CPU 47.53%
内存 1.81 / 72 GB 2.51%
GPU-1: 30% Nvidia A100 0%
显存 15950 / 24566 MiB 64.93%

```
  warnings.warn(
/root/.conda/envs/InternLM2_Huixiangdou/lib/python3.10/site-packages/langchain/embeddings/__init__.py:29: LangChainDeprecationWarning: Importing em
beddings from langchain is deprecated. Importing from langchain will no longer be supported as of langchain==0.2.0. Please import from langchain-co
mmunity instead:

`from langchain_community.embeddings import HuggingFaceEmbeddings`.

To install langchain-community run `pip install -U langchain-community`.
  warnings.warn(
2024-04-12 14:41:08.930 | INFO     | __main__:run:180 - waiting for server to be ready..
2024-04-12 14:41:11.933 | INFO     | __main__:run:180 - waiting for server to be ready..
2024-04-12 14:41:14.936 | INFO     | __main__:run:180 - waiting for server to be ready..
2024-04-12 14:41:17.939 | INFO     | __main__:run:180 - waiting for server to be ready..
04/12/2024 14:41:19 - [INFO] -accelerate.utils.modeling->>>    We will use 90% of the memory on device 0 for storing the model, and 10% for the buf
fer to avoid OOM. You can set `max_memory` in to a higher value to use more memory (at your own risk).
Loading checkpoint shards:   0%|                                                                                             | 0/8 [00:00<?, ?it/s]
2024-04-12 14:41:20.942 | INFO     | __main__:run:180 - waiting for server to be ready..
Loading checkpoint shards:  12%|███████████                                                                        | 1/8 [00:02<00:17,  2.48s/it]
2024-04-12 14:41:23.945 | INFO     | __main__:run:180 - waiting for server to be ready..
Loading checkpoint shards:  25%|█████████████████████                                                              | 2/8 [00:04<00:14,  2.47s/it]
2024-04-12 14:41:26.948 | INFO     | __main__:run:180 - waiting for server to be ready..
Loading checkpoint shards:  50%|██████████████████████████████████████████                                         | 4/8 [00:09<00:09,  2.47s/it]
2024-04-12 14:41:29.951 | INFO     | __main__:run:180 - waiting for server to be ready..
Loading checkpoint shards:  62%|█████████████████████████████████████████████████████                              | 5/8 [00:12<00:07,  2.48s/it]
2024-04-12 14:41:32.954 | INFO     | __main__:run:180 - waiting for server to be ready..
Loading checkpoint shards:  75%|████████████████████████████████████████████████████████████████                   | 6/8 [00:14<00:04,  2.48s/it]
2024-04-12 14:41:35.957 | INFO     | __main__:run:180 - waiting for server to be ready..
Loading checkpoint shards:  88%|██████████████████████████████████████████████████████████████████████████         | 7/8 [00:17<00:02,  2.50s/it]
2024-04-12 14:41:38.960 | INFO     | __main__:run:180 - waiting for server to be ready..
Loading checkpoint shards: 100%|█████████████████████████████████████████████████████████████████████████████████████| 8/8 [00:19<00:00,  2.45s/it]
======== Running on http://0.0.0.0:8888 ========
(Press CTRL+C to quit)
2024-04-12 14:41:41.963 | INFO     | __main__:run:187 - Hybrid LLM Server start.
2024-04-12 14:41:41.970 | INFO     | __main__:run:192 - Config loaded.
2024-04-12 14:41:41.971 | INFO     | huixiangdou.service.retriever:__init__:202 - loading test2vec and rerank models
04/12/2024 14:41:46 - [INFO] -sentence_transformers.SentenceTransformer->>>    Load pretrained SentenceTransformer: /root/models/bce-embedding-base
_v1
/root/.conda/envs/InternLM2_Huixiangdou/lib/python3.10/site-packages/torch/_utils.py:776: UserWarning: TypedStorage is deprecated. It will be remov
ed in the future and UntypedStorage will be the only storage class. This should only matter to you if you are using storages directly.  To access U
ntypedStorage directly, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
```

InternStudio
WebIDE
上海500-800万50-70平、70-
studio.intern-ai.org.cn/console/trinity/20240412-368bebf-40017566
CPU 21.94%
内存 5.47 / 72 GB 7.6%
GPU-1: 30% Nvidia A100 40%
显存 20568 / 24566 MiB 83.73%

```
构建自己的知识助手，而无需编写任何代码。

HuixiangDou 的更多信息可以在 [arxiv2401.08772](https://arxiv.org/abs/2401.08772) 论文中找到。，['README.md']
2024-04-12 14:47:01.516 | 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。')
2024-04-12 14:47:01.516 | 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 timecost 2.692380428314209
04/12/2024 14:47:01 - [INFO] -aiohttp.access->>>    127.0.0.1 [12/Apr/2024:14:46:58 +0800] "POST /inference HTTP/1.1" 200 681 "-" "python-requests/2.31.0"
2024-04-12 14:47:01.824 | 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。')
2024-04-12 14:47:01.824 | 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 0.30391120910644453
04/12/2024 14:47:01 - [INFO] -aiohttp.access->>>    127.0.0.1 [12/Apr/2024:14:47:01 +0800] "POST /inference HTTP/1.1" 200 242 "-" "python-requests/2.31.0"
2024-04-12 14:47:02.178 | INFO     | huixiangdou.service.retriever:query:158 - target README_zh.md file length 11523
2024-04-12 14:47:02.178 | DEBUG    | huixiangdou.service.retriever:query:185 - query:主题：茴香豆的微信部署。 top1 file:README_zh.md
2024-04-12 14:47:05.163 | 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&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>  \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llm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。')
2024-04-12 14:47:05.163 | 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llm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 local timecost 2.9819862842559814
04/12/2024 14:47:05 - [INFO] -aiohttp.access->>>    127.0.0.1 [12/Apr/2024:14:47:02 +0800] "POST /inference HTTP/1.1" 200 721 "-" "python-requests/2.31.0"
2024-04-12 14:47:05.167 | WARNING  | huixiangdou.service.llm_client:generate_response:95 - disable remote LLM while choose remote LLM, auto fixed
```

InternStudio
WebIDE
127.0.0.1
上海500-800万50-70平、70-
studio.intern-ai.org.cn/console/trinity/20240412-368bebf-40017566
14
CPU 26.74%
内存 6.43 / 72 GB 8.93%
GPU-1: 30% Nvidia A100 4%
显存 15310 / 24566 MiB 62.32%

```
/root/.conda/envs/InternLM2_Huixiangdou/lib/python3.10/site-packages/langchain/embeddings/__init__.py:29: LangChainDeprecationWarning: Importing em
beddings from langchain is deprecated. Importing from langchain will no longer be supported as of langchain==0.2.0. Please import from langchain-co
mmunity instead:

`from langchain_community.embeddings import HuggingFaceEmbeddings`.

To install langchain-community run `pip install -U langchain-community`.
  warnings.warn(
/root/.conda/envs/InternLM2_Huixiangdou/lib/python3.10/site-packages/langchain/embeddings/__init__.py:29: LangChainDeprecationWarning: Importing em
beddings from langchain is deprecated. Importing from langchain will no longer be supported as of langchain==0.2.0. Please import from langchain-co
mmunity instead:

`from langchain_community.embeddings import HuggingFaceEmbeddings`.

To install langchain-community run `pip install -U langchain-community`.
  warnings.warn(

2024-04-12 14:52:29.306 | INFO     | __main__:<module>:59 - waiting for server to be ready..
2024-04-12 14:52:32.309 | INFO     | __main__:<module>:59 - waiting for server to be ready..
2024-04-12 14:52:35.312 | INFO     | __main__:<module>:59 - waiting for server to be ready..
2024-04-12 14:52:38.315 | INFO     | __main__:<module>:59 - waiting for server to be ready..
2024-04-12 14:52:41.318 | INFO     | __main__:<module>:59 - waiting for server to be ready..
04/12/2024 14:52:42 - [INFO] -accelerate.utils.modeling->>>    We will use 90% of the memory on device 0 for storing the model, and 10% for the buf
fer to avoid OOM. You can set `max_memory` in to a higher value to use more memory (at your own risk).
Loading checkpoint shards:   0%|                                                                                             | 0/8 [00:00<?, ?it/s]
2024-04-12 14:52:44.321 | INFO     | __main__:<module>:59 - waiting for server to be ready..
Loading checkpoint shards:  12%|███████████                                                                      | 1/8 [00:02<00:19,  2.71s/it]
2024-04-12 14:52:47.324 | INFO     | __main__:<module>:59 - waiting for server to be ready..
Loading checkpoint shards:  25%|█████████████████████                                                            | 2/8 [00:05<00:15,  2.64s/it]
2024-04-12 14:52:50.327 | INFO     | __main__:<module>:59 - waiting for server to be ready..
Loading checkpoint shards:  50%|██████████████████████████████████████████                                       | 4/8 [00:10<00:10,  2.66s/it]
2024-04-12 14:52:53.330 | INFO     | __main__:<module>:59 - waiting for server to be ready..
Loading checkpoint shards:  62%|█████████████████████████████████████████████████████                            | 5/8 [00:13<00:07,  2.62s/it]
2024-04-12 14:52:56.333 | INFO     | __main__:<module>:59 - waiting for server to be ready..
Loading checkpoint shards:  75%|████████████████████████████████████████████████████████████████                 | 6/8 [00:15<00:05,  2.64s/it]
2024-04-12 14:52:59.336 | INFO     | __main__:<module>:59 - waiting for server to be ready..
Loading checkpoint shards:  88%|██████████████████████████████████████████████████████████████████████████       | 7/8 [00:18<00:02,  2.64s/it]
2024-04-12 14:53:02.339 | INFO     | __main__:<module>:59 - waiting for server to be ready..
Loading checkpoint shards: 100%|█████████████████████████████████████████████████████████████████████████████████| 8/8 [00:20<00:00,  2.60s/it]
======== Running on http://0.0.0.0:8888 ========
(Press CTRL+C to quit)
```

127.0.0.1:7860
Run

输入你的提问

huixiangdou擅长什么

生成结果

```
{
 "text": "HuixiangDou擅长处理群聊天场景，并能够回答用户的问题，而不会导致信息过载。它使用了一个两阶段的管道，首先进行拒绝，然后进行响应，以确保在群聊天中提供有价值的信息。HuixiangDou基于LLM（大型语言模型），并具有低成本的特点。如果您想了解更多信息，请参考[arxiv2401.08772](https://arxiv.org/abs/2401.08772)。",
 "code": 0,
 "references": [
  "huixiangdou-inside.md",
  "README.md"
 ]
}
```