基于大语言模型和知识库的工程施工规范问答系统构建研究

肖皓

北京建筑大学城市经济与管理学院，中国·北京 100032

摘要：本研究利用 Langchain 框架与 Qwen-max 大模型结合，通过搭载施工规范知识库创建一个施工规范问答系统，通过引入住房和城乡建设部发布的施工规范类文件和工程论坛中的问答对创建了施工规范本地知识库，通过施工规范本地知识库增强了以 Qwen-max 模型为基座模型的施工规范问答系统的应用能力，实现 Qwen-max 模型在施工规范问答领域的应用。使用评价响应的事实准确性（FA）和完整性（CR）两项指标分别评价施工规范问答系统、Qwen-max 模型和 Deepseek-R1 模型对施工规范测试集的响应，施工规范问答系统的准确性指标远强于 Qwen-max 模型和 Deepseek-R1 模型，完整性指标稍逊于 Deepseek-R1 模型，但也强于 Qwen-max 模型，总体来说施工规范问答系统的表现强于 Qwen-max 模型和参数量更大的 Deepseek-R1 模型，说明通过检索增强生成（RAG）的方法能提高通用模型在施工规范问答领域的应用能力。

关键词：大语言模型；RAG 施工规范；问答系统；本地知识库

Research on the Construction of an Engineering Construction Code Q&A System Based on Large Language Models and Knowledge Bases

Xiao Hao

School of Urban Economics and Management, Beijing University of Civil Engineering and Architecture, China Beijing 100032

Abstract: This study combines the Langchain framework with the Qwen-max large model to create a construction code question-and-answer system by incorporating a construction code knowledge base. A local construction code knowledge base was created by integrating documents on construction codes issued by the Ministry of Housing and Urban-Rural Development and Q&A pairs from engineering forums. The construction code local knowledge base enhances the application capability of the Qwen-max-based construction code question-and-answer system, enabling the application of the Qwen-max model in the field of construction code Q&A. The factual accuracy (FA) and completeness (CR) of the responses were used to evaluate the responses of the construction code Q&A system, the Qwen-max model and the Deepseek-R1 model to the construction code test set. The accuracy of the construction code Q&A system is significantly better than that of the Qwen-max and Deepseek-R1 models, while its completeness is slightly lower than that of Deepseek-R1 but still better than that of Qwen-max. Overall, the construction code Q&A system outperforms the Qwen-max model and the larger-parameter Deepseek-R1 model, indicating that retrieval-augmented generation (RAG) methods can improve the application capability of general models in the field of construction code Q&A.

Keywords: Large language model; RAG construction standards; Q&A system; Local knowledge base

0 引言

目前基于大语言模型构建问答系统主流范式有基于知识图谱、基于微调、基于检索增强生成（RAG，retrieval augmented generation）三种类型[1]。基于知识图谱的问答系统利用实体和关系进行查询、推理和答案生成[6]。但在知识图谱的构建过程中，实体的识别与链接、关系抽取和动态更新等过程复杂，需要大量专家成本与计算资源，同时基于知识图谱增强检索生成方法（RAG）向问知识问答系统方向的适应性和拓展性较差。微调方法能训练专业领域大模型，但微调可能产生过拟合、灾难性遗忘等问题，需要特定领域数据库构建、层冻结、层级学习率衰减等复杂手段优化，并且很难保证整体运行稳定性[7-9]。现有大语言模型在垂直领域生成的结果不符合事实或者提示词内容，这种现象称之为幻觉（hallucination）[10]。

RAG 是一种结合大语言模型和领域知识的有效方法，可以缓解幻觉问题，其核心思想是从本地知识库中检索与

问题相似性最高的文段[2]，并使用 LLM 通过上下文学习的方式生成回答。基于知识库的大模型问答系统在中医药、法律、矿物等专业取得了较大的进展，能够为领域内研究人员或者从业人员提供问答支持[3-5]。

相比于基于知识图谱和微调构建问答系统，基于 RAG 更加简洁、经济、高效，且施工规范语料丰富，适合创建知识库，所以，基于 RAG 创建施工规范知识库和构建施工规范问答系统具备技术上的便捷性与可行性。

1 资料与方法

1.1 数据集构建

本研究在住房和城乡建设部工程建设标准化信息网上搜集 20 份施工及验收标准，包括大体积混凝土施工标准、民用建筑工程修缮施工标准、住宅装饰装修工程施工规范等常用标准及规范，之后对文档进行了切分、保留了术语解释、基本规定和详细规范，本地知识库文件详细情况如表 1 所示、

其中施工规范的问答对来自于一级建造师考试论坛、施工经验交流网站的提问与回答，表 1 中的内容共同构成了施工规范问答系统的知识库数据集。

表1 本地知识库文件

数据集类型	文件形式	页数	大小（MB）
施工规范	PDF	902	260
问答对	Text	30	0.3

1.2 知识向量库构建

构建工程施工规范问答系统知识向量库包括对数据集文本的的加载、读取和分割。其中文本分割是影响知识向量库质量的关键因素，将文本均匀分割会导致文本语义断裂，在中文语境中，一个完整的语义片段往往集中在同一个段落中，由于一个段落内字符数通常不会过长也能降低了模型输入超限的风险，所以本研究采用基于段落分割文本的策略来保证分割后文本语义及情境的完整。

构建知识库需要使用到向量模型将分割后的文本向量化后存储入向量数据库，在本研究中，每一本规范对应一个知识库，通过 API 调用阿里云通义实验室推出的 text-embedding-v1 模型将分割后的文本转化为 1536 维的向量表示后存储入对应规范的知识库，用于后续的知识匹配查询。

1.3 知识匹配

知识匹配需要通过调整历史对话轮数、知识匹配分数阈值、知识库管理等知识库配置来实现，增加历史对话轮数可以增强模型的情境学习能力，提高知识匹配分数阈值会增加知识的匹配精度，同时也减少知识的匹配条目，针对不同领域构建的不同知识库则可以通过知识库管理来使模型实现不同领域的知识问答。

在知识匹配过程中，核心思想是通过计算问题向量与知识向量的余弦相似度来实现知识的关联性匹配。具体过程是对比问题向量 Q（Q1，Q2，…，Qm）分别于知识库中的文本向量 T（T1，T2，…，Tm）的余弦相似度。余弦相似度越高代表问题与知识文本越匹配[3]。公式如下：

$$similarity(Q,T_j)_{j=1}^{j=n}=\left(\frac{\sum_{i=1}^{m}(Q_i\times T_{F_i})}{\sqrt{\sum_{i=1}^{m}(Q_i)^2}\times\sqrt{\sum_{i=1}^{m}(T_{F_i})^2}}\right) \quad (1)$$

知识的匹配过程具体是筛选出达到余弦相似度阈值的前 k 条知识，此过程运用到了 TOPk 匹配算法，算法如图 1 所示，最终筛选出与用户提问相似性最高的 k 条知识并展示给用户。

```
输入：Problems in the dialogue system (Problems)
输出：k pieces of knowledge in the knowledge base with the highest similarity to the problem and above the score - threshold(Knowledge segment)
1.def top_k(Documents, k): //用于取得分最高的k条文档
2.sorted_documents = sorted(Documents, key = lambda x:x.score, reverse = True)//按照得分进行降序排序
3.top_k_documents = sorted_documents[:k]//获取前k条文档
4.return top_k_documents//返回得分最高的k条文档
5.Select the number of knowledge entries to match, k.//选择要匹配的知识条目数量k
6.Select the score_threshold used for filtering knowledge entries.//选择用于过滤知识条目的分数阈值
7.Query = Problems;//将问题设为查询(Query)
8.Initialize an empty list Result;//初始化一个空列表Result
9foreach Text in List[Document]do//对于每个文本，计算其与查询的相似度
10.if calculate_similarity(Query, Text)>=score_threshold then
11.Add Text to Result;//如果相似度高于阈值，将文本添加到结果列表中
12.end
13.end
14.return top_k(Result, k);//返回得分最高的k条结果
```

图1 Topk算法图

1.4 LangChain 框架

LangChain 是一个开源框架，专为简化基于大型语言模型（LLM）的应用程序开发而设计。它通过模块化组件和标准化接口，能提供 RAG 相关 API 工具，在大语言模型的应用领域有重要作用[12]。它能连接生成模型与数据库并实现模型与数据库的交互，本研究利用了 LangChain 框架实现了 Qwen-max 模型与外部知识库的集成与交互，完成了 Qwen-max 在施工规范的增强检索生成。

图 2 施工规范问答系统界面

1.5 大模型输出

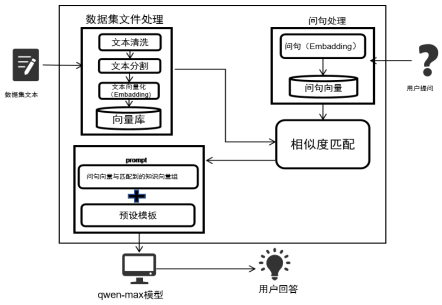

图 3 施工规范问答系统总体流程图

通过 TOPk 匹配算法筛选出 K 个相似度靠前的知识条目形成最终的检索结果，随后将检索结果与预设的模板组成提示词输入大语言模型，利用大语言模型的理解与推理能力输出答案。在本研究中，通过 API 调用阿里云百炼平台的 Qwen-max 模型，将检索结果与预设的模板组成提示词传入 Qwen-max 模型，利用其理解与推理能力得到输出，具体流程如图 3 所示。

2 实验

大模型在面对施工规范的开放性问答任务时，生成的响应是自然语言文本，回答质量不能用简单的正确或错误来判定，同时，在施工规范问答这种依赖垂直领域知识来进行问答对话的任务中，传统的评估指标也难以评价回复质量。

为了更好地评价大模型在建筑领域的输出质量，覃思中等人通过设置判断题和填空题让大模型作答，再分别通过机器和人工判定答案的方法来评价大模型的响应质量[13]。丁志坤等使用 GPT4. 0 作为裁判模型对测试模型在 BIM 正向设计中的实际问题从技术和工程知识、法规和标准等 6 个维度打分对回复质量进行评价[14]。本研究计划从准确性和完整性两个方面对施工规范问答系统、Qwen-max、Deepseek-R1 模型进行对比实验。

2.1 案例实验

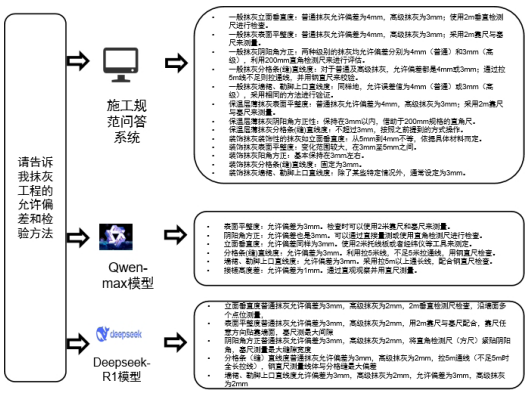

图 4 实验过程展示图

对施工规范问答系统与未加载知识库的 Qwen-max 模型以问题“请告诉我抹灰工程的允许偏差和检验方法”分别提问，未加载知识库 Qwen-max 和施工规范问答系统的回复如图 4 所示

2.2 案例分析

查阅建筑装饰装修工程质量验收标准（GB50210-2018）知，在抹灰工程中，一般抹灰、保温层抹灰、装饰抹灰的的允许偏差和检验方法均不相同，分别如表 2 表 3 表 4 所示：

根据建筑装饰装修工程质量验收标准（GB50210-2018）规定可知，Qwen-max 模型和 Deepseek-R1 模型只回答了一般抹灰工程中的高级抹灰允许的偏差以及检测方法，并未区分抹灰类型予以回答，且在 Deepseek-R1 模型的恢复中出现了事实性错误，一方面是回复内容出现了错误，另一方面提示的《GB50210-2025》规范也并不存在，而施工规范问答系统则明确了抹灰工程的允许偏差和检验方法根据不同类型的抹灰而有所区别，并且分类型给予了回复。

在响应的完整性和事实准确性均优于 Qwen-max 模型与 Deepseek-R1 模型，可见加载的施工规范知识的施工规范问答系统的响应更加准确和专业，也更加符合使用者查阅的实际需求。

2.3 施工规范问答系统与、Qwen-max 模型、Deepseek-R1 模型性能对比

本研究使用相同数据集对三个实验对象进行测试，手工构建了 100 个施工规范的问题，以分析该方法搭载了知识库的施工规范问答系统、Qwen-max 模型、Deepseek-R1 模型对于施工规范类问题响应的准确性和完整性，相关问题示例和参数如表 5 所示：

图中蓝、绿、黄分别表示施工规范问答系统、Qwen-max 模型、Deepseek-R1 模型，用上述测试集分别测试，百分比表示在准确性或完整性上的到最优评价的测试问题数占测试集中问题总数的比例。

图 5 的结果表明施工规范问答系统的准确性和完整性要优于 Qwen-max 模型与 Deepseek-R1 模型，但是 Deepseek-R1 模型比施工规范问答系统的基座模型 Qwen-max 模型训练参数更大，性能更强。因此，如图 6 所示，施工规范问答系统的回复完整性不如 Deepseek-R1 模型，但稍强于没有加载知识库的 Qwen-max 模型。由于施工规范问答系统加载了施工规范知识库，其在施工规范领域回复的专业性远远强于没有加载知识库的 Qwen-max 模型和

表2 一般抹灰的允许偏差和检查方法

编号	项目	最大允许偏差值		检查方法
		普通抹灰	高级抹灰	
1	表面平整度	4mm	3mm	用2m规格的垂直度检测尺检查表面平整度
2	阴阳角方正度	4mm	3mm	用规格为200mm的直角检测尺检查
3	立面垂直度	4mm	3mm	用2m规格的靠尺和塞尺检查立面垂直度
4	分隔条、分隔缝、墙裙、勒脚以及上口的直线度	4mm	3mm	用钢直尺检查，拉5m长直线，当检查部位长度不足5m时，拉通线

表3 保温层抹灰的允许偏差和检验方法

编号	项目	最大允许偏差值	检查方法
1	表面平整度	3mm	用规格为2m的垂直检测尺检查
2	分隔条或分隔缝的直线度	3mm	用钢直尺检查，拉5m长直线，当检查部位长度不足5m时，拉通线，
3	立面垂直度	3mm	用规格为2m的垂直检测尺检查
4	阴阳角方正	3mm	用规格为200mm的直角检测尺检查

表 4 装饰抹灰的允许偏差和检验方法

编号	最大允许偏差值				检验方法
	水刷石	斩假石	干粘石	假面砖	
1	5mm	4mm	5mm	5mm	用规格为2m的垂直检测尺检查
2	3mm	3mm	5mm	4mm	用规格为用2m的靠尺和塞尺检查
3	3mm	3mm	4mm	4mm	用规格为用200mm直角检测尺检查
4	3mm	3mm	3mm	3mm	用钢直尺检查，拉5m长直线，当检查长度不足5m时，拉通线
5	3mm	3mm	–	–	用钢直尺检查，拉5m长直线，当检查长度不足5m时，拉通线

表 5 测试集问题示例

编号	问题
1	请告诉我抹灰工程的允许偏差和检验方法
2	采用电弧焊方法的栓钉焊接接头最小焊脚尺寸是多少
3	预埋件和预留孔洞的安装允许偏差是多少
4	强夯地基质量检验标准有哪些项目
5	土工合成材料地基的质量检验标准是什么
6	建筑防腐工程结合层厚度和灰缝宽度是怎样规定的
7	单位工程的划分原则是什么
8	建筑节能工程的分部工程有哪些
9	木结构外观质量如何分级验收
10	在滑坡地段挖方有哪些规定

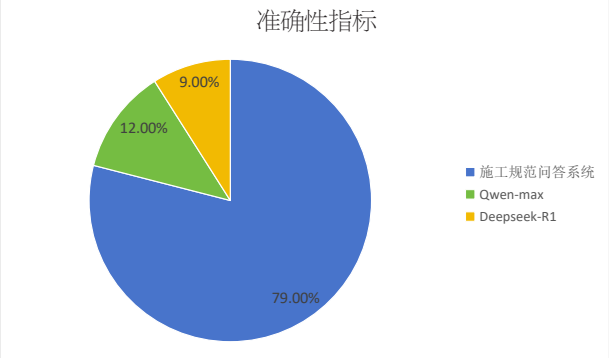

图 5 施工规范问答系统、Qwen-max、Deepseek-R1响应准确性对比图

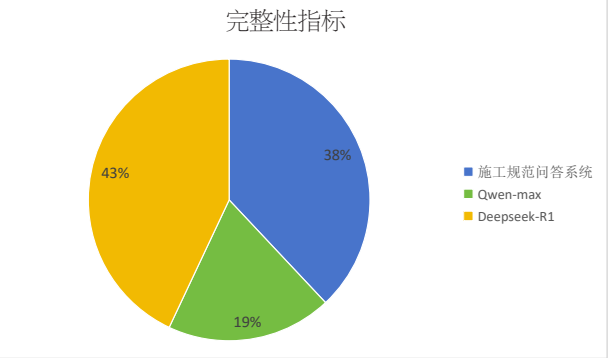

图 6 施工规范问答系统、Qwen-max、Deepseek-R1响应完整性对比图

Deepseek-R1模型，总体来看，施工规范问答系统的整体表现优于未加载知识库的Qwen-max模型和Deepseek-R1模型

3 结语

在这项研究中，使用RAG方法基于Qwen-max模型开发了一个施工规范问答系统，为了这一研究，首先搜集了大量知识文件，然后对其进行文本清洗、文本分割、向量化，最后存入向量库，通过TOPk匹配算法匹配到相关知识条目。本研究发现，加载了垂直领域知识的施工规范问答系统在施工规范问答领域的回复整体质量好于没有加载知识库的基座模型和训练参数量远大于其基座模型但未加载知识库的Deepseek-R1模型，因此，在特定领域，为大模型加载知识库可以提高模型在特定领域回复的完整性和准确性，妥善处理知识文本和恰当的知识匹配算法是提高模型问答质量方案之一，下一步的工作是创新文本清洗与文本分割方法，提高知识库的质量。

参考文献：

[1] 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]. 北京邮电大学学报，2024,47(04):50-56.DOI:10.13190/j.jbupt.2023-279.

[2] Wang S H, Xu Y C, Fang Y W, et al. Training Data is More Valuable than You Think: A Simple and Effective Method by Retrieving from Training Data[A]. //The 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)[C]. Stroudsburg: ACL, 2022: 3170-3179.

[3] 王文湖，韦昌法 . 基于大语言模型和知识库的阿尔茨海默病智能问答系统构建研究[J]. 世界科学技术 - 中医药现代化，2025,27(03):856-866.

[4] 李明达，邸洪波，孙媛媛等 . 基于法条检索的生成式法律问答研究[J/OL]. 山西大学学报 (自然科学版),1-13[2025-05-18]. https://doi.org/10.13451/j.sxu.ns.2024159.

[5] 季晓慧，刘成健，杨眉等 . 大语言模型及其在矿物问答系统中的应用[J/OL]. 矿物岩石地球化学通报，1-9[2025-05-18].http://kns.cnki.net/kcms/detail/52.1102.P.20250305.1707.001.html.

[6] 乔少杰，杨国平，于泳等 . QA-KGNet: 一种语言模型驱动的知识图谱问答模型[J]. 软件学报，2023,34(10):4584-4600.DOI:10.13328/j.cnki.jos.006882.

[7] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(13): 3521-3526. DOI:10.1073/pnas.1611835114.

[8] TINN R, CHENG H, GU Y, et al. Fine-tuning large neural language models for biomedical natural language processing[J]. Patterns, 2023, 4(4): 100729. DOI:10.1016/j.patter.2023.100729.

[9] MOSBACH M, ANDRIUSHCHENKO M, KLAKOW D. On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines[EB/OL]. (2020-06-16)[2024-06-30]. arXiv:2006.04884v3.

[10] JI Z W, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[EB/OL].2022:2202.03629.[2024-06-30]. https://arxiv.org/abs/2202.03629v7.

[11] CHEN J W, LIN H Y, HAN X P, et al. Benchmarking large language models in retrieval-augmented generation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16): 17754- 17762. DOI: 10.1609/aaai.v38i16.29728.

[12] Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey. arxiv preprint arxiv:2312.10997, 2023.

[13] 覃思中，郑哲，顾燚等 . 大语言模型在建筑工程中的应用测试与讨论[J]. 工业建筑，2023,53(09):162-169.DOI:10.13204/j.gyjzg23081006.

[14] 丁志坤，李金泽，刘明辉 . 基于大语言模型的BIM正向设计问答系统研究[J]. 土木工程与管理学报，2024,41(01):1-7+12.DOI:10.13579/j.cnki.2095-0985.2024.20240046.

作者简介：肖皓（1999-），男，湖南湘潭人，工程管理专业在读硕士研究生，研究方向：智能建造及AI大模型赋能城市更新。