

1.RAG 技术是通过检索和生成,并利用外部知识库来增强 LLMs 的性能

OpenMMLab Bilibili

## RAG 技术概述

**定义**

RAG ( Retrieval Augmented Generation ) 是一种结合了检索 ( Retrieval ) 和生成 ( Generation ) 的技术,旨在通过利用**外部知识库**来增强大型语言模型 ( LLMs ) 的性能。它通过检索与用户输入相关的信息片段,并结合这些信息来生成更准确、更丰富的回答。

**知识**

解决LLMs在处理**知识密集型任务**时可能遇到的挑战。提供更准确的回答、降低成本、实现外部记忆。

**VS**

- 生成幻觉 ( hallucination )
- 过时知识
- 缺乏透明和可追溯的推理过程

**应用**

问答系统    文本生成    信息检索    图片描述

2.RAG 的工作原理:

- ① 将文档或网页分割成各向量存储在数据库中
- ② 将用户问题也编码成向量,并在数据库找出与之最相关的文档块
- ③ 将文档块和问题作为提示一起输入到大语言模型中

3.向量数据库:

OpenMMLab Bilibili

## 向量数据库 ( Vector-DB )

**数据存储**

将文本及其他数据通过其他预训练的模型转换为固定长度的向量表示,这些向量能够捕捉文本的语义信息。

**相似性检索**

根据用户的查询向量,使用向量数据库快速找出最相关的向量的过程。通常通过计算余弦相似度或其他相似性度量来完成。检索结果根据相似度得分进行排序,最相关的文档将被用于后续的文本生成。

**向量表示的优化**

包括使用更高级的文本编码技术,如句子嵌入或段落嵌入,以及对数据库进行优化以支持大规模向量搜索。

Images from : <https://github.com/chenzoni12/AISystem/blob/main/06Foundation/05Dataset/04VectorDB.pdf>

dot product

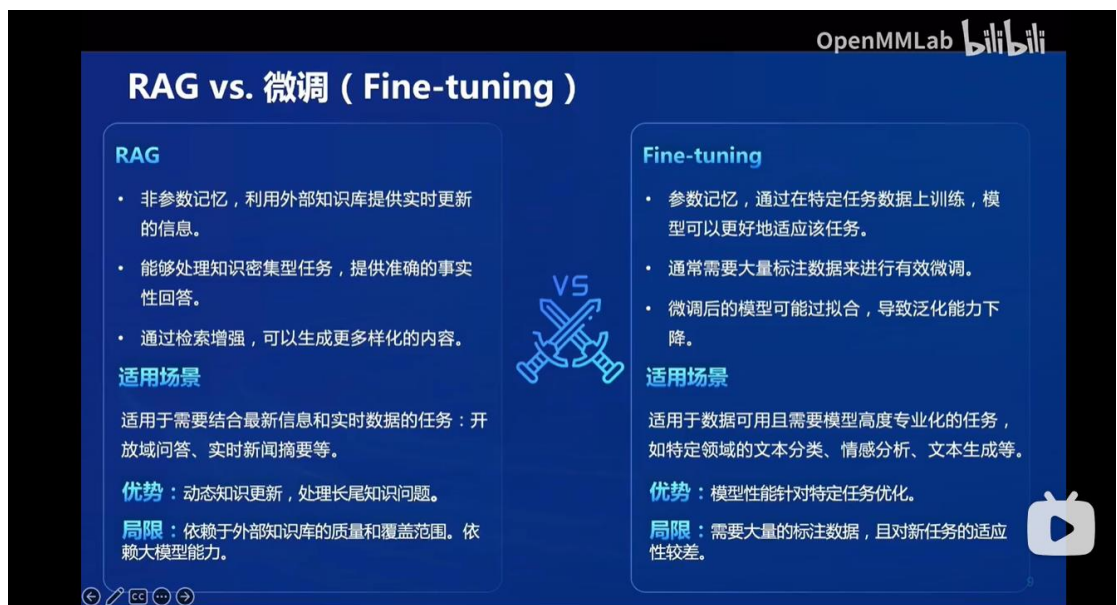
Cosine Distance

Query vector

4.RAG 常见优化方法



#### 4.RAG 与微调



#### 利用茴香豆搭载 RAG 智能助手

##### 1. 环境配置

###### 1.1

以在 Intern Studio 服务器上部署茴香豆为例。

首先，打开 In 进入开发机后，从官方环境复制运行 InternLM 的基础环境，命名为 InternLM2\_Huixiangdou，在命令行模式下运行：tern Studio 界面，点击 创建开发机 配置开发机系统。

## 1.2

复制茴香豆所需模型文件，为了减少下载和避免 HuggingFace 登录问题，所有作业和教程涉及的模型都已经存放在 Intern Studio 开发机共享文件中。本教程选用 InternLM2-Chat-7B 作为基础模型。

## 1.3 下载安装茴香豆

```
# 安装 python 依赖
# pip install -r requirements.txt
```

```
pip install protobuf==4.25.3 accelerate==0.28.0 aiohttp==3.9.3 auto-gptq==0.7.1
bcembedding==0.1.3 beautifulsoup4==4.8.2 einops==0.7.0 faiss-gpu==1.7.2
langchain==0.1.14 loguru==0.7.2 lxml_html_clean==0.1.0 openai==1.16.1
openpyxl==3.1.2 pandas==2.2.1 pydantic==2.6.4 pymupdf==1.24.1 python-
docx==1.1.0 pytoml==0.1.21 readability-lxml==0.8.1 redis==5.0.3 requests==2.31.0
scikit-learn==1.4.1.post1 sentence_transformers==2.2.2 textract==1.6.5 tiktoken==0.6.0
transformers==4.39.3 transformers_stream_generator==0.0.5 unstructured==0.11.2
```

## 因为 Intern Studio 不支持对系统文件的永久修改，在 Intern Studio 安装部署的同学不建议安装 Word 依赖，后续的操作和作业不会涉及 Word 解析。

## 想要自己尝试解析 Word 文件的同学，uncomment 掉下面这行，安装解析 .doc .docx 必需的依赖

```
# apt update && apt -y install python-dev python libxml2-dev libxslt1-dev antiword
unrtf poppler-utils pstotext tesseract-ocr flac ffmpeg lame libmad0 libsox-fmt-mp3 sox
libjpeg-dev swig libpulse-dev
```

## 2. 使用茴香豆搭载 RAG 助手

### 2.1

提取知识库特征，创建向量数据库。数据库向量化的过程应用到了 LangChain 的相关模块，默认嵌入和重排序模型调用的网易 BCE 双语模型，如果没有在 config.ini 文件中指定本地模型路径，茴香豆将自动从 HuggingFace 拉取默认模型。

除了语料知识的向量数据库，茴香豆建立接受和拒答两个向量数据库，用来在检索的过程中更加精确的判断提问的相关性，这两个数据库的来源分别是：

接受问题列表，希望茴香豆助手回答的示例问题

存储在 huixiangdou/resource/good\_questions.json 中

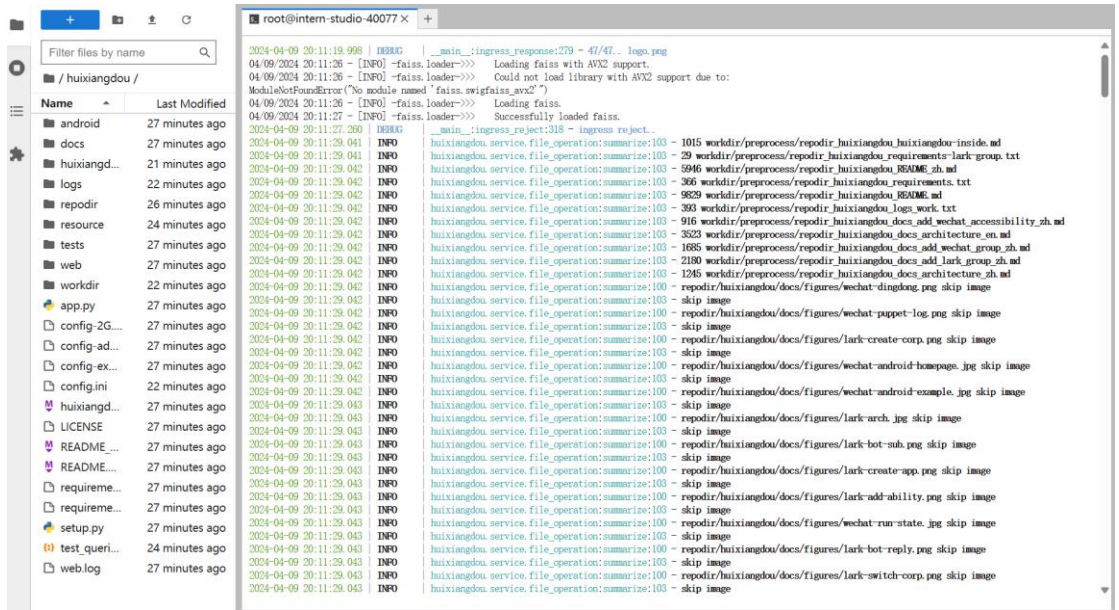
拒绝问题列表，希望茴香豆助手拒答的示例问题

存储在 huixiangdou/resource/bad\_questions.json 中

其中多为技术无关的主题或闲聊

如："nihui 是谁"，"具体在哪些位置进行修改？"，"你是谁？"，"1+1"

### 2.2. 创建知识库



## 2.3. 运行茴香豆知识助手

```
InternLM2_Huixiangdou) root@intern-studio-40077417:~/huixiangdou#
InternLM2_Huixiangdou) root@intern-studio-40077417:~/huixiangdou# # 填入问题
InternLM2_Huixiangdou) root@intern-studio-40077417:~/huixiangdou# sed -i '74s/.*/ queries = ["huixiangdou 是什么?", "茴香豆怎么部署到微信群", "今天天气怎么样?"]/'
~/huixiangdou/huixiangdou/main.py
InternLM2_Huixiangdou) root@intern-studio-40077417:~/huixiangdou#
InternLM2_Huixiangdou) root@intern-studio-40077417:~/huixiangdou# # 运行茴香豆
InternLM2_Huixiangdou) root@intern-studio-40077417:~/huixiangdou# cd /root/huixiangdou/
InternLM2_Huixiangdou) root@intern-studio-40077417:~/huixiangdou# python3 -m huixiangdou.main --standalone
root/.conda/envs/InternLM2_Huixiangdou/lib/python3.10/site-packages/langchain/embeddings/_init_.py:29: LangChainDeprecationWarning: Importing embeddings from langch
s deprecated. Importing from langchain will no longer be supported as of langchain==0.2.0. Please import from langchain-community instead:

from langchain_community.embeddings import HuggingFaceEmbeddings

> 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 embeddings from langch
s deprecated. Importing from langchain will no longer be supported as of langchain==0.2.0. Please import from langchain-community instead:

from langchain_community.embeddings import HuggingFaceEmbeddings

> install langchain-community run `pip install -U langchain-community`.
warnings.warn(
2024-04-09 20:12:56.327 | INFO | __main__:run:180 - waiting for server to be ready..
2024-04-09 20:12:59.330 | INFO | __main__:run:180 - waiting for server to be ready..
2024-04-09 20:13:02.355 | INFO | __main__:run:180 - waiting for server to be ready..
2024-04-09 20:13:05.358 | INFO | __main__:run:180 - waiting for server to be ready..
2024-04-09 20:13:08.361 | INFO | __main__:run:180 - waiting for server to be ready..
4/09/2024 20:13:11 - [INFO] -accelerate.utils.modeling>>> We will use 90% of the memory on device 0 for storing the model, and 10% for the buffer to avoid OOM. Yo
set 'max_memory' in to a higher value to use more memory (at your own risk).
```

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）的群聊聊天助手。它使用两阶段管道来处理群聊聊天场景，以回答用户问题，而不会导致信息过载。该工具的优点包括：\n\n1. 两阶段管道：HuixiangDou 使用一个两阶段管道来处理群聊聊天。首先，它使用一个预训练的模型来快速判断一个消息是否需要被处理。如果消息需要被处理，它将使用另一个模型来生成回答。这种设计可以避免信息过载，并提高处理效率。2. 低成本：HuixiangDou 只需要 1.5GB 的内存，并且不需要进行训练，因此它的成本相对较低。3. 完整套件：HuixiangDou 提供了 Web、Android 和管道源代码的完整套件，这些代码是工业级和商业上可行的。4. 运行场景：HuixiangDou 已经在多个场景中运行，包括 WeChat 和 Feishu 群组。5. 开源：HuixiangDou 的源代码是开源的，并且可以在 [OpenXLab] (https://openxlab.org.cn/apps/detail/tpoisonoo/huixiangdou-web) 上使用，用户可以轻松地构建自己的知识助手，而无需编写任何代码。'\n\nHuixiangDou 的更多信息可以在 [arxiv2401.08772] (https://arxiv.org/abs/2401.08772) 论文中找到。'\n\n2024-04-09 20:14:00.645 | ERROR | huixiangdou.service.llm\_server\_hybrid.generate\_response:599 - OSError: [Errno 11] Resource temporarily unavailable

## 结果显示：

问题：HuixiangDou 是什么？\n 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）的群聊聊天助手。它使用两阶段管道来处理群聊聊天场景，以回答用户问题，而不会导致信息过载。该工具的优点包括：

- 两阶段管道：HuixiangDou 使用一个两阶段管道来处理群聊聊天。首先，它使用一个预训练的模型来快速判断一个消息是否需要被处理。如果消息需要被处理，它将使用另一个模型来生成回答。这种设计可以避免信息过载，并提高处理效率。
- 低成本：HuixiangDou 只需要 1.5GB 的内存，并且不需要进行训练，因此它的成本相对较低。
- 完整套件：HuixiangDou 提供了 Web、Android 和管道源代码的完整套件，这些代码是工业级和商业上可行的。
- 运行场景：HuixiangDou 已经在多个场景中运行，包括 WeChat 和 Feishu 群组。
- 开源：HuixiangDou 的源代码是开源的，并且可以在 [OpenXLab] (https://openxlab.org.cn/apps/detail/tpoisonoo/huixiangdou-web) 上使用，用户可以轻松地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。 remote local timestamp 22.593166828155518

2024-04-09 20:14:31.414 | 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：有主语谓语句并且是疑问句得 10 分；缺少主语宾扣分；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 timestamp 2.7529218196868896



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 A: 要部署茴香豆到微信群, 请按照以下步骤操作:

1. **\*\*安装茴香豆\*\***:
  - 首先, 您需从 GitHub 上下载茴香豆的源代码。
  - 在您的计算机上安装 Python 3.8 或更高版本。
  - 使用 pip 安装茴香豆的依赖项: `pip install -r requirements.txt`。
2. **\*\*准备数据\*\***:
  - 您需要为茴香豆提供一些数据, 例如问题、答案和相关知识。
  - 将数据整理成 JSON 格式, 并保存为 `data.json` 文件。
3. **\*\*配置茴香豆\*\***:
  - 打开 `config.ini` 文件, 并根据您的需求配置参数。
  - 配置 `model` 参数, 以选择您要使用的模型。
  - 配置 `data` 参数, 以指定您要使用的数据文件。
  - 配置 `log` 参数, 以指定日志文件的路径。
4. **\*\*运行茴香豆\*\***:
  - 在终端中, 导航到茴香豆的根目录。
  - 运行 `python huixiangdou.py` 启动茴香豆。
  - 茴香豆将开始处理数据, 并准备回答用户的问题。
5. **\*\*集成到微信群\*\***:
  - 在您的微信中, 创建一个新的群聊。
  - 将茴香豆的 QQ 号添加到该群聊中。
  - 茴香豆将开始接收来自微信群的消息, 并尝试回答用户的问题。
6. **\*\*测试和优化\*\***:
  - 测试茴香豆的性能, 并根据需要进行调整。
  - 您可以通过添加更多数据、调整参数或使用更高级的模型来提高茴香豆的性能。

请注意, 以上步骤仅提供了一个基本的部署流程, 您可能需要根据您的具体需求进行一些调整。同时, 为了确保茴香豆的正常运行, 请确保您的计算机具有足够的计算资源, 并且您的数据集是干净、有组织且格式正确的。  
remote local timecost 45.48393630981445  
04/09/2024 20:15:21 - [INFO] -aiohttp.access>>> 127.0.0.1 [09/Apr/2024:20:14:35 +0800] "POST /inference HTTP/1.1" 200 3315 "-" "python-requests/2.31.0"  
2024-04-09 20:15:21.252 | INFO | \_\_main\_\_:lark\_send\_only:79 - ErrorCode.SUCCESS, 茴香豆怎么部署到微信群, 要部署茴香豆到微信群, 请按照以下步骤操作:

在茴香豆 Web 版中创建自己领域的知识问答助手

首先建立自己的知识库,在自己的知识库中上传与自己想创建内容相关的文档,该文档中需要包含你所想提问的具体知识,同时可以设计自己的拒答主题的正反例,而该模型会根据你设置的主题在相应的对话框中选择不回答相关问题.

小时候, 我想当开源人, 朋友给我鼓励和我最爱!  
小星星 🌟 🍌

不再显示 前往鼓励



computer science

.....

取消

前往

作  
相



请简单介绍一下书生浦语大模型



书生浦语大模型是一款通用人工智能的重要途径，具有超长上下文能力和综合性能全面提升的特点。它能够实现结构化创作和可靠的数据分析，并且强化了内生计算能力和代码解释器。模型在多个能力维度上都取得了不错的评测结果，包括语言知识、推理、数学、代码等方面。书生浦语大模型2.0提供不同尺寸和类型的模型，支持多语言和多模态任务，并且有完整的开源体系，包含从数据、预训练、微调、部署、评测、应用等环节。此外，还有Open Compass 2.0思南大模型评测体系、英特尔开源模型推理和部署工具、智能体框架和多媒体多模态智能体工具箱等配套工具和资源。

参考文档:

书生浦语大模型.docx  
茴香豆.docx



今天中午吃日料吗



参考文档:

茴香豆.docx

[Empty]: Topics unrelated to the knowledge base. Updating good\_questions and bad\_questions can improve accuracy.



我可以利用书生浦语大模型做些什么



根据提供的参考材料，您可以利用书生浦语大模型进行多种任务，包括但不限于以下方面：

1. **语言处理**：书生浦语大模型能够进行语言知识、推理、数学、代码等方面的处理，这意味着您可以利用它来进行文本翻译、摘要生成、对话系统构建、情感分析、问题回答等基于语言的任务。
2. **数据分析**：该模型可以用于可靠的数据分析，您可以利用它来处理大量文本数据，提取有用信息，进行数据挖掘和预测分析。
3. **结构化创作**：书生浦语大模型具备结构化创作的能力，您可以利用它生成新闻文章、科技报告、市场分析等专业文档。
4. **代码生成与解释**：模型强化了内生计算能力和代码解释器，您可以使用它来生成或解释代码，特别是在需要快速开发原型或自动化代码生成的场景中。
5. **多语言和多模态任务**：书生浦语大模型2.0支持多种语言 and 多种模态任务，因此您可以在跨语言和多模态（如文本、图像、音频等）环境中使用该模型。

支持输入文字、emoji 和粘贴图片



我可以利用茴香豆做什么



在参考材料中并没有提及“茴香豆”这一概念，因此无法提供直接的答案。但是，茴香豆是一种常见的食材，您可以使用它来制作各种美食，如茴香豆炒肉、茴香豆炖排骨等。如果您能提供更多的背景信息或者具体的问题，我将能够为您提供更准确的答案。

参考文档:

茴香豆.docx  
书生浦语大模型.docx



茴香豆是什么



参考文档:

茴香豆.docx

[Empty]: Topics unrelated to the knowledge base. Updating good\_questions and bad\_questions can improve