

基于本地 RAG 的井控知识问答系统的实现

刘月月

川庆钻探工程有限公司长庆钻井总公司

DOI:10.12238/pe.v2i5.9882

[摘要] 构建智能高效的井控专业知识问答系统可有效帮助企业进行员工培训,并进行实时井控作业在线指导。基于大语言模型(large language model,LLM)的知识问答系统相较传统方法具有更高的灵活性和丰富性,检索增强生成(Retrieval-Augmented Generation,RAG)技术是一种结合外部知识库与LLM构建知识问答系统的有效方法。本研究应用了开源RAG框架Dify与大模型部署平台Ollama搭建了本地化的井控知识问答系统,实现了井控知识的精准高效问答并可保护企业的隐私数据,具有较强的应用价值。

[关键词] 人工智能RAG井控知识问答系统

中图分类号: TP18 文献标识码: A

Implementation of the Well Control Knowledge Q&A System Based on Localized RAG Framework

Yueyue Liu

Changqing Drilling Company, CNPC Chuanqing Drilling Engineering Co., Ltd

[Abstract] Building an intelligent and efficient well control professional knowledge Q&A system can effectively assist enterprises in employee training and provide real-time online guidance for well control operations. The knowledge question answering system based on the Large Language Model (LLM) has higher flexibility and richness compared to traditional methods. Retrieval Augmented Generation (RAG) technology is an effective method for constructing knowledge question answering systems by combining external knowledge bases with LLM. This study applied the open-source RAG framework Dify and the large model deployment platform Olama to build a localized well control knowledge Q&A system, achieving accurate and efficient Q&A of well control knowledge and protecting the privacy data of enterprises, which has strong application value.

[Key words] Artificial Intelligence RAG Well Control Knowledge Q&A System

引言

井控是石油天然气勘探开发中重要的安全工作,与石油勘探生产的各个环节密切相关。人为因素在井控事故成因中占很大比例,因此强化井控安全教育培训有着重要意义^[1]。构建智能高效的井控专业知识问答系统,既能用于员工的教育培训,改变原有相对枯燥的学习方式,又能直接指导井控相关的生产活动。但传统的知识库问答系统仅能完成简单问题的问答,无法有效解决较复杂的问题,随着深度学习技术的兴起,近年来提出了一系列的较为有效的方法,如模板匹配、语义解析和embedding等,这些方法虽都取得了一些效果,但仍然存在准确率不高、通用性不强等问题^[2]。

大语言模型出现后,基于LLM的知识问答方法开始出现,其特点是显著提升了问答系统响应的灵活性与丰富性。但LLM在回答专业问题时会出现“幻觉”问题,即生成虚假、错误或陈旧的信息^[3]。微调LLM是一种解决方法,但微调LLM既需要大量标注好

的专用数据集和强大的GPU算力,还需要具备一定的模型微调经验,使得这种方法的应用受到一定限制。

RAG技术最先由Patrick Lewis等^[4]提出,其基本思想是将检索到的外部知识送入LLM生成反馈,这样可以保证答案的准确性和专业性。该方法无需再次训练模型,只要更新知识库即可应用新知识生成响应。因此,该方法是一种高效的知识问答方案。

1 基于RAG技术的知识问答系统

RAG技术是利用外部知识源和大语言模型共同生成合理答案的技术。外部数据源如文本知识、数据资料等在转化为文本格式后又切成小的文本块,再通过文本嵌入模型转化为向量并存储于向量数据库中。用户提出的问题同样通过嵌入模型编码为向量,利用向量相似性检索技术在数据库中查找与问题最为相近的知识向量。之后将检索结果送入大语言模型,由大语言模型总结并生成出答案,从而实现对专业知识领域的问答。具体流程如图1所示。

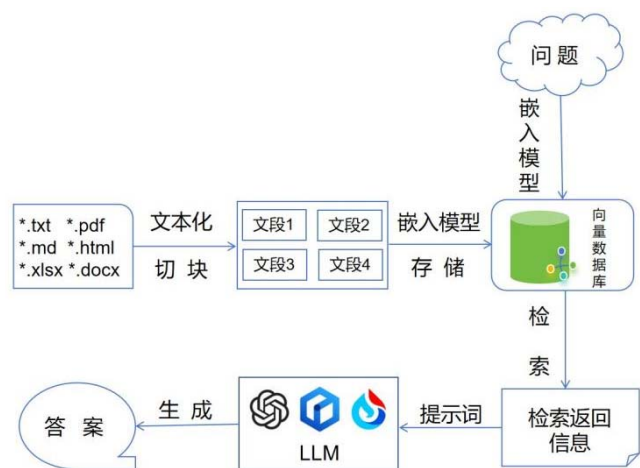


图1 RAG知识问答系统结构图

外部数据源可为多种格式的文件,如pdf,txt,word,excel以及html等,采用版面解析技术直接提取文本内容,对于pdf文件中不可直接提取的文本,则需通过光学字符识别技术(Optical Character Recognition, OCR)转换。之后将提取后的文本内容进行切块,切分方法既可根据自然段落,也可根据字符数量,但切分时不宜过大或者过小。大语言模型的输入字符数限制决定了文本块的上限,同时过大的文本块也会引入语义噪音;过小的文本块则有可能无法包含完整的知识信息。选择合适的切分方法有助于匹配关联度最高的文本内容,从而生成更优的结果。嵌入前,还需对文本块进行简单的清洗,比如去掉不需要的字符和空行等。文本块送入嵌入模型后会转化为一个高维向量并存储在向量数据库中。

当用户输入一个问题后,该问题也会通过嵌入模型转化为向量,在向量数据库中利用先进的检索算法,检索出与该问题相似度最高的几个的知识向量,这种相似性通常用欧式距离和余弦相似度等度量。检索得到的文本块将与提示词一起送入大语言模型,大语言模型分析其中的信息,利用其强大的语言处理能力和逻辑推理能力,产生语法通顺、内容准确且有意义的回答。通过这种方式,系统根据外部知识库生成了专业领域的回答,更贴合用户的需求。

2 RAG框架Dify与大模型管理软件Ollama

Dify是一个开源的大语言模型应用开发平台,开发者可以利用这个平台快速搭建RAG应用。其主要优势在于:(1)零代码开发和流程的可视化。开发过程中无需编写任何代码即可搭建RAG应用,提示词编排、大模型参数设置和知识库创建均以可视化界面操作。(2)该平台既支持以API的形式远程调用LLM,也支持接入多种本地模型部署框架(如Xinference、OpenLLM和Ollama等),从而可以实现模型的本地化部署。(3)只需将应用部署在服务器上,就可便捷的将开发好的RAG应用以网页的形式分享给用户或是直接嵌入到业务网站中。

Ollama是一个本地化部署和运行大模型的平台,能以极简的指令帮助用户在本地运行大模型。其使用的模型既可以是多

种官方的LLM,如Llama 2、Qwen 2和Gemma等,也可以使用自己训练或微调过的模型。Ollama因其易用性和良好的生态支持等优势,一经推出就广受好评。以Dify与Ollama相结合的方式,可以快捷的实现本地化搭建RAG应用。

3 构建本地化井控知识问答系统

本研究应用Dify与Ollama相结合的方式构建本地化井控知识问答系统。其步骤主要包含了井控知识库的构建和RAG系统的本地化搭建两部分内容。

3.1 井控知识库的构建

本文采用的知识库中包含《陆上石油天然气开采安全规程》(GB42294-2022)、《钻井井控技术规范》(Q/SY 02552-2022)和《放喷管线现场布局与安装规范》(Q/SYCQZ 1298-2021)等10余份与井控安全制度和作业规范相关的文档。这些文档包含word和pdf两种格式,并且除了文字外,还有图表形式的内容。虽然dify可直接读入pdf文件,但是文字转换效果不佳,因此本研究使用百度智能云千帆大模型平台的文档识别模型实现pdf到word格式的转换,之后将word格式的文件导入到井控知识库中,用Dify提供的自动分段与清洗方法实现对文本的切分和清洗,该方法会根据自然段落和字符数将文件自动截断为文本块。

上述文本块将进一步通过文本嵌入模型编码为向量,由于文本嵌入模型决定了向量检索的最终质量,其选择十分重要。当前很多嵌入模型主要基于英文数据集训练,所以中文支持能力较弱。本研究在测试了多种模型后,选择了北京数元灵科技有限公司开源的嵌入模型Dmeta-Embedding,并通过Ollama部署。然后这些向量被存储在Dify自带的开源向量数据库Weaviate中,该数据库既实现了向量数据的存储,同时也是一个向量检索引擎,可以高效实现相似向量的检索。

3.2 RAG系统的本地化搭建

设计RAG系统最核心的就是要选择一个合适的LLM,其性能很大程度上决定了最终输出的效果。考虑到本地的GPU为NVIDIA RTX3060 Laptop (6G显存),本文采用了参数量相对较少的阿里通义千问模型Qwen2-7B,与相近参数量的其他模型比较,该模型性能先进且对中文支持较好。同样,我们采用Ollama在本地加载运行此模型。同时我们还将模型的三个主要参数(温度,Top P和Top K)设置为较小的值,分别为0.1,0.2和1,其余参数取默认值。当这三个参数在取值较大时,意味着模型将产生更具创造性或多样性的答复;取值较小时,生成的结果更保守、更专注于原文。对于井控知识问答,我们希望获得的结果更忠实于原文件中的描述,因此选择较小的参数值。

在Dify中预先设置好提示词模板,从向量数据库检索到的文本块先被填充到提示词模板中,在人工加入的部分中,我们让模型称呼自己为“井控小助手”,并且用中文回答只与上下文知识相关的内容。随后上述内容被提交给LLM并生成答复。

为了提升用户的体验,我们应用了Dify提供的对话开场白功能,让模型主动同用户交流。此外,我们还开启了引用和归属功能,可以从回复中看到引用的具体文档及段落,让用户可以清

晰看到所给出答复的真实依据。Dify框架可将开发的井控问答系统一键发布,既支持直接以站点的形式显示,也支持将此应用嵌入到业务网站中以窗口的形式出现,这使得我们的系统能便捷、灵活的分享给用户。

4 结语

本研究成功应用Dify框架与Ollama平台开发了一种本地化部署的RAG井控知识问答系统。本系统利用了Ollama平台的模型管理功能,在本地部署了语言大模型和文本嵌入模型,再通过Dify框架提供的RAG工具,快速实现了井控知识问答RAG系统的搭建与发布。本系统首先可用于员工的教育培训,可改变以往刻板的学习方式,方便员工主动探索求知,提升学习效率。其次,对于井控相关的生产活动也有重要意义,方便工作人员随时随地以灵活的方式的查找资料。由于Dify是开源且可商用的,将此应用部署到企业服务器上是非常方便的,并且消费级的显卡即可完成一般的大模型推理任务,更重要的是,我们实现的本地化部署方案可以有效的保护企业的隐私数据。

然而,当前的RAG系统仍存在一些问题尚待突破,比如文本切分时的最理想的情况应为相对独立的知识点被分割成文本块,当前的分段方法还无法有效实现。又如对于整个文档提纲挈领的理解无法完成,导致各文本块内在的联系容易被忽略。目前,一些研究尝试在RAG中通过引入知识图谱来实现对全文知识结构的梳理,从而提升生成能力^[5]。随着技术的发展,基于专用知

识库的问答系统将越来越完善。

[参考文献]

[1]喻海霞.井控培训在井控风险控制中的重要性[J].西部探矿工程,2014,26(05):19-22.

[2]张金营,王天堃,么长英,等.基于大语言模型的电力知识库智能问答系统构建与评价[J/OL].计算机科学,1-10[2024-07-14].<https://libresource.xiyi.edu.cn:443/http/80/net/cnki/kns/yitlink/kcms/detail/50.1075.TP.20240528.0931.002.html>.

[3]齐思洋,胡慧云,李洪冰,等.融合大语言模型的领域问答系统构建方法[J/OL].北京邮电大学学报,1-7[2024-07-14].<https://libresource.xiyi.edu.cn:443/https/443/org/doi/yitlink/10.13190/j.jbupt.2023-279>.

[4]Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J].Advances in Neural Information Processing Systems,2020,33:9459-9474.

[5]Edge D,Trinh H,Cheng N,etal.From local to global:A graph rag approach to query-focused summarization[J].arxiv preprint arxiv:2404.16130,2024.

作者简介:

刘月月(1981--),女,汉族,陕西省西安市人,大学本科,高级工程师,研究方向:信息工程。