Operations Research and Fuzziology 运筹与模糊学, 2024, 14(6), 637-644
Published Online December 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.146564

基于RAG的供应链智能问答模型

王博楷[1]，牛　本[2]，刘玲丽[1*]

[1]武汉科技大学汽车与交通工程学院，湖北　武汉
[2]中国人民解放军78006部队，四川　成都

收稿日期：2024年11月7日；录用日期：2024年12月10日；发布日期：2024年12月16日

摘　要

随着人工智能技术的快速发展，问答模型已成为信息检索和知识获取的重要工具。在供应链管理相关领域，由于市场环境瞬息万变、专业术语复杂且相关流程繁多，传统的查询方式难以满足管理者和研究人员高效、准确获取信息的需求。相较于传统搜索引擎和原生开源LLMs，基于RAG的智能问答模型能够提供更高质量的答案，大幅提升知识检索的效率。因此本研究提出一种基于大模型、RAG技术的供应链智能问答模型，应用Embedding等方法，构建供应链领域相关知识文本块的向量数据库以及检索和生成模块，设计和优化Prompt提示词，提升大语言模型生成更精准和高质量的回答。结果表明，Ragas评测效果指标较好，忠实度分数，答案相关性分数，上下文相关性分数分别为0.73、0.8、0.83。模型应用测试回答准确度，专业度，可迁移性较原生LLMs具有显著优势，模型与数据集可以完全本地化，数据安全程度较高。基于RAG的供应链智能问答模型，验证了其能够充分利用供应链管理领域的大规模知识库，结合先进的自然语言处理技术和强化学习算法，实现对复杂供应链问题的深度理解和精准回答。

关键词

供应链，大语言模型，RAG检索增强生成

Supply Chain Intelligent Question and Answer Model Based on RAG

Bokai Wang[1], Ben Niu[2], Lingli Liu[1*]

[1]School of Automotive and Traffic Engineering, Wuhan University of Science and Technology, Wuhan Hubei
[2]Unit 78006 of PLA, Chengdu Sichuan

Received: Nov. 7th, 2024; accepted: Dec. 10th, 2024; published: Dec. 16th, 2024

Abstract

With the rapid development of artificial intelligence technology, question and answer model has

*通讯作者。

文章引用: 王博楷, 牛本, 刘玲丽. 基于 RAG 的供应链智能问答模型[J]. 运筹与模糊学, 2024, 14(6): 637-644.
DOI: 10.12677/orf.2024.146564

become an important tool for information retrieval and knowledge acquisition. In the field of supply chain management, due to the rapidly changing market environment, complex terminology and numerous related processes, traditional query methods are difficult to meet the needs of managers and researchers to obtain information efficiently and accurately. Compared with traditional search engines and native open source LLMs, RAG-based intelligent question and answer model can provide higher-quality answers and greatly improve the efficiency of knowledge retrieval. Therefore, this study proposes an intelligent question and answer model of supply chain based on large model and RAG technology. It applies the Embedding method to build a vector database of text blocks of relevant knowledge in the field of supply chain and search and generate modules, design and optimize Prompt words, and improve the large language model to generate more accurate and high-quality answers. The results show that the Ragas evaluation effect index is better, the scores of fidelity, answer relevance and context relevance are 0.73, 0.8 and 0.83, respectively. Compared with native LLMs, the model application test answer accuracy, professionalism, and portability have significant advantages. The model and data set can be fully localized, and the data security is higher. The supply chain intelligent question and answer model based on RAG proves that it can make full use of the large-scale knowledge base in the field of supply chain management, combine advanced natural language processing technology and reinforcement learning algorithm, and realize the deep understanding and accurate answer of complex supply chain questions.

Keywords

Supply Chain, Large Language Model, RAG Search Enhancement Generation

1. 引言

在生成式人工智能(AIGC)时代，ChatGPT (智能问答系统)为代表的大语言模型(LLMs)飞速发展，由于其卓越的文本理解与生成能力，这一技术极大地促进了人工智能与科学研究的深度融合，同时提升了各类企事业单位员工的工作效率，推动了知识管理与共享的进一步发展。尽管 LLM 大语言模型在生成内容方面表现出色，但它仍然无法处理超出其训练数据范围的任务。此外，仅依赖 LLM 大语言模型进行人机交互智能问答时，仍然存在大量的“幻觉”问题，即生成的答案可能不准确或与现实不符，因此检索增强式(Retrieval-Augmented Generation, RAG)开源 LLMs 在各行业各系统中的探索使用日益增多。如洪亮[1]等针对煤矿安全智能辅助预警决策的需求，基于煤矿安全隐患知识数据源，构建了一种基于 RAG 的煤矿安全智能问答模型。通过验证，该模型在瓦斯超限等煤矿安全隐患的智能辅助决策中，展示了其高效性、可靠性及良好的迁移能力。关殿玺[2]等，开发了一种基于 RAG 的岩土工程问答机器人，能够充分利用岩土工程领域的大规模知识库，实现对复杂岩土问题的深度理解和精准回答。Jaeyeon Byun [3]等提出了一种将 RAG 与个性化数据库系统相结合的个性化信息检索方法，通过在个人文档中标记关键字并将信息组织到基于上下文的类别中，使用户可以在其数据存储库中进行有效的搜索。刘彦宏[4]等通过使用 Jieba 分词，Word2Vec 模型对文本数据进行词嵌入，计算句子间的向量相似度并做重排序，帮助大语言模型快速筛选出最可靠可信的模型外部的医疗知识数据，再根据编写相关的提示词(Prompt)，可以使大语言模型针对医生或患者的问题提供令人满意的答案。José Benzinho [5]等创建了一个基于大语言模型会

话代理的解决方案，将用户查询与向量数据库中的邻近搜索相结合，并将其馈送到 LLM 中，同时 LLM 在其回复中会考虑与用户的对话历史。Alex Thomo [6]讨了 RAG 在增强 PubMed 数据库医学信息检索中的应用，通过将 RAG 与大型语言模型相结合，提高了向医疗专业人员提供的医疗信息的准确性和相关性。郝世博[7]等基于开源检索增强生成问答架构 FastGPT 集成的大模型文本表征、知识库检索和文本生成等核心能力，构建适用于校企技术合作信息推荐场景的智能问答系统并开展领域应用实践，整体问答流程简明且易于操作。张丽静[8]等设计并实现了基于大语言模型、LangChain 框架、pgVector 向量数据库以及表示学习等技术的智能客服系统原型，旨在利用大模型理解用户复杂的自然语言输入，随时随地以更高效、准确的方式回答用户所遇到的问题，在降低中邮网院客服人工成本的同时，提升用户体验感和满意度。Cheng Ye [9]通过使系统学习用户不同的搜索语义来解决增强电子病历(EMR)与 RAG 搜索引擎结合的问题，使 RAG 模型对特定问题生成具有临床意义的答案，强调了用户定制的学习排序方法在临床实践中的潜力。

通过梳理相关文献发现，目前学者从多行业多角度对 RAG 与 LLMs 相关内容进行研究，在供应链管理领域，由于市场环境瞬息万变、专业术语复杂且相关流程繁多，数据安全问题较为重要，传统的查询方式难以满足管理者和研究人员高效、准确获取信息的需求。相较于传统搜索引擎，面向供应链领域的智能问答模型能够提供更高质量的答案，大幅提升知识检索的效率。因此，开发一种基于先进技术的供应链问答模型显得尤为重要。为解决该问题，本文提出一种基于 LLMs、RAG 技术的供应链问答模型，能够充分利用供应链管理领域的大规模知识库，结合先进的自然语言处理技术和强化学习算法，实现对复杂供应链问题的深度理解和精准回答。

2. RAG 检索增强生成模型

2.1. 简介

检索增强生成(RAG)是一种结合了信息检索技术与语言生成模型的人工智能技术。该技术通过从外部知识库中检索相关信息，并将其作为提示(Prompt)输入给大型语言模型(LLMs)，以增强模型处理知识密集型任务的能力，如问答、文本摘要、内容生成等。RAG 模型由 Facebook Al Research (FAIR)团队于 2020 年首次提出，并迅速成为大模型应用中的热门方案。主要步骤主要分为以下三步：

第一步：检索。检索是 RAG 流程的第一步，从预先建立的知识库中检索与问题相关的信息。这一步的目的是为后续的生成过程提供有用的上下文信息和知识支撑。

第二步：增强。RAG 中增强是将检索到的信息用作生成模型(即大语言模型)的上下文输入，以增强模型对特定问题的理解和回答能力。这一步的目的是将外部知识融入生成过程中，使生成的文本内容更加丰富、准确和符合用户需求。通过增强步骤，LLM 能够充分利用外部知识库中的信息。

第三步：生成。生成是 RAG 流程的最后一步。这一步的目的是结合 LLM 生成符合用户需求的回答，生成器会利用检索到的信息作为上下文输入，并结合大语言模型来生成文本内容。

应用 RAG 技术的 LLM 可有效解决原生 LLM 的幻觉问题，即模型容易出现一本正经的“胡说八道”并提供虚假信息的现象；时效性问题，即不能回答时效性比较强的题目；数据安全问题，即用户需将信息数据上传至互联网。

2.2. 基于 RAG 的供应链智能问答模型

根据 RAG 基本原理以及供应链管理相关特点，构建基于 RAG 的供应链智能问答模型，主要分为三部分内容，包括数据准备与知识库构建部分，检索模块设计部分，生成模块设计部分，具体设计结构如图 1 所示。

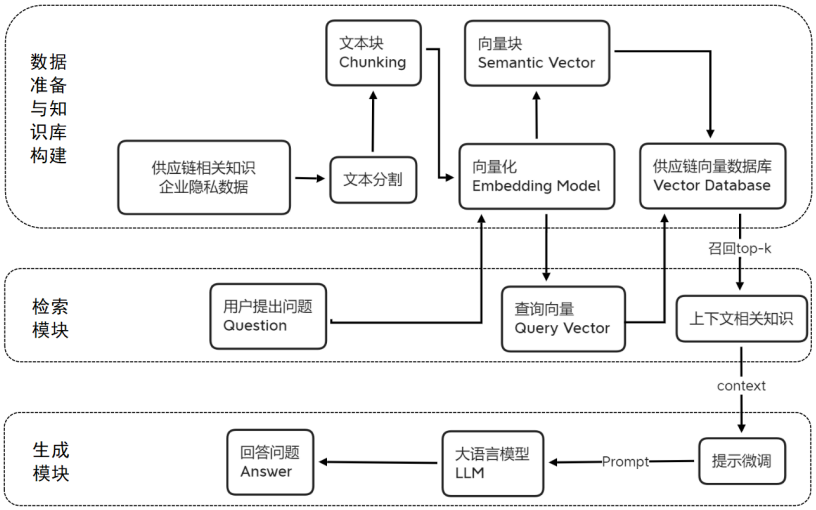

Figure 1. Supply chain intelligent question and answer model structure
图 1. 供应链智能问答模型结构

2.2.1. 数据准备与知识库构建

数据准备与知识库构建是一个离线的过程，主要是将供应链相关知识用文字数据图片等形式文本化后再进行向量化，并构建索引并存入数据库的过程。主要包括数据提取环节、文本分割环节、向量化环节以及数据入库等环节。

数据提取环节是对供应链相关数据进行加载，预处理与关键信息提取的过程。需调整所有数据为同一范式，并进行压缩与格式化，然后截取数据中的关键信息。文本分割环节主要是借助 langchain 的字符分割器，将供应链相关数据进行切割的过程。主要有两种分割方式，第一种方式是以“句”的粒度进行切分，保留一个句子的完整语义。常见切分符包括：句号、感叹号、问号、换行符等。第二种方式是固定长度分割，这种切分方式会损失很多语义信息，一般通过在头尾增加一定冗余量来缓解。向量化环节与数据入库环节是对分割后的数据进行 embedding，并写入数据库的过程。

2.2.2. 检索模块设计

检索模块的设计主要由两个部分构成：数据检索模块和 Prompt 注入微调模块。其中数据检索模块是在用户提出问题后，系统选择 embedding 模型，从上述构建的供应链管理数据库中检索出一个与问题最相关的内容作为回答的上下文，其中相似性的计算方式包括：余弦相似性、欧氏距离、曼哈顿距离等。而 Prompt 注入微调模块则通过在检索到的结果基础上，动态调整和优化输入提示，进一步提升模型在特定任务上的性能和准确性。这一设计不仅提高了系统的响应速度，还增强了模型对复杂问题的处理能力，从而在实际应用中能够提供更为精准的智能决策支持。

2.2.3. 生成模块设计

生成模块设计部分即将检索器，提示模板(Prompt Template)以及 LLM 有机结合在一起以实现更加高效和精准的答案生成过程。首先，检索器从海量数据中获取与用户查询相关的片段，这些片段包含了可能有助于解答问题的信息。接着，提示模板将这些检索到的相关片段与用户提出的原始问题进行融合，生成更加丰富和全面的上下文信息。这一过程不仅为模型提供了更广泛的知识背景，还帮助模型理解问题的上下文，从而提高其回答的准确性和相关性。最后使用 LLM 根据生成的上下文信息来生成最终的回答。通过这种方式，LLM 能够学习如何在具体问题的上下文中，结合检索到的知识，生成更加准确、实

用的回答。这种设计能够有效提升系统的智能化水平，使得生成的回答不仅更具针对性，还能适应不同场景和复杂问题的需求，从而增强系统的实用性和灵活性。

3. 实验设计与结果分析

3.1. 实验设置与数据集

本文使用的 LLM 模型与 Embedding 模型均为 qwen2-7b，向量数据库使用 LanceDB，设置 LLM Temperature 为 0.7，最大 Embedding chunk 长度为 8192，文本分割块长度为 8192，分块期间相邻文本块之间的最大字符重叠数量为 20。实验数据集包括供应链管理相关知识数据库以及模拟供应链公司运行数据库。评估数据集包括问题(question)，即用户输入的问题；答案(answer)即系统生成的答案；检索上下文(contexts)即根据提出的问题所检索到的与问题相关的文档；正确答案(ground_truths)即人类提供的问题真实答案。

3.2. 评价指标

本文选择 Ragas 评估方法对模型结果进行评估，评价指标主要包括忠实度(faithfulness)，答案相关性(Answer relevancy)，上下文相关性(Context relevancy)。

忠实度衡量生成的答案与给定的上下文的事实一致性，具体公式为：

$$F = \frac{|C|}{|A|} \tag{1}$$

其中 F——忠实度分数；C——能从给定上下文推断的答案数量；A——答案总数量。

答案相关性评估系统生成的答案与问题之间的关联程度，具体公式为：

$$A_r = \frac{1}{n}\sum_{i=1}^{n} sim(q, q_i) \tag{2}$$

其中 A_r——答案相关性分数；q_i——查询样本 i 的编码向量。

上下文相关性评估检索到上下文的相关性，根据问题与上下文计算得出，具体公式为：

$$C_r = \frac{|S|}{|R|} \tag{3}$$

其中 C_r——上下文相关性分数；S——检索上下文与答案相关的句子数量；R——检索上下文中句子总数。

3.3. 实验结果

本文使用 qwen2-7b 作为 LLM 模型进行 Ragas 评测，评测结果如表 1 所示。

Table 1. Ragas evaluation data
表 1. Ragas 评测数据

LLM	忠实度	答案相关性	上下文相关性
qwen2-7b	0.73	0.8	0.83

根据实验结果，qwen2-7b 模型忠实度分数，答案相关性分数，上下文相关性分数分别为 0.73、0.8、0.83。Ragas 评测结果总体较好，可作为供应链智能问答模型的 LLM 模型使用。

4. RAG 智能问答模型应用

根据上述模型构建以及 Ragas 评测结果，在知识库中导入供应链相关知识书籍、文献等资料例如《供

应链管理》[10]或企业内部模拟数据资料，对供应链智能问答模型进行回答准确程度、时效性和数据安全程度等方面的验证，共完成 200 余次人机交互测试，部分测试结果如图 2、图 3 所示。

Figure 2. Model accuracy test
图 2. 模型准确度测试

根据图 2 所示，当用户向系统提问“在《供应链管理》这本书中，福特公司的例子说明了什么”时，系统能够准确地理解并分析该问题。系统通过检索相关的书籍内容，结合上下文信息，生成了一个可靠的回答。与此同时，系统还标注了答案的来源，明确指出该回答是基于《供应链管理》一书中的相关章节或内容。这不仅展示了系统在信息检索和上下文理解方面的能力，也有效地增强了回答的透明度和可信度。通过提供答案来源，用户能够清楚地了解模型生成答案的依据，从而判断回答的准确性和可靠性。整体来看，系统生成的回答反映了较高的准确程度，能够有效满足用户的查询需求，并提供有价值的参考信息。

Figure 3. Model timeliness test
图 3. 模型时效性测试

根据图 3 所示，在知识库中导入企业模拟数据资料或文献后，向系统提问“宜家平均每年的货物运输量有多少，其中船舶、铁路、公路运输占比分别为多少？”时，系统对数据类问题做出精确回答，并标注出来源于文献[11]，体现了模型的良好时效性，即能够准确处理最新导入的数据资料，并将其有效整合用于回答。通过这种方式，系统展示了其强大的可迁移性，能够灵活应对不同领域和任务中的信息需求，从而为用户提供高质量、可靠的解答。

根据图 2、图 3 以及其他测试结果，供应链智能问答模型展现出强大的性能，能够准确解答一般性问题以及供应链领域的专业问题。无论是简单的常规问题，还是复杂的行业专业问题，模型都能迅速且精确地提供答案。这表明该模型在理解和处理供应链相关知识方面具备了高水平的智能，能够满足实际应用中多种场景的需求。其中，模型不受限于数据集的规模或内容。一旦知识库得到更新，系统即可迅速吸收新信息，适应新变化，回答针对新问题的需求。这种灵活性不仅增强了模型的时效性，还使其在面对动态变化的环境时，能及时提供更新的解决方案。此外，该系统采用了完全私有化的本地部署方式，用户可以根据需要将其应用于个人或企业内部环境中，从而确保对于不公开的敏感数据进行安全处理。由于模型完全在本地运行，不依赖于外部云端服务，这就有效避免了可能发生的隐私数据泄露问题。在回答涉及公司机密或个人隐私的相关问题时，系统能够通过内置的安全机制保障数据的隐私性和保密性，从而降低数据外泄的风险。在技术层面，模型成功缓解了原生 LLMs 所普遍存在的“幻觉”问题。传统 LLMs 可能在回答时产生不准确或错误的内容，而该供应链智能问答模型则通过与更新后的知识库相结合，确保其回答基于最新且准确的知识资源，避免了“幻觉”现象的发生。综合来看，该模型在处理复杂问题时展现出极高的准确性和可靠性，同时拥有私有化本地部署的特点，为用户提供了更高的安全保障，显著缓解了数据隐私泄露的潜在风险，有效提升了系统在实际应用中的时效性和实用性。

5. 结论

本文基于大模型、RAG 技术构建了一种关于供应链的智能问答模型，并对模型进行 Ragas 评测以及应用模拟测试。结果表明使用 qwen2-7b 作为 LLM 模型的 Ragas 评测结果较好，同时模型回答准确程度，专业程度较原生 LLMs 具有显著优势，模型不受限于数据集，时效性和可迁移性较高，模型与数据集可以完全本地化，数据私密程度较高。该研究为供应链相关企业或个人提供了一种问答效果更好且更为安全的参考模型，但对于 RAG 与 LLMs 在供应链领域的相关应用仍需进一步研究，在模型建立方面，可根据硬件条件使用较新的开源大模型例如 Qwen2-72b，使模型在自然语言理解，知识，数学以及多语言等多项能力上有更好的表现。在模型评价方面，可根据应用场景对更多 LLMs 进行 Ragas 评测，丰富评价指标。在数据获取方面，仅采用书籍或模拟企业运营数据，数据获取范围与类型存在局限性，可与相关企事业单位或个人建立连接，扩大数据的获取范围同时保持数据的时效性。

参考文献

[1] 洪亮, 郭瑶, 刘兴丽, 等. 基于 RAG 的煤矿安全智能问答模型[J]. 黑龙江科技大学学报, 2024, 34(3): 487-492.

[2] 关殿玺, 黄琨, 崔年治, 等. 基于大模型、RAG 和智能体技术的勘察岩土问答机器人研究[J]. 中国勘察设计, 2024(8): 101-104.

[3] Byun, J., Kim, B., Cha, K. and Lee, E. (2024) Design and Implementation of an Interactive Question-Answering System with Retrieval-Augmented Generation for Personalized Databases. *Applied Sciences*, **14**, Article No. 7995. https://doi.org/10.3390/app14177995

[4] 刘彦宏, 崔永瑞. 基于 Word2Vec 模型与 RAG 框架的医疗检索增强生成算法[J]. 人工智能与机器人研究, 2024, 13(3): 479-482.

[5] Benzinho, J., Ferreira, J., Batista, J., Pereira, L., Maximiano, M., Távora, V., *et al*. (2024) LLM Based Chatbot for Farm-to-Fork Blockchain Traceability Platform. *Applied Sciences*, **14**, Article No. 8856. https://doi.org/10.3390/app14198856

[6] Thomo, A. (2024) PubMed Retrieval with RAG Techniques. In: Mantas, J., *et al*., Eds., *Ebook*: *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, IOS Press, 652-653. https://doi.org/10.3233/shti240498

[7] 郝世博, 史东昊, 唐裕晨. 基于开源 RAG 架构的校企专利技术合作问答应用研究[J]. 技术与市场, 2024, 31(5): 1-11.

[8] 张丽静, 杜冬梅, 刘庆芳, 等. 基于 LLM 和 RAG 的中邮网院智能客服系统研究[J]. 邮政研究, 2024, 40(4): 66-72.

[9] Ye, C. (2024) Exploring a Learning-to-Rank Approach to Enhance the Retrieval Augmented Generation (Rag)-Based Electronic Medical Records Search Engines. *Informatics and Health*, **1**, 93-99. https://doi.org/10.1016/j.infoh.2024.07.001

[10] 苏尼尔·乔普拉. 供应链管理[M]. 北京: 中国人民大学出版社, 2021.

[11] 王倩. 宜家供应链管理体系研究[D]: [硕士学位论文]. 青岛: 中国海洋大学, 2013.