

Research on the Construction and Reliability Assessment of RAG Dataset for Engineering Structural Design Specifications

Yida Li

Fujian Zhongzi Planning & Design Research Group Co., Ltd., Zhangzhou, Fujian, 363000, China

Abstract

Structural design codes organize knowledge through clause numbers, chapter hierarchies, tables, formulas, and explanatory notes. Mechanical segmentation may fragment retrieval evidence, weaken table semantics, and obscure answer sources. To improve traceability and reliability in engineering question answering, this paper builds a chapter-aware RAG dataset for structural design codes. The workflow integrates clause-level semantic chunking, metadata annotation, FAISS vector indexing, and hybrid retrieval, and evaluates reliability through evidence coverage, answer correctness, and hallucination rate. Results show that catalogues, clauses, tables, formulas, and explanatory notes can be connected within one retrieval chain, providing a practical data basis for code-oriented intelligent question answering, design checking, and knowledge-base updating.

Keywords

structural design codes; RAG dataset; FAISS index; reliability assessment; code-oriented question answering

工程结构设计规范 RAG 数据集构建与可靠性评估研究

李益达

福建中咨规划设计研究集团有限公司, 中国·福建 漳州 363000

摘 要

工程结构设计规范具有条文编号严密、层级稳定、表格公式密集和条文说明并存等特点, 普通文本分块易造成依据断裂、表格语义丢失和来源不清。为提高工程领域大语言模型问答的可追溯性与可靠性, 本文构建章节结构化RAG数据集, 提出条文语义单元切片、元数据标注、FAISS向量索引和混合召回方法, 并从证据覆盖率、答案正确性和幻觉率3个维度开展人工评估。结果表明, 该方法可将目录、条文、表格公式和条文说明纳入同一检索链路, 为工程规范智能问答、设计校核和知识库更新提供数据基础。

关键词

工程结构设计规范; RAG数据集; FAISS索引; 可靠性评估; 智能问答

1 引言

工程结构设计活动高度依赖规范条文。设计人员进行材料取值、构件验算、构造审查和耐久性判断时, 需同时查找总则、术语、计算条文、参数表和条文说明。传统关键词检索虽能定位页面, 却难以识别条文层级, 也难判断表格或公式是否属于必要依据。随着建筑工程标准数字化推进, 标准文本正由可阅读文档转向可检索、可理解、可计算数据^[1]。

检索增强生成成为规范问答提供了新路径。其思路不是让大语言模型凭训练记忆回答, 而是先从标准库召回依据, 再将依据与问题共同输入模型。工程结构设计规范由章、节、条、款、表、公式、注和条文说明构成, 切片过短会丢失完整约束, 切片过长又会稀释相似度。已有规范智能解译研究

表明, 条文拆解、语义标注和规则处理, 是标准知识问答与设计审查的基础^[2]。因此, 建立面向工程结构设计规范的RAG数据集和可靠性评估方法, 具有工程应用价值。

2 数据集构建思路

2.1 规范文本解析与层级切片

本文以工程结构设计规范文本为对象, 先开展文档清洗和结构识别。清洗过程保留条文编号、章节标题、表号、公式号、注释和页码, 不将其作为无意义符号删除。目录建立章、节、条索引, 正文确认条文边界, 表格和公式补充计算依据, 条文说明解释适用条件。对电子化文本, 系统先抽取版面内容, 再校验条文编号连续性, 避免页眉页脚、扫描换行和分栏文字破坏结构。

切片粒度以条文级 chunk 为主。每个 chunk 优先保留完整条文及必要表格、公式和说明。条文较短时可并入相邻注释, 条文较长时按款项或计算步骤拆为子 chunk, 但仍保

【作者简介】李益达(1997-), 男, 中国福建漳州人, 硕士, 助理工程师, 从事工程智能算法、通信研究。

留原条文号和上级标题。这样既便于向量模型匹配，也便于生成答案时追溯原始位置。对“应”“宜”“可”等规范用

语，系统单独标记约束强度，避免把推荐性条文误写为强制性结论。

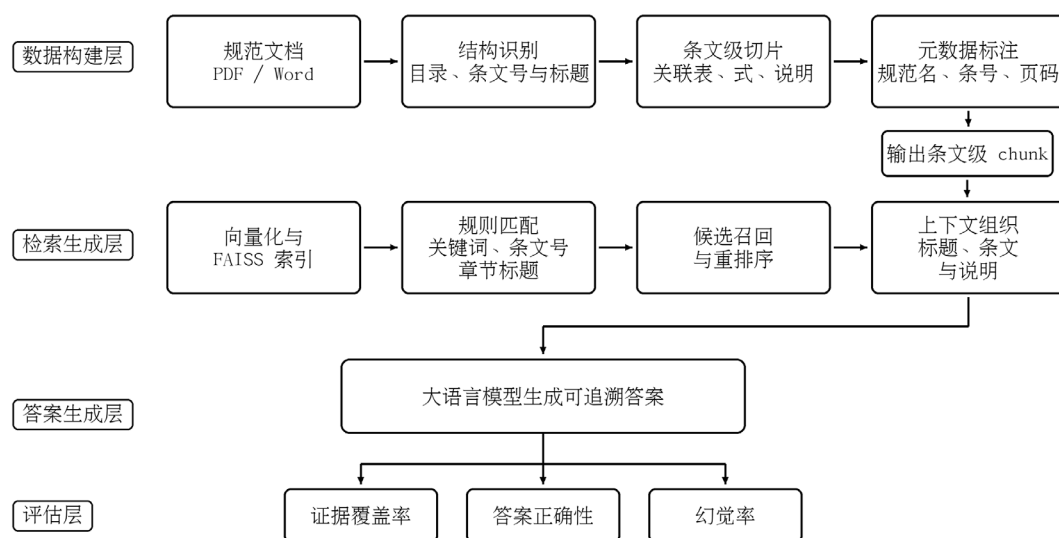


图1 工程结构设计规范章节层级切片与 RAG 构建流程

2.2 元数据设计与问答样本生成

结构化数据集采用文本内容与元数据并行存储。核心字段包括 chunk_id、规范名称、规范编号、版本、层级类型、章号、节号、条文号、页码、是否含表、是否含公式、关键词、正文、条文说明和关联依据。chunk_id 按照规范编号和条文号生成，保证同一条文在版本更新后仍可追踪。关键词由条文标题、专业名词和公式符号共同生成，避免只依靠模型嵌入造成专业符号检索失败。对于表格数据，除保存原文外，还拆分表名、列名、行名和单位字段，使系统能够回答“某强度等级对应设计值是多少”这类精确取值问题。

问答样本按照工程使用场景构建，包括取值题、条件判断题、公式适用题、章节定位题和多条文综合题。取值题检验表格召回能力，条件判断题检验约束理解能力，公式适用题检验变量含义与计算式对应关系，多条文综合题检验跨章节证据整合能力。每条样本均标注标准答案、必要依据、可选补充依据和不可接受答案类型。不可接受答案包括无依据扩展、误把条文说明当成强制要求、遗漏适用条件和编造参数。

数据入库前设置质量检查规则：完整性检查核对章、节、条、页码和正文是否缺失；一致性检查核对目录标题与正文标题是否对应；可追溯性检查核对问答样本能否回到原始条文位置；重复性检查合并同一条文在不同页面抽取中形成的重复片段。经检查后的数据再进入训练和评测流程，可减少将检索误差误判为生成误差的情况。

3 检索数据库与问答生成

3.1 向量化与 FAISS 索引

完成切片后，系统将 chunk 正文、标题和关键词拼接为向量化输入，并使用适合中文技术文本的嵌入模型生成向

量。FAISS 用于建立本地向量索引，支持大规模规范片段的快速近邻搜索。为减少向量表示对表格数字和条文号的不敏感问题，向量字段不单独承担全部召回任务，而是与关键词倒排、条文号精确匹配和章节标题匹配共同组成候选池。RAG 知识服务研究已经指出，外部知识库的规范化、向量化与可验证来源是提高专业问答质量的重要环节^[9]。

索引构建采用增量更新方式。新增规范或修订版本进入数据库时，系统先比较规范编号、版本号和条文号，再判断 chunk 内容是否发生变化。未变化条文保留原向量，变化条文重新嵌入并更新索引。对于废止条文，不直接删除，而是标记状态并限制召回范围，以免历史项目复核时失去依据。该策略能够同时支持现行规范问答和历史版本追溯，适合工程项目周期长、规范版本多的实际情况。

数据库运行时保存检索日志，内容包括用户问题、召回片段编号、相似度分值、重排序结果、最终上下文和生成答案。日志虽不替代人工评估，但可帮助定位错误发生在切片、索引、召回、重排序还是生成环节。对工程规范问答系统而言，过程记录有助于解释系统为何采用某条依据。

3.2 混合召回与上下文组织

检索阶段采用两级处理。第1级从 FAISS 相似度、关键词、条文号和章节标题4个入口召回候选片段；第2级通过重排序模型和规则校验筛选最终依据。问题含明确条文号时，条文号匹配优先；问题含构件、材料、作用效应或验算类型时，提高章节标题和关键词权重；问题表达较口语化时，由向量相似度承担主要召回任务。铁路行业知识库问答研究也说明，行业 RAG 系统需协同设计知识库组织、检索策略和问答流程^[4]。

答案生成阶段不把候选片段简单堆入提示词，而按“上

级标题、条文正文、表格或公式、条文说明”组织上下文。模型先判断问题类型，再列出依据并生成简明答案。对于规范未明确规定的问题，系统输出“未在当前知识库中检索到直接依据”，不作经验性推断。涉及数值的回答须同时给出数值、单位、来源条文和适用条件，以降低概念混淆风险。

4 可靠性评估方法

4.1 测试集设计与人工评分

可靠性评估以人工标注测试集为基础。测试集从真实

规范条文中抽取，覆盖材料性能、结构设计基本规定、承载能力验算、正常使用验算、构造要求和质量检验等主题。每个问题至少对应1条必要依据，复杂问题可对应多条依据。标注人员给出问题意图、标准答案、必要依据、可接受表述和典型错误。答案正确性采用人工5分制：5分表示答案完整、依据准确、无多余扩展；4分表示结论正确但略有缺项；3分表示方向正确但遗漏关键条件；2分表示依据不充分；1分表示结论错误或与规范冲突。

表1 RAG 可靠性评价指标与判定口径

评价指标	计算口径	人工判定要点	主要用途
证据覆盖率	被召回的标准依据数量 / 标准依据总数量	检查必要条文、表格、公式和条文说明是否进入最终上下文	评价检索充分性
答案正确性	人工5分制评分均值	检查结论、条件、数值、单位和规范依据是否一致	评价生成质量
幻觉率	出现幻觉的回答数量 / 总回答数量	识别编造条文、编造数值、错误引用和无依据扩展	评价安全边界

人工评分采用双人复核。若两个评分差值超过1分，则由第3名复核人员裁定。评分表除分数外，还记录错误来源，包括检索遗漏、召回片段无关、表格解析错误、上下文排序错误、模型生成扩展和单位表达错误。检索增强 Text-to-SQL 问答研究表明，垂直领域测试集应把查询意图、外部数据调用和执行结果统一评价，这一思路同样适用于工程规范问答可靠性验证^[5]。

4.2 指标体系与结果解释

本文设置证据覆盖率、答案正确性和幻觉率3个核心指标。证据覆盖率等于被召回的标准依据数量除以标准依据总数量，用于判断检索是否找到所需条文、表格或说明。答案正确性采用人工评分均值，用于判断回答是否符合规范原意。幻觉率等于出现幻觉的回答数量除以总回答数量，幻觉包括编造条文、编造数值、错误引用规范名称、将未检索到的信息写成确定结论，以及把经验判断冒充规范规定。

3个指标需要联动解释。证据覆盖率高但正确性低，说明检索找到了依据，但模型理解或组织上下文不足；证据覆盖率低但答案看似正确，说明模型可能依赖训练记忆或偶然猜测，不能视为可靠；幻觉率低但正确性也低，说明系统较谨慎但回答能力不足，需要优化切片和重排序。工程结构设计场景宁可提示依据不足，也不能给出无来源的确定结论。智能审图研究强调，工程规则审查不仅要求结果可用，还要求知来源、判断逻辑和算法过程可解释^[6]。

在样本构建测试中，系统形成168条条文级 chunk，其中含表格 chunk28 条、含公式 chunk38 条、含条文说明 chunk40 条，平均 chunk 长度约 209.6 字符。围绕材料参数取值、结构设计判断、耐久性要求、施工控制和质量检验5类场景设置15个测试问题，采用 Top3 召回结果人工核验，命中必要依据14个，证据覆盖率为93.33%。未完全命中的问题主要涉及耐久性技术措施与构造条文关联判断。结果表明，材料取值和条文定位问题召回较稳定，跨章节综合问题

仍需通过章节标题权重和重排序规则优化。

5 结论

工程结构设计规范 RAG 数据集建设不能简单等同于固定长度切片。规范文本的价值存在于层级结构、条文编号、表格公式和条文说明之间的关系中。本文提出的章节结构感知切片方法，能够将规范文本整理为可追踪的条文级 chunk，并通过元数据保存规范名称、版本、条文号、页码、关键词和内容类型。基于 FAISS 的向量索引与关键词、条文号、章节标题混合召回相结合，可以提高专业问题召回稳定性。

在可靠性评估方面，证据覆盖率衡量检索是否充分，答案正确性衡量生成是否符合规范原意，幻觉率衡量模型是否生成无依据内容。3项指标共同构成工程规范问答的基础评价框架。后续研究可扩展多规范联合检索、表格结构化解析、公式变量约束校验和版本差异比对，并在设计咨询、施工审查和教育培训场景中持续测试，使 RAG 系统由“能回答”逐步走向“有依据、可复核、可追责”。

参考文献

- [1] 林佳瑞,陈柯吟,潘鹏. 建筑工程标准数字化与智能化:现状与未来[J]. 东南大学学报(自然科学版),2025,55(1):16-29.
- [2] 林佳瑞,陈柯吟,郑哲,周育丞,陆新征. 建筑工程标准规范智能解译关键技术及应用[J]. 工程力学,2025,42(2):1-14.
- [3] 王亮. 检索增强生成(RAG)驱动的知识服务:原理、范式及评估[J]. 科技与出版,2025(4):37-46.
- [4] 骆铭,林铭,刘思行,等. 基于检索增强的铁路行业知识库问答系统设计与实现[J]. 铁路计算机应用,2025,34(12):38-43.
- [5] 王震宇,朱学芳,张君冬,杨睿,刘崧印. 基于检索增强Text-to-SQL生成的书目搜索对话问答系统研究[J]. 数据分析与知识发现,2025,9(11):165-174.
- [6] 林佳瑞,周育丞,郑哲,陆新征. 自动审图及智能审图研究与应用综述[J]. 工程力学,2023,40(7):25-38.