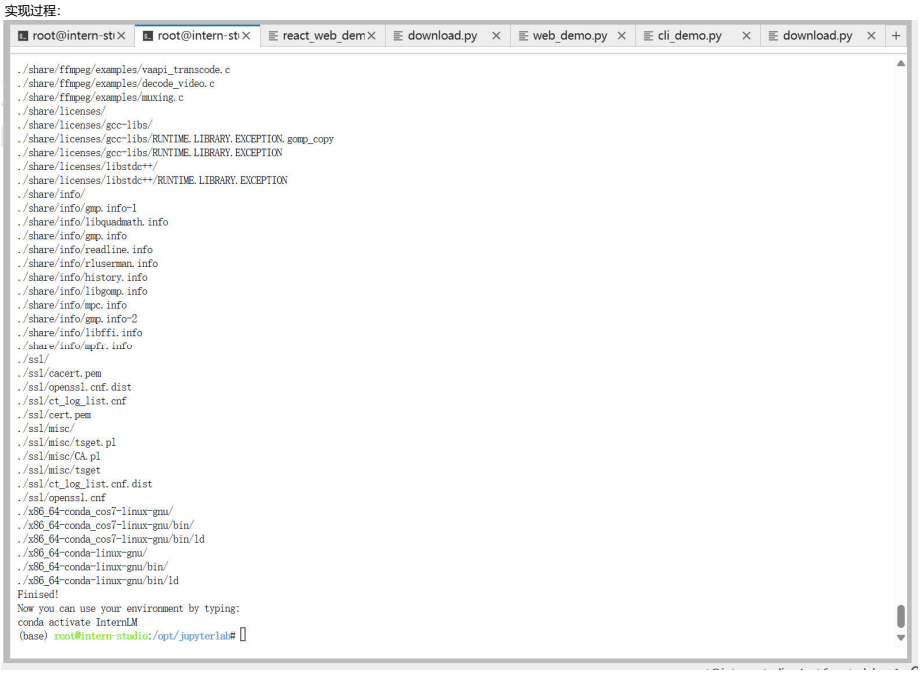
RAG过程：

基于LangChain:



Limitations:

实现过程：



配置环境，安装依赖：

```
root@intern-st× | root@intern-st× | react_web_dem× | download.py × | web_demo.py × | cli_demo.py × | download.py × | +

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
(InternLM) root@intern-studio:/opt/jupyterlab# pip install accelerate==0.24.1
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Collecting accelerate==0.24.1
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/13/9e/ee987874058f2d93006961f6ff49e0cb60ab9c26709ebe06bfa8707a4d8/accelerate-0.24.1-py3-none-any.whl (261 kB
)
Requirement already satisfied: numpy>=1.17 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from accelerate==0.24.1) (1.26.2)
Requirement already satisfied: packaging>=20.0 in /root/.local/lib/python3.10/site-packages (from accelerate==0.24.1) (23.2)
Requirement already satisfied: psutil in /root/.local/lib/python3.10/site-packages (from accelerate==0.24.1) (5.9.7)
Requirement already satisfied: pyyaml in /root/.local/lib/python3.10/site-packages (from accelerate==0.24.1) (6.0.1)
Requirement already satisfied: torch>=1.10.0 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from accelerate==0.24.1) (2.0.1)
Requirement already satisfied: huggingface-hub in /root/.local/lib/python3.10/site-packages (from accelerate==0.24.1) (0.20.2)
Requirement already satisfied: filelock in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from torch>=1.10.0->accelerate==0.24.1) (3.13.1)
Requirement already satisfied: typing-extensions in /root/.local/lib/python3.10/site-packages (from torch>=1.10.0->accelerate==0.24.1) (4.9.0)
Requirement already satisfied: sympy in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from torch>=1.10.0->accelerate==0.24.1) (1.11.1)
Requirement already satisfied: networkx in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from torch>=1.10.0->accelerate==0.24.1) (3.1)
Requirement already satisfied: jinja2 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from torch>=1.10.0->accelerate==0.24.1) (3.1.2)
Requirement already satisfied: fsspec>=2023.5.0 in /root/.local/lib/python3.10/site-packages (from huggingface-hub->accelerate==0.24.1) (2023.12.2)
Requirement already satisfied: requests in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from huggingface-hub->accelerate==0.24.1) (2.31.0)
Requirement already satisfied: tqdm>=4.42.1 in /root/.local/lib/python3.10/site-packages (from huggingface-hub->accelerate==0.24.1) (4.66.1)
Requirement already satisfied: MarkupSafe>=2.0 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from jinja2->torch>=1.10.0->accelerate==0.24.1) (2.1.1)
Requirement already satisfied: charset-normalizer<4,>=2 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from requests->huggingface-hub->accelerate==0.24.1)
 (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from requests->huggingface-hub->accelerate==0.24.1) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from requests->huggingface-hub->accelerate==0.24.1) (1.26
.18)
Requirement already satisfied: certifi>=2017.4.17 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from requests->huggingface-hub->accelerate==0.24.1) (2023
.11.17)
Requirement already satisfied: mpmath>=0.19 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from sympy->torch>=1.10.0->accelerate==0.24.1) (1.3.0)
Installing collected packages: accelerate
  Attempting uninstall: accelerate
    Found existing installation: accelerate 0.26.0
    Uninstalling accelerate-0.26.0:
      Successfully uninstalled accelerate-0.26.0
Successfully installed accelerate-0.24.1
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
(InternLM) root@intern-studio:/opt/jupyterlab#
(InternLM) root@intern-studio:/opt/jupyterlab#
(InternLM) root@intern-studio:/opt/jupyterlab#
(InternLM) root@intern-studio:/opt/jupyterlab#
```

安装命令行工具：

```
(InternLM) root@intern-studio:/opt/jupyterlab# pip install -U huggingface_hub
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Requirement already satisfied: huggingface_hub in /root/.local/lib/python3.10/site-packages (0.20.2)
Requirement already satisfied: filelock in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from huggingface_hub) (3.13.1)
Requirement already satisfied: fsspec>=2023.5.0 in /root/.local/lib/python3.10/site-packages (from huggingface_hub) (2023.12.2)
Requirement already satisfied: requests in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from huggingface_hub) (2.31.0)
Requirement already satisfied: tqdm>=4.42.1 in /root/.local/lib/python3.10/site-packages (from huggingface_hub) (4.66.1)
Requirement already satisfied: pyyaml>=5.1 in /root/.local/lib/python3.10/site-packages (from huggingface_hub) (6.0.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /root/.local/lib/python3.10/site-packages (from huggingface_hub) (4.9.0)
Requirement already satisfied: packaging>=20.9 in /root/.local/lib/python3.10/site-packages (from huggingface_hub) (23.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from requests->huggingface_hub) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from requests->huggingface_hub) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from requests->huggingface_hub) (1.26.18)
Requirement already satisfied: certifi>=2017.4.17 in /root/.conda/envs/InternLM/lib/python3.10/site-packages (from requests->huggingface_hub) (2023.11.17)
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
(InternLM) root@intern-studio:/opt/jupyterlab#
```

手动下载nltk资源：

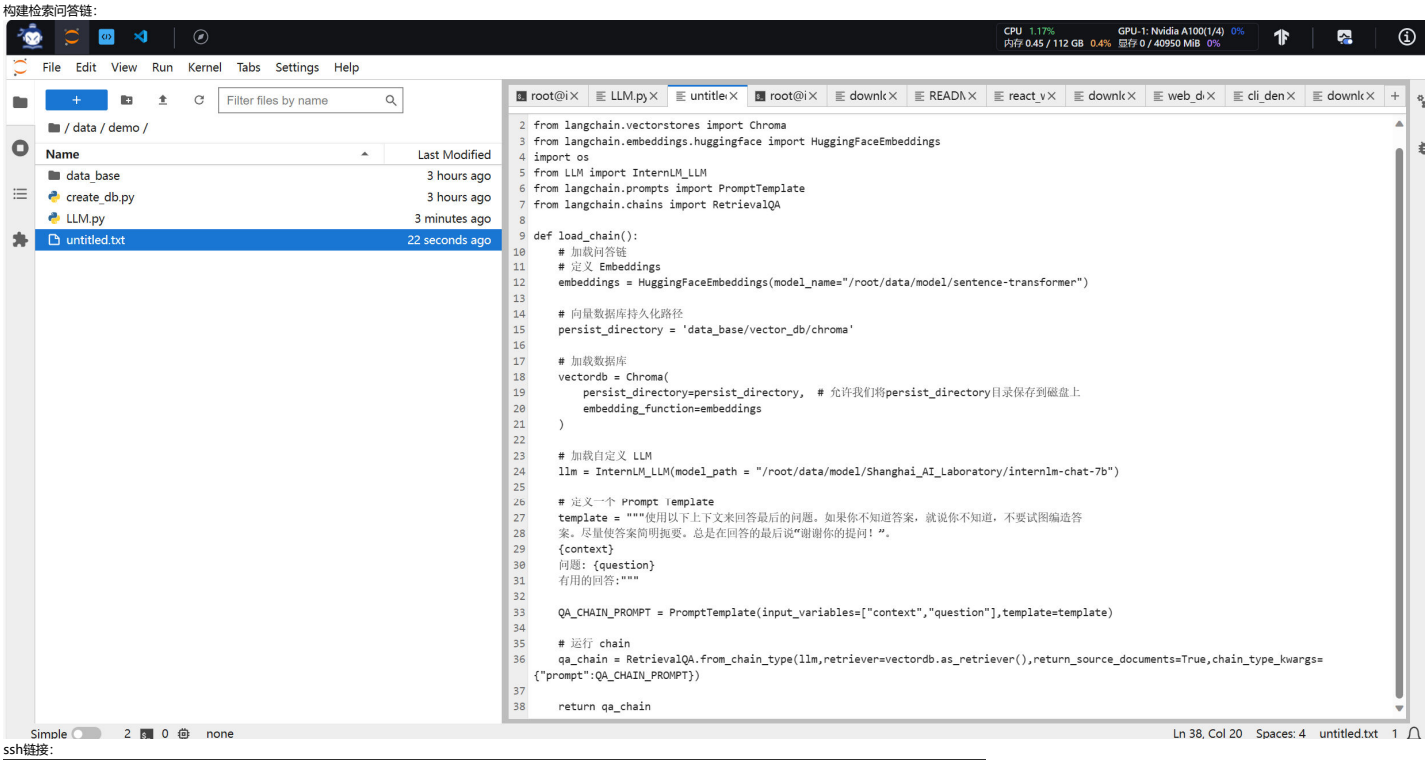
```
  inflating: punkt/malayalam.pickle
(InternLM) root@intern-studio:~/nltk_data/tokenizers# cd ../taggers
(InternLM) root@intern-studio:~/nltk_data/taggers# unzip averaged_perceptron_tagger.zip
Archive:  averaged_perceptron_tagger.zip
   creating: averaged_perceptron_tagger/
  inflating: averaged_perceptron_tagger/averaged_perceptron_tagger.pickle
(InternLM) root@intern-studio:~/nltk_data/taggers#
(InternLM) root@intern-studio:~/nltk_data/taggers#
(InternLM) root@intern-studio:~/nltk_data/taggers#
(InternLM) root@intern-studio:~/nltk_data/taggers# cd /root/data
(InternLM) root@intern-studio:~/data# python download_hf.py
Consider using `hf_transfer` for faster downloads. This solution comes with some limitations. See https://huggingface.co/docs/huggingface_hub/hf_transfer for more deta
ils.
Fetching 14 files: 100%|████████████████████████████████████████████████████████████| 14/14 [00:00<00:00, 14.11it/s]
/root/data/model/sentence-transformer
(InternLM) root@intern-studio:~/data#
(InternLM) root@intern-studio:~/data#
(InternLM) root@intern-studio:~/data#
(InternLM) root@intern-studio:~/data#
(InternLM) root@intern-studio:~/data#
(InternLM) root@intern-studio:~/data# cd /root
(InternLM) root@intern-studio:~#
```

下载本项目代码：

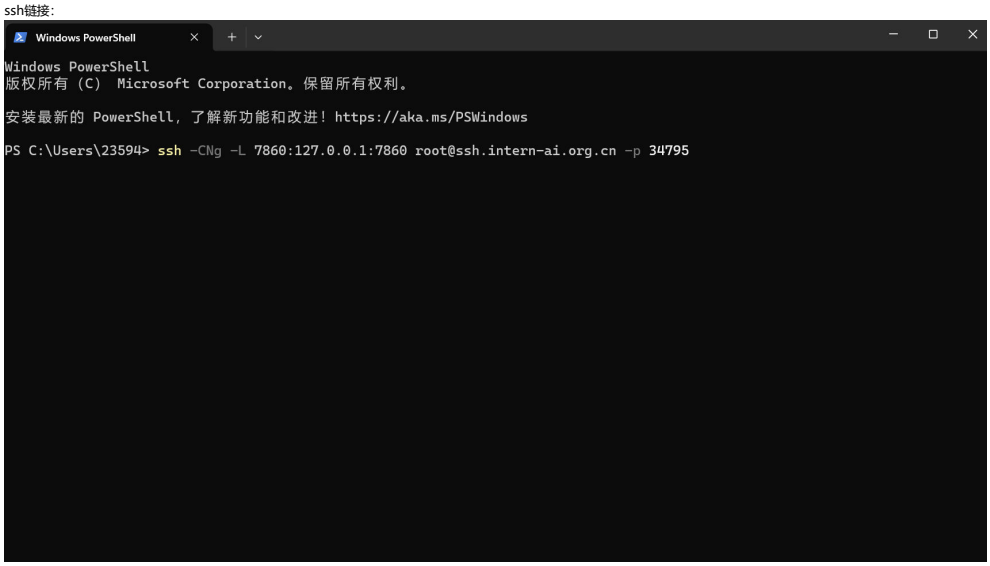
```
(base) root@intern-studio:/opt/jupyterlab# conda activate InternLM
(InternLM) root@intern-studio:/opt/jupyterlab# cd /root/data
(InternLM) root@intern-studio:~/data# git clone https://github.com/InternLM/tutorial
Cloning into 'tutorial'...
fatal: unable to access 'https://github.com/InternLM/tutorial/': Received HTTP code 503 from proxy after CONNECT
(InternLM) root@intern-studio:~/data# git clone https://kgithub.com/InternLM/tutorial
Cloning into 'tutorial'...
fatal: unable to access 'https://kgithub.com/InternLM/tutorial/': Received HTTP code 503 from proxy after CONNECT
(InternLM) root@intern-studio:~/data# git clone https:/https://mirror.ghproxy.com/github.com/InternLM/tutorial
Cloning into 'tutorial'...
fatal: protocol 'https:/https' is not supported
(InternLM) root@intern-studio:~/data# https://mirror.ghproxy.com/https://github.com/InternLM/tutorial
bash: https://mirror.ghproxy.com/https://github.com/InternLM/tutorial: No such file or directory
(InternLM) root@intern-studio:~/data# git clone https://mirror.ghproxy.com/https://github.com/InternLM/tutorial
Cloning into 'tutorial'...
remote: Enumerating objects: 548, done.
remote: Counting objects: 100% (317/317), done.
remote: Compressing objects: 100% (183/183), done.
remote: Total 548 (delta 233), reused 186 (delta 134), pack-reused 231
Receiving objects: 100% (548/548), 11.41 MiB | 4.36 MiB/s, done.
Resolving deltas: 100% (271/271), done.
(InternLM) root@intern-studio:~/data#
```

构建向量数据库：

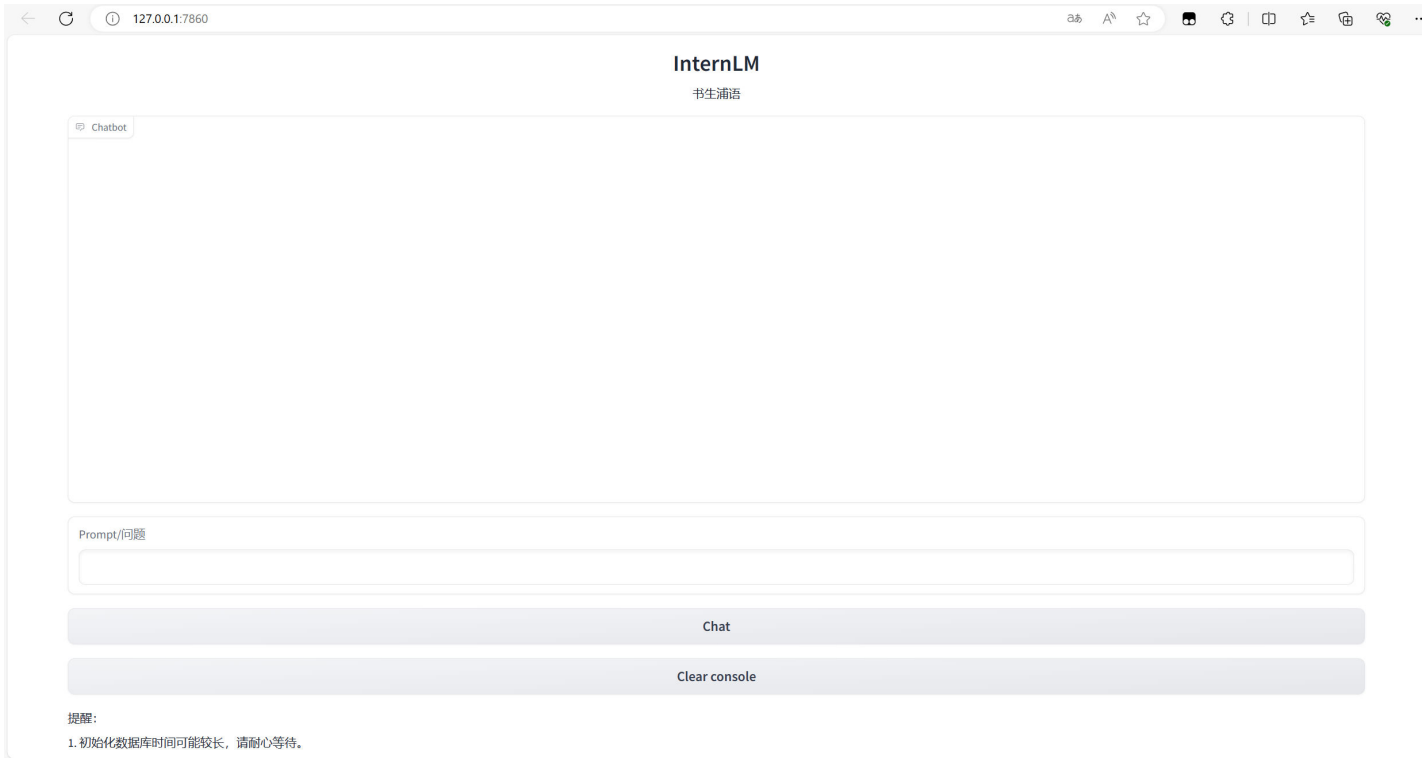
构建检索问答链：



ssh链接：



在本地电脑打开web demo:


127.0.0.1:7860
InternLM
书生浦语
Chatbot
Prompt/问题
Chat
Clear console
提醒:
1. 初始化数据库时间可能较长，请耐心等待。


InternLM
书生浦语
Chatbot
InternLM是什么
仩嘀认竺潪能exp1031實験屋上s海h人r工g智z能n实s验y室s原ori创gin
InternLM是什么
InternLM是一个开源的轻量级训练框架，旨在支持大模型训练，可在拥有数千个GPU的大型集群上进行预训练，并在单个GPU上进行微调，同时实现了卓越的性能优化。该框架在1024个GPU上训练时，可以实现近90%的加速效率。
Prompt/问题
Chat
Clear console
提醒:
1. 初始化数据库时间可能较长，请耐心等待。