11.20.2023

建立向量数据库存千万片文献，检索，生成 tf-idf算法 分层：先用abstract建一个表，再用doi找全文 技术点：1 embedding search 2 目前是openai的embedding 找跨语言的embedding 3 faiss延迟 找一个低延迟向量数据库 4 生成大语言模型 openai llama 百川 5 langchain llama-index weblangchain

两个基于baichuan 13B的领域大模型：Disc-FinLLM Disc-MedLLM

11.22.2023

Embedding 范围讨论

Title(精确) abstract(模糊，存embedding) date authors doi

embedding模型调研和选择 bge m3e gte jina

底层语料 输入长度限制

测试：1000-2000 abstract embedding 测试搜索结果如何

搜索内容以及搜索方式

某个疾病相关的文献

具体文献的内容(作者、单位、结果)

预计文献数量：3000万

预计token数量：80亿

预计embedding维度数：1024

OceanGpt Knowledge-QA-LLM

11.24.2023

CellGPT搜索枢纽

融合海量生物医学公开文献+生物医学专业数据库和海量生物信息学代码库+打造新一代智能搜索引擎

RAG：知识搜索+大语言生成

指令微调：低成本获得领域大模型

扩散模型：蛋白质生成 问题速度慢、领域泛化性、精度低

蛋白质结构数据：alphafold预测

数据挖掘 大模型 社区枢纽 蛋白质生成 可编程设计

What would you like ChatGPT to know about you to provide better responses?

Given an abstract of a pubmed article, generate 10 questions and corresponding questions that 10 professional researchers will ask. Do not ask simple conceptual questions. The questions should be longer than 20 words, and the answers should be in json format, such as {"question": "", "answer": ""}, {"question":"", "answer":""}Give the answer directly, don't say anything else, answer in the language of the abstract, not in Chinese

How would you like ChatGPT to respond?

list of json with abstract and question answer pair

{

"title": "",

"authors":"",

"journal":"",

"abstract": "",

qa\_pairs: {

{

"question":{}

"answer":{}

},

{

"question":{}

"answer":{}

},

....

}

}

2023.12.1

访问数据库的api

生成数据库

2023.12.5

本地模型+本地embedding

Llama2 13G+RAG +langchain/llama index

魔塔社区modelscope [Llama-2-13b-ms · 模型库 (modelscope.cn)](https://modelscope.cn/models/modelscope/Llama-2-13b-ms/summary)

Local llama2+vectorstoreindex

本地大模型：RAG using local models

提问路由：提问路由的目标是找到一个路由函数，对于提问集中的每个，该函数能够确定最适合处理该提问的模型 。如果不同的 LLM 在不同领域和任务中具有专业知识，路由函数将能够预测提问属于 LLM 专业知识的概率。与现有的 RMR 方法相比，提问路由函数缓解了计算效率问题。路由函数通过分析每个提问的属性，尝试为其分配最适合的特定领域或任务的模型，从而在整个集成过程中提高效率。

Documentation QA using LlamaIndex and Milvus

生成问答对 instruct tuning

2023.12.7

每日论文：爬虫、文档处理、向量化->向量数据库、端口

10万综述：10万关键词、每个关键词匹配100篇文章、生成综述

RAG问答对话交互：意图识别、Prompt加工、论文代码数据库问答

生成引用：paper-qa

GeneGPT Gorilla LLM citationqueryengine

2023.12.8

ProteinMPNN

2023.12.12

结构化数据库 clinical traial gov 非结构化数据库 pubmed

综述agent：

Title(clinical traial给出)->Search term search term->doi(从pubmed检索) doi->full text 构建训练集abstract对应到reference finetune 生成目录

2023.12.15

Finetune ner llama