基于 RAG 的建筑施工安全规范智能咨询系统设计

颜斌斌
北京建筑大学城市经济与管理学院，中国 · 北京 100044

摘要：本研究针对当前建筑施工领域中安全规范文档繁杂、一线管理与施工人员检索效率低下且难以快速获取精准指导的问题，设计了一款基于检索增强生成（RAG, Retrieval-Augmented Generation）的智能咨询系统。该系统以大语言模型（LLM）为核心处理引擎，结合向量数据库技术，能够对海量的建筑施工安全国家标准、行业规范及企业内部制度进行深度解析与存储。通过用户自然语言提问，系统利用 RAG 技术在本地知识库中进行语义检索，并将匹配到的规范原文作为上下文约束 LLM 的生成结果，从而有效克服了大语言模型的“幻觉”问题，实现了高准确率、低延迟的智能问答交互。本设计极大降低了安全管理的信息获取成本，提升了现场作业的规范化程度，为建筑施工行业的智能化安全管理提供了切实可行的技术方案。

关键词：检索增强生成 (RAG)；大语言模型 (LLM)；建筑施工安全；向量数据库；智能咨询

Design of an Intelligent Consultation System for Construction Safety Regulations Based on RAG

Yan Binbin
School of Urban Economics and Management, Beijing University of Civil Engineering and Architecture, China Beijing 100044

Abstract: Aiming at the problems of complex safety specification documents, low retrieval efficiency of front-line management and construction personnel, and the difficulty in quickly obtaining precise guidance in the current building construction field, this study designs an intelligent consulting system based on Retrieval-Augmented Generation (RAG). The system takes the Large Language Model (LLM) as the core processing engine, combined with vector database technology, to deeply analyze and store massive national standards for building construction safety, industry specifications, and internal corporate regulations. Through users' natural language queries, the system utilizes RAG technology to perform semantic retrieval in the local knowledge base, and uses the matched original text of the specifications as context to constrain the generation results of the LLM. This effectively overcomes the "hallucination" problem of large language models and realizes high-accuracy, low-latency intelligent question-and-answer interaction. This design greatly reduces the information acquisition cost of safety management, improves the standardization of on-site operations, and provides a feasible technical solution for intelligent safety management in the building construction industry.

Keywords: Retrieval-augmented generation (RAG); Large language model (LLM); Building construction safety; Vector database; Intelligent consultation

0 引言

随着建筑行业规模的不断扩大和施工技术日趋复杂，施工安全管理的重要性日益凸显 。在现代工地，超高层建筑、深基坑工程以及多工种交叉作业已成常态，这使得施工现场的安全风险变得动态且多变。传统的安全管理模式高度依赖管理人员的个人经验，以及对纸质或电子版规范手册的翻阅 。然而，现行建筑施工安全规范（如《建筑施工安全检查标准》）[1] 条款繁多、专业性强。在实际作业中，这种模式正面临严峻挑战。首先是效率问题：在面临突发情况或复杂工况（如塔吊遇强风、脚手架变形）时，管理人员往往需要在成千上万条规范中寻找应对依据，人工翻阅速度慢，极易延误最佳处置时机。其次是理解偏差：安全条款大多采用法律与技术性语言编写，非专业人员或经验不足的施工员在解读时，容易产生主观误读，导致现场执行的措施与标准的偏差，埋下事故隐患。

近年来，大语言模型（LLM）[2] 在自然语言处理领域展现出强大的理解和生成能力。然而，若将其直接应用于严谨的工业安全领域，存在两个关键问题：一是知识更新不及时，模型无法自动感知最新修订的国家标准或企业内部规定 ；二是“幻觉”问题，模型有时会编造看似合理但

完全脱离事实的错误指令 。在“生命至上”的建筑安全场景下，哪怕 1% 的信息差错都可能引发无法挽回的后果。

鉴于此，本文提出了一种基于 RAG 架构[3] 的智能咨询系统总体设计方案。该系统不再让大模型凭空记忆知识，而是给它配备一个“建筑安全规范知识库”，通过文档解析、向量化存储和精准的语义检索，确保模型在生成答案时必须基于真实的规范原文。系统涵盖了提示词工程、大模型推理及极简的前端交互界面，致力于为一线施工人员提供一个即问即答、有据可查的安全专家系统，从而真正实现建筑施工安全的数字化智能辅助 。

1 智能咨询系统总体设计

图1 基于 RAG 的建筑施工安全规范智能咨询系统总体设计框架

本设计以建筑施工现场的实际需求为研究对象，提出了一套基于 RAG 技术和微服务架构的软件系统方案。系统主要功能包括安全规范的智能检索、自然语言问答、多轮对话上下文理解以及知识库的动态更新。总体设计框架如图 1 所示。根据功能划分，系统架构可主要分为五个模块：数据预处理模块、向量检索引擎模块、核心大模型生成模块、系统管理与评估模块以及前端交互模块。

数据预处理模块：主要负责收集建筑施工安全相关的国标、行标及企业规定，进行文本的清洗、段落切分（Chunking）以及格式统一，为后续的向量化做准备。由于建筑规范中包含大量表格和特殊排版，本模块需具备强大的版面分析能力。

向量检索引擎模块：将切分后的文本块通过 Embedding 模型转化为高维向量，并存入向量数据库中。当用户发起提问时，该模块负责对问题进行同样的向量化计算，并在数据库中进行相似度检索，召回最相关的规范片段。

核心大模型生成模块：系统的大脑。它负责接收前端传入的用户问题，并融合向量检索模块召回的相关规范文本，通过预设的系统提示词模板重组后，交由 LLM 进行逻辑推理与自然语言组织，输出最终的准确回答。

系统管理与评估模块：用于管理员对知识库内规范文档的增删改查。当有新的安全标准出台时，可通过此模块进行热更新，确保系统提供的信息具备时效性。同时监控系统的回答准确率。

前端交互模块：作为用户与系统沟通的桥梁，通常设计为 Web 端或移动端 App，提供类似聊天界面的极简交互体验，支持语音输入和文本输入。

2 系统核心组件选型与设计

智能咨询系统的“硬件”在于其底层的算力支撑与核心技术组件的选型 。本设计在组件选择上兼顾了系统的响应速度、推理准确性以及部署成本 。在此基础上，系统架构还充分考虑了建筑施工企业对数据安全、系统高并发以及高可用性的工程化需求。

2.1 核心大语言模型 (LLM)

本设计选用具有强大中文理解能力的开源大语言模型（如 Qwen [4]）作为核心基座。选用该系列模型不仅因为其在中文自然语言基准测试中表现优异，更因为它具有强大的上下文理解和复杂逻辑推理能力，能够更好地适配 RAG 架构下的规范文本重组任务。在实际部署中，通过采用 8-bit 或 4-bit 量化技术，大幅压缩了模型运行所需的显存占用，使其能够在消费级或入门级企业 GPU（如 RTX 4090 或 A10）上流畅运行 。这种本地化的轻量级部署方案，不仅极大降低了系统的硬件门槛和运营成本 ，更重要的是实现了核心数据的物理隔离，彻底避免了企业内部安全规范、内部制度等敏感数据上传至公有云所带来的数据泄露风险。

2.2 向量数据库

为了实现对海量切分文本的毫秒级检索，本设计选用 Milvus [5] 作为底层的向量数据库引擎。建筑施工安全规范种类繁多且更新频繁，对底层数据库的扩展性要求极高。Milvus 作为云原生的向量数据库，支持百亿级向量的存储与快速近似最近邻（ANN）搜索。其分布式架构允许系统随着知识库的扩容而弹性扩展。在业务处理时，该数据库能够高效处理用户查询，将安全规范文档的文本嵌入向量与用户提问的语义进行匹配。相较于传统关系型数据库的全文检索，Milvus 能在保证检索精度的同时，维持较低的查询延迟，确保现场管理人员能够即查即得。

2.3 文本嵌入 (Embedding) 模型

Embedding 模型的质量直接决定了 RAG 系统的检索召回率。本系统采用 BGE (BAAI General Embedding)[6] 系列大模型，该模型在中文语义表征方面表现优异 。在实际施工现场，一线工人的提问往往高度口语化，且包含大量特定的专业词汇。BGE 模型能够精准捕捉施工专业术语的深层语义，避免传统基于关键词匹配导致的漏检问题 。例如，当用户搜索“防坠落措施”时，该模型能准确识别包

含“安全带挂靠”“临边防护”等未直接包含搜索词但语义高度相关的规范条款。这种强大的特征提取能力，使得系统能够生成高质量的向量，大幅提升后续检索的准确度和鲁棒性。

2.4 数据存储与缓存组件

系统采用 MySQL 数据库存储用户的账号信息、历史对话记录以及文档的元数据（如文档来源、生效日期等）。作为成熟的关系型数据库，MySQL 可确保系统基础业务数据的稳定与规范溯源的可控性。同时，引入 Redis 作为缓存层，对于高频的常规安全提问（如“施工现场必须佩戴的安全护具包含哪些”）进行答案缓存 。这种架构设计，能够有效拦截重复性或通用型查询请求，从而减少大语言模型的重复推理计算，降低算力开销，并提升系统在突发流量下的并发响应能力 。

3 系统软件工作流程与功能设计

3.1 核心检索与生成主流程

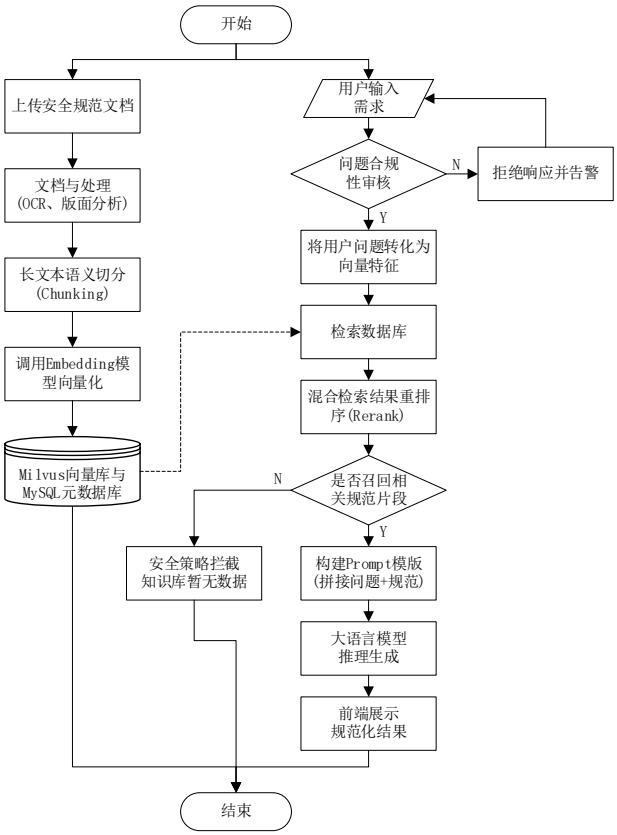

图2 系统软件工作总流程图

智能咨询系统的软件运行逻辑可解耦为离线与在线两条主干链路。离线链路侧重于安全规范文档的结构化处理与知识资产沉淀；在线链路则聚焦于用户意图的精准捕捉与大模型的受控生成。系统总体的软件数据流转与工作框架如图 2 所示。在离线阶段，系统通过解析、切片与向量化处理，将实体安全规范转化为大模型可读的底层资产；在在线咨询阶段，系统接收用户提问并实时将其向量化，在知识库中匹配最相关的规范原文。为确保安全领域的严谨性，系统加入了合规性审核与安全策略拦截机制，当知识库未命中时主动拒绝回答，从而在工作流层面彻底阻断了由于大模型“幻觉”引发的安全事故风险。

3.2 离线知识库构建与预处理流程

本流程主要面向系统管理员，是保障整个 RAG 系统回答准确性的数据底座。当最新的国家安全标准或企业内部施工规范发布时，管理员通过后台将 PDF、Word 或 HTML 格式的文档批量导入系统。

由于建筑规范中通常包含大量的复杂表格、图注及非标准排版，单纯的文本提取极易造成信息丢失。因此，系统首先调用版面分析与 OCR 引擎剥离非文本元素，并进行数据清洗。随后，系统采用“带重叠量的滑动窗口策略”对长文本进行语义切分（Chunking）。例如，设定分块长度为 500 个字符，并保留 50 个字符的重叠量，以此确保相邻规范条款之间语义的连续性与完整性。切分完成的文本块经由 Embedding 模型转化为高维特征向量，并持久化写入 Milvus 向量数据库中；同时，文档的元数据（如规范名称、发文单位、生效时间等）同步存入 MySQL 数据库以备溯源。

3.3 在线问答检索与大模型生成流程

在线问答流程是系统面向一线施工人员的核心业务链路。当用户通过前端发起自然语言提问时（如：“塔吊遇六级大风应如何处理？”），系统首先对问题进行安全合规性与领域相关性审核。若问题脱离建筑施工范围，系统将触发防御策略并主动拒绝回答；若判定合规，系统则调用与离线阶段相同的 Embedding 模型，将用户的自然语言问题转化为高维特征向量。接着，检索引擎以该向量为查询条件，在 Milvus 数据库中发起 Top-K 近似最近邻（ANN）检索，迅速召回语义相似度最高的若干条规范原文片段。最后，系统将“用户原问题”与“召回的规范片段”按预设的 Prompt（提示词）模板进行拼接重组，共同作为上下文输入给核心大语言模型（LLM）。在已知规范文本的强力约束下，LLM 负责梳理逻辑、提炼重点，并生成符合安全标准的自然语言回答，从而实现“有据可查”的智能咨询。

3.4 检索增强 (RAG) 核心策略设计

在实际业务场景中，由于建筑施工领域包含大量高度专业化的术语（如“满堂支撑架”“悬挑式操作平台”

等），单一的向量语义检索在面对生僻词或极简短句提问时，容易出现匹配偏移或召回精度不足的问题。为了进一步提升复杂工况下系统的检索精准度，本设计在标准 RAG 流程基础上引入了“混合检索与重排序（Rerank）”策略。系统将基于关键词词频的传统 BM25 算法与基于 Embedding 模型的深度语义检索相结合，对知识库进行双路召回。融合后的粗排结果再经过特定的 Rerank 交叉编码器模型进行细粒度打分与重新排序。该策略能够有效弥补单一检索方式的短板，确保最权威、最核心的安全条款能够被优先提取并输送给大模型进行推理。

3.5 前端界面设计

前端交互模块作为人机沟通的桥梁，基于 Vue 等跨平台框架开发，全面适配 Windows 桌面端与移动智能终端，以适应施工现场复杂多变的使用环境。考虑到一线作业人员的操作习惯，界面设计采用极简的对话流布局，支持文本直输与语音转写功能，极大降低了系统的使用门槛。在展示逻辑上，系统秉持“结论先行，依据兜底”的原则：除了以流式打字效果输出大模型生成的直白解答外，系统还会在回答末尾以独立的高亮卡片形式，精准附上该回答所参考的具体规范条款出处（例如展示：参考依据：《建筑施工高处作业安全技术规范》[7] JGJ 80-2016 第 X.X 条）。这种双轨制的展示形态既保证了信息获取的高效性，又赋予了底层安全指导极高的公信力与可追溯性。

4 系统测试

测试开始前，首先启动数据库服务与大模型推理服务端，控制台输出成功加载权重的日志后说明系统初始化成功。测试分为两个阶段：

基础知识库测试：向系统中导入包含土方工程、脚手架、高处作业等 10 部国家及行业核心安全规范文档。观察预处理模块是否能够正确切分并入库。

多维度问答测试：由测试人员模拟现场施工员输入复杂问题，如：“塔吊遇到 6 级大风还可以继续吊装作业吗？”。系统测试结果显示：系统能够在 2 秒内响应，并在回答中明确指出“根据相关规定，遇有六级及以上大风或恶劣气候时，应停止塔式起重机露天起重吊装作业”，同时准确附上规范出处。在百余次涵盖不同工种的安全测试中，凭借 RAG 架构的约束，系统对安全规范的解释准确率达到 95% 以上，且未出现严重违背安全常识的幻觉回答。

5 结语

本文研究并设计了一种基于 RAG 架构的建筑施工安全规范智能咨询系统，该系统通过将大语言模型的生成能力与外挂垂直知识库的精准度相结合，实现了自动化、专业化的安全问答。

基于微服务架构搭建了系统底层，使得 LLM 推理、向量检索与应用服务可独立扩展。

采用混合检索和动态上下文注入技术，有效解决了通用大模型在建筑施工垂直领域中存在的知识盲区与幻觉问题。

通过系统测试表明，该设计逻辑清晰、功能完备，大幅降低了施工一线人员获取专业安全指导的门槛。综上所述，本设计为推动建筑施工行业的安全管理信息化与智能化提供了一套极具应用价值的技术方案。

参考文献：

[1] 中华人民共和国住房和城乡建设部. 建筑施工安全检查标准：JGJ 59-2011[S]. 北京：中国建筑工业出版社，2011.

[2] 赵磊，武彦清，周大伟等. 建筑施工安全领域大语言模型构建思路与方法研究[J]. 土木工程学报，2025，58(10): 144-152.

[3] 李圣飞，卢昱杰，陈晓莹等. 基于检索增强生成的建筑工程管理知识问答模型实现[J]. Science Technology & Engineering，2025, 25(25).

[4] Bai J, Bai S, Chu Y, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.

[5] Wang J, Yi X, Guo R, et al. Milvus: A purpose-built vector data management system[C]//Proceedings of the 2021 international conference on management of data. 2021: 2614-2627.

[6] Chen J, Xiao S, Zhang P, et al. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation[J]. arXiv preprint arXiv:2402.03216, 2024, 4(5).

[7] 中华人民共和国住房和城乡建设部. 建筑施工高处作业安全技术规范：JGJ 80-2016[S]. 北京：中国建筑工业出版社，2016.