Computer Science and Application 计算机科学与应用, 2026, 16(6), 177-190
Published Online June 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.166218

Hans汉斯

面向多版本法规问答的版本感知知识图谱增强检索方法

陈晓宇[1,2*]，花　蕾[1,2]

[1]上海市数字城市规划研究中心，上海
[2]上海市量子城市空间智能创新重点实验室，上海

收稿日期：2026年5月12日；录用日期：2026年6月16日；发布日期：2026年6月26日

摘　要

法律法规在多次修订后，同一条款编号在不同版本中内容存在差异，且条款之间交叉引用频繁。现有基于BM25和稠密向量的检索增强生成(RAG)系统在两类问题上表现不佳：其一为时效性问题(即给定日期下应当适用条款的哪个版本)，其二为版本差异问题(即条款在不同版本之间发生了哪些变化)。本文分析上述不佳的根本原因，提出一套有针对性的解决方案。首先，构建跨版本法律知识图谱，包含条款引用(references)、版本修订(amends)和未变延续(inherits)三类边。其次，将问题划分为精确、版本差异、推理和时效四类，每类配置固定的跳数和边类型策略。当问题中出现具体条款编号时，跳过稠密种子检索，按article_no_int字段直接执行SQL查询以获取同条款的多版本内容，从而规避稠密检索因抽象关键词导致的语义偏移。本文在《中华人民共和国城乡规划法》四个版本(1989、2007、2015、2019，共256条)上构建67题基准开展评估。实验结果显示，本方法Recall@5达到0.963, MRR达到0.934，引用F1达到0.897，均高于BM25(0.761/0.692/0.718)和稠密检索(0.918/0.872/0.867)。在版本差异类别上，两类基线方法的Recall@5仅为0.167和0.500，本方法达到1.000，引用F1由0.133和0.500提升至0.933。在时效性类别上，本方法与稠密检索持平于Recall@5 = 1.000和引用F1 = 0.972。为验证泛化性，本文在《测绘法》(32题，修订密集)和《土地权属争议调查处理办法》(15题，修订稀疏)两部法律上补充实验。前者四项指标均达到1.000，后者退化为稠密检索水平但不劣化。双侧Wilcoxon配对检验显示，本方法相对BM25在六项指标上均$p < 0.01$(效应量$r \in [0.659, 0.883]$)。本文实验环境完全可复现：嵌入采用本地Qwen3-Embedding-4B，生成采用Gemma-4-31B。

关键词

知识图谱，检索增强生成，法律问答，多版本检索，本地大模型推理

Version-Aware Knowledge-Graph-Enhanced Retrieval for Multi-Version Legal Question Answering

*通讯作者。

文章引用: 陈晓宇, 花蕾. 面向多版本法规问答的版本感知知识图谱增强检索方法[J]. 计算机科学与应用, 2026, 16(6): 177-190. DOI: 10.12677/csa.2026.166218

Xiaoyu Chen[1,2], Lei Hua[1,2]

[1]Shanghai Digital City Planning & Research Center, Shanghai
[2]Shanghai Key Laboratory of Quantum City Spatial Intelligence Innovation, Shanghai

Received: May 12, 2026; accepted: June 16, 2026; published: June 26, 2026

Abstract

Legal statutes are amended repeatedly. Across versions the same article number can carry different text and cross-references. BM25 and dense RAG both fail on two practical question types in this setting: temporal applicability (i.e., which version of an article should be applied on a given date) and version differences (i.e., how an article has changed across different versions). This paper analyzes the root causes of these failures and proposes a targeted solution. First, we build a cross-version legal knowledge graph containing references, amends, and inherits edges. Second, we classify questions into four types (precise, version difference, reasoning, and temporal applicability) and configure fixed hop and edge type policies for each. Finally, when an explicit article number appears in the question, we bypass dense seeding retrieval and directly execute a SQL query on the article_no_int field to obtain multi-version contents of the same article, thereby circumventing the semantic shift in dense retrieval caused by abstract keywords. On the Chinese Urban-Rural Planning Law (4 versions, 256 articles, 67-question benchmark), our method reaches Recall@5 = 0.963, MRR = 0.934, Hit@1 = 0.896, Citation F1 = 0.897, against 0.761/0.692/0.612/0.718 for BM25 and 0.918/0.872/0.821/0.867 for dense retrieval. On version-diff queries, the Recall@5 of the two baselines is only 0.167 and 0.500, while our method reaches 1.000, and the citation F1 is improved from 0.133 and 0.500 to 0.933. On the temporal applicability category, our method ties with dense retrieval at Recall@5 = 1.000 and citation F1 = 0.972. To verify generalization, we conduct supplementary experiments on two additional laws: the Surveying and Mapping Law (32 questions, densely amended) and the Measures for the Investigation and Handling of Land Ownership Disputes (15 questions, sparsely amended). The former achieves 1.000 on all four metrics, while the latter regresses to the dense retrieval level without degradation. Two-sided Wilcoxon paired tests show that our method outperforms BM25 on all six metrics with $p < 0.01$ (effect size r in [0.659, 0.883]). The experimental setup is fully reproducible, using local Qwen3-Embedding-4B (MLX backend) for retrieval and Gemma-4-31B (Google AI Studio free tier) for generation.

Keywords

Knowledge Graph, Retrieval-Augmented Generation, Legal Question Answering, Multi-Version Retrieval, Local LLM Inference

1. 引言

1.1. 问题背景与动机示例

法律法规生效后通常经多轮修订。以《中华人民共和国城乡规划法》[1]为例，前身为 1989 年颁布的

[1]上海市规划和自然资源局. 中华人民共和国城乡规划法[EB/OL]. https://ghzyj.sh.gov.cn/zcwj/cxgh/20241205/3a637454c6984d6b8610e4d00b2b24b9.html, 2019-04-24.

《城市规划法》，经 2007 年重新制定并更名，又于 2015 年和 2019 年两次修正，累计形成四个有效版本。跨版本演化中条款编号变化显著(1989 版 46 条，2007 版扩至 70 条)，条款之间还存在大量交叉引用，形成一张跨版本、跨条款的复杂关系网络。

检索增强生成(RAG) [1]是当前法律问答主流路线，但 BM25 与稠密向量两类方法在多版本场景下表现不佳。考察动机示例：“2019 年修正前后，《城乡规划法》第三十八条关于建设单位领取建设用地规划许可证的顺序有何调整？”正确答案依赖 2007 与 2019 版第 38 条。稠密检索返回第 42、59、70 条，全部错失。由嵌入向量分析可知，“变化”“修正前后”等抽象词占主要权重，而“第三十八条”仅 5 字，在均值池化中被显著稀释；BM25 不佳模式类似。在 4 道版本差异类问题上，两类基线 Recall@5 均为 0.00。

类似问题反复出现，主要分两类：1) 时效性(time_applicable)，即给定日期适用条款的哪个版本；2) 版本差异(version_diff)，即条款在两版间如何变化。两者均非单版本检索所能解决。CALRK-Bench [2]在韩国法律上独立识别出相同模式，表明该问题具有跨语系普遍性。

1.2. 研究问题

本文围绕以下三个研究问题展开。RQ1：在多版本法规语料下，BM25 和稠密检索是否在时效性与版本差异两类问题上系统性不佳？RQ2：跨版本知识图谱与问题分类驱动的检索策略能否同时弥补上述两类不佳？RQ3：图扩展机制是否会对其他类别(精确、推理)的检索性能造成负面影响？

1.3. 主要贡献

本文核心贡献三方面。1) 跨版本法律知识图谱：从 PDF 抽取条款，将条款编号整数化为 article_no_int，依据文本差异在同编号条款的版本对之间建立引用(references)、修订(amends)、延续(inherits)三类边。2) 问题分类驱动的图谱增强检索：规则分类器将问题划为四类并抽取 mentionArticleNo、atTime 特征；显式含条号时跳过稠密种子直接执行 SQL，规避抽象关键词在稠密空间对正确条款的语义偏移。3) 可复现基准：将 BM25、稠密、长上下文、KG-RAG 置于同一题库与生成模型下横评。主基准《城乡规划法》67 题，加《测绘法》[2]32 题与《土地权属争议调查处理办法》[3]15 题做泛化验证，共 114 题。实验环境完全可复现：嵌入用本地 Qwen3-Embedding-4B，生成用 Gemma-4-31B。

1.4. 论文组织

本文余下章节安排如下：第 2 章综述法律问答、检索增强生成与时序检索三个相关方向；第 3 章给出方法的形式化定义与三个核心组件；第 4 章描述实验设置(语料、题库、对比方法、评估协议)；第 5 章报告主对比、分类别对比、消融实验与统计显著性结果；第 6 章讨论稠密检索失败根因、残留失败案例、跨法规泛化性与方法局限；第 7 章总结全文并给出未来工作。

2. 相关工作

2.1. 法律问答

法律问答已是 NLP 中相对独立的评测分支。英文基准包括 LegalBench [3]、LexGLUE [4]与 COLIEE 系列[5]；中文有 CAIL [6] (判决预测)、CJRC [7] (司法阅读理解)与 LawBench [8] (按记忆/理解/应用 20 类任务)。面向中文法律的检索增强大模型有 Lawyer LLaMA [9]、ChatLaw [10]、DISC-LawLLM [11]。上述

[2]上海市规划和自然资源局. 中华人民共和国测绘法[EB/OL]. https://ghzyj.sh.gov.cn/zcwj/chgl/20241206/20ccf288b2a140498c5ebab69d1306c5.html, 2017-04-27.
[3]土地权属争议调查处理办法[EB/OL]. https://www.gov.cn/gongbao/content/2003/content_62339.htm, 2003-01-03.

工作多假设单版本条文，未显式建模版本演化。Louis 等[12]研究可解释的长篇法律问答，LegalLens [13]识别违法行为，二者仍以静态快照为基础。CALRK-Bench [2]在韩国法律上独立提出考察 temporal validity 与 norm scope 的评测，验证了版本漂移问题在不同法系中的普遍性。

2.2. 检索增强生成

RAG 范式由 Lewis 等[1]奠基。稠密检索从 DPR [14]经 Contriever [15]发展至 BGE-M3 [16]；本文采用 Qwen3 嵌入模型[17]。查询侧改进包括 HyDE [18]、Query2doc [19]与 RAG-Fusion [20]。图增强 RAG 是与本工作最近的方向：GraphRAG [21]抽实体图做社区摘要；KAPING [22]与 Sen 等[23]将三元组注入提示词；Think-on-Graph [24]在图上做受控束搜索。法律领域内，Domain-Partitioned Hybrid RAG [25]拆分印度法律为三模块并用大模型路由；SEARCHFiRESAFETY [26]要求消防法规引用链在层级结构内闭合。与上述工作相比，本文路由规则为确定性而非大模型驱动，无额外推理开销。重排方法(ColBERT [27]、monoT5 [28]、RankGPT [29])留作未来工作。

2.3. 时序与多版本文档检索

时序信息检索在网页搜索领域积累深厚(Alonso 等[30]，Kanhabua 等[31])。法律历时性语料需要版本感知索引：Akoma Ntoso [32]强制条款级版本与时间元数据；LegalRuleML [33]形式化法律规范；EUR-Lex 通过 CELEX [34]追踪跨文档版本。法律引用网络方面，Fowler 与 Jeon [35]分析判例权威度，Sadeghian 等[36]做语义边标注。时序知识图谱有 Leblay 与 Chekol [37]推导事实有效区间、García-Durán 等[38]学习时序补全嵌入。Kanapala 等[39]综述将版本追踪列为开放问题。据本文 2026 年 4 月文献调研，尚无工作同时具备：(i) 显式跨版本 amends 与 inherits 边；(ii) 问题类型驱动检索；(iii) 条号直达快路径。最相关工作[2] [25] [26]均未将同一法规多版本演化纳入检索对象。

3. 方法

3.1. 问题定义

本文将法律查询形式化为三元组 $q = (\text{text}, \text{law\_scope}, \text{atTime})$，其中 text 表示自然语言问题，law_scope 为目标法律集合，atTime 为适用时间点。整个任务可划分为两个阶段：首先从多版本法规语料 A 中召回候选集 R 作为检索结果；随后由 R 生成答案 y 与引用集 $C \subseteq R$。标准答案由必需引用集 C*与参考回答 y*组成。本文采用 Recall@K、MRR、Hit@1 与引用 Precision/Recall/F1 作为评价指标。

3.2. 多版本语料构建

PDF 到结构化条款的流水线：opendataloader-pdf 或 MinerU 抽取文本，chunker 按章节与条款正则切分，条款编号整数化为 article_no_int，大模型辅助抽取条款引用并标注 scope (internal/external/unknown)，写入 copus.articles 与 copus.article_edges。跨版本对齐依赖 article_no_int (由中文序数前缀解析的整数 ID)，共享相同 article_no_int 的两条款即为 amends 或 inherits 候选。PDF 中有时混入其他法规，本文将条款编号回退现象(如第七十条后出现第一条)视为异物注入信号，截断尾部并隔离至 corpus/_bycatch/待人工复核。

3.3. 跨版本知识图谱

节点定义为 article (article_id, law_id, version_id, article_no_int, text, chapter)。边三类：references (A 引用 B，版本内或跨版本)；amends (同条号文本不同，视为修订)；inherits (同条号文本相同，视为延续)。

构建过程：对任意版本对(v_i, v_j)i < j 中相同 article_no_int 的两条款计算文本编辑距离，低于阈值建 inherits，否则建 amends。边属性 metadata.gap 记录版本索引距离，支持仅直接前驱或全部祖先等过滤。时效性建模上，每个版本携带(promulgate_at, effective_at, repeal_at)时间戳(bigint, 秒)；给定 atTime，选择满足 effective_at ≤ atTime < repeal_at 的版本，repeal_at = 0 视为永久有效。

3.4. 问题分类器

本文采用规则分类器，依据词汇线索将问题划为四类(详见表 1)：精确类(precise)、版本差异类(version_diff)、推理类(reasoning)、时效类(time_applicable)。对每个问题，分类器抽取 mentionArticleNo (整数)、atTime、mentionVersion 与 confidence 四项特征。其中 mentionArticleNo 对所有类别均进行抽取，以保证条号直达路径在版本差异类下能够正常触发。

Table 1. Question types and retrieval policies
表 1. 问题类别与检索策略

类别	触发词	种子	边类型	跳数
precise	默认；含具体条号	稠密	无	0
version_diff	修正/修改/前后/变化	条号直达	amends, inherits	2
reasoning	违反/责任/处罚	稠密	References (+amends)	2
time_applicable	日期 + 适用提示	scope 按 atTime 过滤 + 条号直达	无	0

3.5. 知识图谱增强检索

对于一个已分类的问题(q, type, features)，检索流程包含以下步骤。首先，进行种子检索得到候选集 S0；其次，按类别策略沿对应类型的边进行 k 跳扩展得到 Sk；随后，种子节点保留原始得分，扩展节点的得分按 0.5/(hop+1)衰减；最后，按得分排序并截取 Top-K。默认种子检索器为稠密检索(Qwen3-Embedding-4B, 2560 维向量，余弦距离，基于 pgvector 顺序扫描)。在单方法对比中，稠密检索性能优于 BM25 (详见第 5.1 节)。

3.5.1. 条号直达快路径

条号直达快路径是本方法的核心环节。当 features.mentionArticleNo 非空，且 type ∈ {precise, version_diff, time_applicable}时，系统跳过稠密种子，直接执行 SQL: SELECT articles WHERE article_no_int = $1 AND (law_id, version_id) IN scope ORDER BY version_id。对版本差异类，单次查询即覆盖全部版本的标准答案；对精确类与时效类，scope 层已按最新版或 atTime 过滤，通常返回单条(hops = 0)。针对版本差异类问题，额外扩展一跳 amends 边作为兜底，计算开销较低。将时效类纳入快路径白名单是相对 v3 初版的修订：初版扩一跳 references 致 Recall@5 达 1.00 但生成端出现版本漂移，检索选中的生效版本被扩展邻近条款挤占，引用精确率反而降低。

3.5.2. 为何奏效

稠密检索在“第三十八条的变化”此类问题上表现不佳，根因在于嵌入向量权重分布：“变化”“修正前后”等抽象词在嵌入空间占主要权重，向量偏向修订密集的邻近条款(如第 42、70 条)，而“第三十八条”仅 5 个有效字符在均值池化中被显著稀释。条号直达快路径规避了这一过程：对条款编号这一确定性结构信号，稠密检索视为软匹配，SQL 视为硬约束—这是性能差距的根本来源。

3.6. 答案生成与引用回填

生成阶段以检索候选为上下文，JSON 模式约束输出{answer, citations, confidence}。当 citations 为空但 answer 中提到条款编号时，系统启用引用回填：扫描“第 X 条”模式按 article_no_int 匹配检索集补齐。该机制避免长上下文方法(候选多，大模型可能省略显式引用)在引用 F1 上受不公平扣分。

4. 实验设置

4.1. 数据集

本文主语料为《中华人民共和国城乡规划法》四个版本(版本信息见表 2)，共计 256 条款。原始 PDF 文件来自政府官网。构建得到的知识图谱共包含 256 个条款节点。其中 references 边共 22 条，包含已解析至具体条款的内部边 21 条与一条未解析的外部引用边；amends 边共 4 条，主要修订集中在第 24 条与第 38 条；inherits 边共 206 条，覆盖全部版本对而非仅相邻版本。

Table 2. Law version metadata
表 2. 法律版本信息

版本	颁布日期	生效日期	条数	类型
1989-12-26 (《城市规划法》)	1989-12-26	1990-04-01	46	原法
2007-10-28	2007-10-28	2008-01-01	70	新法(更名重构)
2015-04-24	2015-04-24	2015-04-24	70	修正
2019-04-23	2019-04-23	2019-04-23	70	修正

题库共 67 道，涵盖四类(分布见表 3)。其中 26 道由具有法律背景的作者人工撰写，借助大模型辅助润色；其余 41 道由 Deepseek V3.2 批量合成(精确类按条款随机采样，版本差异类沿 amends 边枚举，推理类沿 references 链遍历，时效类按版本生命周期采样)，按大模型自评置信度 ≥ 0.5 过滤后由作者人工抽检。

Table 3. Question benchmark distribution
表 3. 问题基准分布

类别	数量	示例
precise	35	2019 年修正版《城乡规划法》第十九条规定了什么？
reasoning	14	违反城市规划进行建设的法律责任是什么？
time_applicable	12	2017 年 3 月办理的建设工程规划许可证适用哪个版本的第 38 条？
version_diff	6	2015 年修正前后，第二十四条关于规划编制单位的条件要求发生了哪些变化？

4.2. 对比方法与实现

四路对比方法(表 4)共用同一生成模型(Gemma-4-31B-it, Google AI Studio)，仅替换检索模块，以确保对比数据反映检索本身差异。运行时 Bun 1.3 + PostgreSQL 16 + pgvector 0.8。本地推理在 Apple M1 Max 上，嵌入用 mlx-community/Qwen3-Embedding-4B-mxfp8，单条约 230ms。随机种子固定 42。每个(题，方法)组合执行完整端到端流程：检索 top-K = 5→JSON 模式生成→引用回填→写 eval_runs 表；配置快照写 eval_run_configs 表作为复现锚点。

Table 4. Four retrieval baselines
表 4. 四路对比方法

方法	说明
BM25	PostgreSQL pg_trgm 词频相似度
稠密(Qwen3-4B)	Qwen3-Embedding-4B (mxfp8, 2560 维)，余弦距离，pgvector
长上下文	将 scope 内全部条款按字符预算(≤180k 字)塞入 LLM 上下文
KG-RAG (本文)	分类器 + 条号直达 + 图扩展(种子为稠密)

5. 结果

5.1. 主对比

表 5 汇总四路方法的检索与引用性能。长上下文方法按定义返回 scope 内全部条款(平均每题约 43 条)，故 Recall@5 = 1.000 是结构性必然而非排序质量。本方法 Recall@5 达 0.963，相对稠密(0.918)提 4.5%，相对 BM25 (0.761)提 20.2%。MRR 差距更显著(0.934 对稠密 0.872)，表明正确引用排序更靠前；Hit@1 达 0.896 (67 题中 60 题首位命中)。引用 F1 达 0.897，相对稠密提 3.0%，相对 BM25 提 17.9%。长上下文虽 Recall = 1.000 但 MRR 仅 0.059、Hit@1 仅 0.015，约为本方法的 1/16 与 1/60；引用 F1(0.884)反而低于本方法，代价是输入 tokens 约 11 倍、端到端延迟约 1.5 倍。

Table 5. Method-level retrieval and citation performance
表 5. 方法级检索与引用性能

方法	n	Recall@5	MRR	Hit@1	Cite P	Cite R	Cite F1	延迟 p50
BM25	67	0.761	0.692	0.612	0.715	0.746	0.718	20098 ms
稠密(Qwen3-4B)	67	0.918	0.872	0.821	0.853	0.918	0.867	21915 ms
长上下文	66	1.000	0.059	0.015	0.839	0.992	0.884	41292 ms
KG-RAG (本文)	67	0.963	0.934	0.896	0.877	0.963	0.897	27367 ms

5.2. 分类别对比

表 6 与表 7 分别按问题类别给出 Recall@5 与引用 F1。性能差异集中在两个类别。精确与推理类上稠密与本方法的 Recall 与引用 F1 基本持平，原因是这两类不依赖图结构，分类器识别后采零跳或低跳扩展，系统退化为稠密检索。时效类上本方法与稠密持平于 Recall = 1.000 与引用 F1 = 0.972；BM25 仅 0.639，表明词频对时间约束几乎不敏感。差距最显著的是版本差异：BM25 与稠密均明显不佳(引用 F1 分别为 0.133 与 0.500)，本方法稳定 0.933。6 道版本差异题均由条号直达快路径单次查询覆盖，一跳 amends 扩展仅作备选。

Table 6. Recall@5 by question category
表 6. 按问题类别的 Recall@5

类别	题数	BM25	稠密	长上下文	KG-RAG
precise	35	0.914	0.971	1.000	0.971
reasoning	14	0.714	0.893	1.000	0.893
time_applicable	12	0.667	1.000	1.000	1.000
version_diff	6	0.167	0.500	1.000	1.000

Table 7. Citation F1 by question category
表 7. 按问题类别的引用 F1

类别	BM25	稠密	长上下文	KG-RAG
precise	0.871	0.895	0.883	0.878
reasoning	0.652	0.862	0.834	0.862
time_applicable	0.639	0.972	0.944	0.972
version_diff	0.133	0.500	0.883	0.933

5.3. 消融实验

消融结果(表 8)显示，条号直达快路径是版本差异与时效两类的主要性能贡献者：禁用后版本差异引用 F1 由 0.933 降至 0.467，时效类 Hit@1 由 0.833 降至 0.500。当前实现中分类器与快路径耦合：禁用分类器后 mentionArticleNo 不再填充，快路径触发条件隐式失效，故 w/o classifier 与 w/o fast-path 数据接近。amends 与 inherits 边在当前题库几乎未产生贡献(6 道版本差异题均被快路径覆盖)，但当用户用描述性语言询问版本演化时，这些边将作为快路径未命中时的备选机制。

Table 8. Ablation—Citation F1 by category
表 8. 消融实验——按类别的引用 F1

类别	完整 KG-RAG	w/o fast-path	w/o classifier	w/o cross-ver edge
precise	0.878	0.905	0.888	0.892
reasoning	0.862	0.876	0.890	0.891
time_applicable	0.972	0.972	0.972	0.972
version_diff	0.933	0.467 ↓	0.500 ↓	0.933

5.4. 统计显著性

考虑到 n = 67 的样本规模可能导致单点数值的误导，本文为每个指标计算了 bootstrap 95% 置信区间(重采样次数 n = 1000)。同时，本文对 KG-RAG 相对 BM25 与稠密检索的每对同 qid 指标执行双侧 Wilcoxon 符号秩配对检验(含 tied-ranks 修正与连续性校正)。

Table 9. Paired Wilcoxon signed-rank test
表 9. KG-RAG 对基线的 Wilcoxon 配对检验

对比	指标	n_eff	p (双侧)	效应量 r	显著性
vs. BM25	Recall@5	18	0.0003	0.860	***
vs. BM25	MRR	29	0.0001	0.732	***
vs. BM25	Hit@1	27	0.0003	0.701	***
vs. BM25	Cite P	21	0.0022	0.668	**
vs. BM25	Cite R	18	0.0002	0.883	***
vs. BM25	Cite F1	23	0.0016	0.659	**
vs. 稠密	MRR	8	0.0371	0.737	*
vs. 稠密	Hit@1	7	0.0726	0.679	n.s.

配对检验结果汇总于表 9。本方法相对 BM25 在 6 项指标上均显著(4 项 $p < 0.001$, 2 项 $p < 0.01$)，效应量 $r \in [0.659, 0.883]$属中至大效应。对稠密仅 MRR 显著($p = 0.0371$, $r = 0.737$)，其他指标 n_effective 较小，原因是超 85%题目两方法得分相同或差异极小，符合设计预期：精确与推理类持平，差异集中在版本差异与时效类少量题(详见 5.2)。

5.5. 子模块独立评估

5.5.1. 问题分类器性能

为独立评估分类器可靠性，本文在 GS-Q (n = 67，城乡规划法主题库)上分别运行规则分类器与 DeepSeek-v4-pro few-shot 分类器，结果见表 10。

Table 10. Question classifier—by-class P/R/F1
表 10. 问题分类器 by-class P/R/F1(GS-Q, n = 67)

类别	支持数	规则 P	规则 R	规则 F1	LLM P	LLM R	LLM F1
precise	35	0.839	0.743	0.788	1.000	0.314	0.478
version_diff	6	0.208	0.833	0.333	1.000	1.000	1.000
reasoning	14	0.000	0.000	0.000	0.368	1.000	0.538
time_applicable	12	0.875	0.583	0.700	1.000	1.000	1.000
macro	67	0.481	0.540	0.455	0.842	0.829	0.754
accuracy	67	—	—	0.567	—	—	0.642

规则分类器 macro F1 = 0.455，主要失分在 reasoning 类(F1 = 0.000，14 题全被误判为 version_diff)。该误判在实践中无害：推理类问题通常不含明确条号，mentionArticleNo = null，快路径不触发，系统退化到稠密检索(与 reasoning 设计路径相同)；消融实验(表 8)证实关闭分类器后该类 Recall@5 不变。version_diff 类规则分类器 R = 0.833 (6 题中 5 题命中)，足以驱动快路径覆盖全部正确召回。DeepSeek-v4-pro 分类器 macro F1 = 0.754，完美识别 version_diff 与 time_applicable (F1 = 1.000)，但 precise R = 0.314(24/35 题被过分归为 reasoning)，综合延迟(约 2s/题)与可复现性，规则分类器仍为部署默认；跨领域迁移时建议改用 DistilBERT 类微调模型(见 7.1)。

5.5.2. 跨版本边构建敏感性

为验证阈值选择的稳健性，扫描归一化 Levenshtein 阈值 $\tau \in \{0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$，每档重建图谱后仅跑 KG-RAG 一路(主基线数据保持冻结)，结果见表 11。

Table 11. Edge construction sensitivity
表 11. 归一化 Levenshtein 阈值 τ 敏感性(KG-RAG 端到端，n = 67)

τ	模式	amends 数	inherits 数	Cite F1 (整体)	version_diff Cite F1
—	exact (默认)	4	206	0.897	0.933
0.00~0.05	levenshtein	4	206	0.907	0.917
0.10~0.15	levenshtein	2	208	0.904	0.883
0.20~0.30	levenshtein	0	210	0.897	0.850

$\tau \in [0.05, 0.15]$时 2 条 amends 被折叠为 inherits，整体 Cite F1 浮动±1 pp，version_diff Cite F1 从.933 降至.883，鲁棒性良好。$\tau \geq 0.20$ 时全部 4 条 amends 被折叠，version_diff Cite F1 降至.850(↓8pp)，判别力明显下降。结果表明默认精确比对在本语料上是稳健选择；OCR 噪声明显的扫描件语料中，$\tau = 0.10$ 可作 fallback (仅损失 1pp)。

5.5.3. 图谱构建质量

对城乡规划法 3 个修正版本(2007/2015/2019)共 210 个条款构建黄金集 GS-G，由作者人工核对切分边界、条号抽取与跨版本关系判定(含 split/merge/renumber 复杂情形)；评估脚本将系统输出与 GS-G 对照，结果见表 12。

Table 12. Graph construction quality
表 12. 图谱构建质量三项准确率(GS-G, n = 210)

维度	总计	正确	准确率	主要失败模式
切分边界	210	210	1.000	—(正则确定性算法；PDF 页眉页脚噪声在抽取层拦截)
条号抽取	210	210	1.000	—(本语料无“一百二十”等长形中文数字)
跨版本关系	140	140	1.000	split/merge/renumber 0 条

三项准确率均达 1.000。跨版本关系判定覆盖 2007→2015 和 2015→2019 两对共 140 条边；1989→2007 (《城市规划法》→《城乡规划法》)因法规名称更换与条数扩张(46→70 条)，视为新法颁布，不纳入同条号对齐管线。ad-hoc 复杂修订检测脚本(Jaccard ≥ 0.7)在两修正版本对上均输出 0，印证普通修正阶段同条号对齐已完全覆盖；大修或新法颁布阶段 ad-hoc 工具可大幅降低人工搜索空间。

5.6. 效率

BM25、稠密与本方法三条路径端到端延迟处于同一量级(p50 在 20 至 27 秒之间)，大部分时间消耗在 Gemma-4-31B 推理(10~20 秒)，检索仅占 1~2 秒。长上下文 p50 为 41.3 秒，明显高出，主因是处理约 10 倍 prompt tokens。本方法相对长上下文节省约 11 倍输入 tokens 与 1.5 倍延迟，大规模问答场景下可扩展性更好。

6. 讨论

6.1. 稠密检索为何在版本差异上表现不佳

以查询“2015 年修正前后，第二十四条关于规划编制单位的条件要求发生了哪些变化？”为例展开分析。该查询的嵌入向量由“修正”“前后”“变化”“要求”等抽象词主导。上述词汇在语料中广泛出现，与第 24、42、59、63 与 70 条均存在相当程度的语义邻近性。而“第二十四条”这一精确标识在查询中仅占 2 个汉字(约占查询总长的 2%)，在均值池化得到的句向量中被显著稀释。条号直达快路径有效利用了一个确定性的结构信号——整数 24。稠密检索将其视为软匹配信号，而 SQL 查询将其视为硬约束。这一处理方式与自然语言到 SQL 任务中的查询路由策略一致：当查询显式包含精确实体标识时，跳过语义相似检索是更合理的选择。

6.2. 失败案例

本方法残留两类失败。一是无显式条号的精确类问题：少数精确类采用概念性表述(如“规划编制的

原则是什么？”)，无法触发快路径，完全依赖稠密种子质量，本方法在该子集上不优于稠密。二是小生成模型版本漂移，已修复：初版用本地 Gemma-4-E4B (4.5B 激活，MLX 8-bit)时，本方法在时效类 Recall = 1.000 但引用 F1 仅 0.750。两步修复：时效类策略改为 hops = 0 并纳入快路径白名单(F1 升至 0.833)；替换为 Gemma-4-31B (F1 进一步至 0.972)。由此该失败的根因是小模型指令跟随不足，而非方法结构问题。

本文从主 run(run_id 8af64cb8)选取三案例说明基线在多版本场景下的失败模式。案例 A 查询为“2019 年修正前后，《城乡规划法》第三十八条建设用地规划许可证办理顺序的调整？”，标准引用 art.38 (2007/2019)。BM25 检索到 art.42、art.39 多版本，大模型回应“所给条款不足以回答”，Recall/F1 = 0/0；稠密检索到 art.37、art.42，同样回绝，Recall/F1 = 0/0。本方法快路径单次 SQL 命中第 38 条三个版本，大模型据此给出 2007 与 2019 差异，Recall/F1 = 1.0/1.0。案例 B 是 art.24 同模式独立验证，BM25 与稠密 Recall = 0，本方法 Recall = 1.0。案例 C 长上下文在同 art.24 问题：Recall@5 = 1.000，但 MRR = 0.042 (art.24 按自然编号位列候选第 24)、Hit@1 = 0；F1 = 1.000 因大模型在 210 候选中正确选中目标，代价是输入 tokens 约为本方法 11 倍。

6.3. 泛化性

三部法规的横向对比(表 13)显示，《测绘法》(123 条, amends 53 条，修订密集)上本方法四项指标均达 1.000；版本差异类(7 题)再次观察到 BM25(0.143)与稠密(0.571)显著不佳，与主基准(0.167/0.500/1.000)模式一致。《土地权属争议调查处理办法》(72 条，amends 仅 4 条，修订稀疏)上本方法与稠密基本持平：修订少，快路径触发场景有限，图扩展边际收益近零。这一现象印证而非反驳本文主张：性能增益依赖多版本 + 修订量足以形成 amends 边的条件；条款稳定的法规上本方法退化为稠密水平但不劣化。

Table 13. Method-level metrics across three laws
表 13. 三部法规上的方法级指标

语料	n	方法	Recall@5	MRR	Hit@1	Cite F1
城乡规划法	67	BM25	0.761	0.692	0.612	0.718
		稠密	0.918	0.872	0.821	0.867
		KG-RAG	0.963	0.934	0.896	0.897
测绘法	32	BM25	0.797	0.760	0.719	0.781
		稠密	0.906	0.898	0.875	0.885
		KG-RAG	1.000	1.000	1.000	0.990
土地权属争议办法	15	BM25	0.929	0.752	0.643	0.905
		稠密	0.933	0.883	0.867	0.933
		KG-RAG	0.933	0.883	0.867	0.911

6.4. 局限

本文存在以下局限。1) 样本规模累计 114 题偏少，单部法律题量有提升空间；三部法律均属中文大陆法系，英文、欧盟与判例法泛化未验证。2) 规则分类器 macro F1 = 0.455，主要失分在 reasoning 类(F1 = 0.000，详见 5.5 表 11)；该误判在实践中无害(5.5.1 有分析)，跨领域迁移时建议改用 DistilBERT 类轻量

微调模型。3)生成模型规模影响引用 F1：Gemma-4-E4B 时效类有约 14 个百分点版本漂移损失，31B 下消失；检索指标与生成模型无关，不影响核心贡献。

6.5. 应用前景

本文方法对法律实务工具具有直接价值。基于 KG 的版本追踪机制可作为底层组件，支撑回溯性法规合规审计场景：给定历史合同或行政行为的发生日期，系统能精确定位当时适用的条款版本，避免传统检索因版本漂移导致的法律意见偏差。此外，跨版本 amends/inherits 边可服务于立法演进研究：通过遍历 amends 边可量化某条款的修订频率与改动幅度，辅助法学研究者识别活跃修订的法律领域，也可作为法律 AI 训练语料的版本标注信号。

7. 结论

本文提出一套面向多版本法规问答的版本感知知识图谱增强检索方法，含三个核心组件：基于条款引用与跨版本文本差异直接构建的跨版本法律知识图谱、将查询映射至专属策略的规则分类器，以及在查询显式含条款编号时跳过稠密种子的条号直达快路径。

在《城乡规划法》四版本、256 条款、67 道题主基准上，本方法 Recall@5 = 0.963、MRR = 0.934，均高于 BM25 (0.761/0.692)与稠密(0.918/0.872)；提升集中在版本差异类(本方法 1.000，稠密 0.500, BM25 0.167)；时效类与稠密持平于 Recall 1.000、引用 F1 0.972；简单类持平不回退。泛化实验上《测绘法》(amends = 53)本方法四项指标均达 1.000，《土地权属争议调查处理办法》(amends = 4)退化至稠密水平但不劣化，表明性能增益与 amends 边数量正相关。

实验环境完全可复现，嵌入用本地 Qwen3-Embedding-4B，生成用 Gemma-4-31B。代码、数据与评测协议同步公开。

8. 未来工作

未来工作展望：1) 扩展更多法律语料(多辖区成文法与行政法)验证跨领域泛化；2) 采用轻量学习式分类器替代规则分类器，在更大规模标注集上微调提升鲁棒性；3) 将修订差异摘要注入生成提示以缓解版本漂移；4) 探索条号直达快路径在技术标准、学术引用网络等含精确标识符场景的适用性。

参考文献

[1] Lewis, P., Perez, E., Piktus, A., *et al*. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*, Vancouver, 6-12 December 2020, 9459-9474.

[2] Jung, J., Yoon, T. and Cho, H. (2026) CALRK-Bench: Evaluating Context-Aware Legal Reasoning in Korean Law. arXiv:2603.26332.

[3] Guha, N., Nyarko, J., Ho, D.E., Ré, C., Chilton, A., Narayana, A., *et al*. (2023) Legalbench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *SSRN Electronic Journal*, 143 p. https://doi.org/10.2139/ssrn.4583531

[4] Chalkidis, I., Jana, A., Hartung, D., Bommarito, M.J., Androutsopoulos, I., Katz, D.M., *et al*. (2021) LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. *SSRN Electronic Journal*, 17 p. https://doi.org/10.2139/ssrn.3936759

[5] Goebel, R., Kano, Y., Kim, M., Rabelo, J., Satoh, K. and Yoshioka, M. (2023) Summary of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2023. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, Braga, 19-23 June 2023, 472-480. https://doi.org/10.1145/3594536.3595176

[6] Xiao, C., Zhong, H., Guo, Z., *et al*. (2018) CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. arXiv:1807.02478.

[7] Duan, X., Wang, B., Wang, Z., Ma, W., Cui, Y., Wu, D., *et al*. (2019) CJRC: A Reliable Human-Annotated Benchmark

Dataset for Chinese Judicial Reading Comprehension. In: Sun, M., Huang, X., Ji, H., Liu, Z. and Liu, Y., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 439-451. https://doi.org/10.1007/978-3-030-32381-3_36

[8] Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Huang, A., *et al*. (2024) Lawbench: Benchmarking Legal Knowledge of Large Language Models. *Proceedings of the* 2024 *Conference on Empirical Methods in Natural Language Processing*, Miami, 12-16 November 2024, 7933-7962. https://doi.org/10.18653/v1/2024.emnlp-main.452

[9] Huang, Q., Tao, M., An, Z., *et al*. (2023) Lawyer LLaMA Technical Report. arXiv:2305.15062.

[10] Cui, J., Li, Z., Yan, Y., *et al*. (2023) ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. arXiv:2306.16092.

[11] Yue, S., Chen, W., Wang, S., *et al*. (2023) DISC-LawLLM: Fine-Tuning Large Language Models for Intelligent Legal Services. arXiv:2309.11325.

[12] Louis, A., Van Dijck, G. and Spanakis, G. (2024) Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 22266-22275. https://doi.org/10.1609/aaai.v38i20.30232

[13] Bernsohn, D., Semo, G., Vazana, Y., Hayat, G., Hagag, B., Niklaus, J., *et al*. (2024) LegalLens: Leveraging LLMs for Legal Violation Identification in Unstructured Text. *Proceedings of the* 18*th Conference of the European Chapter of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), St. Julian's, 17-22 March 2024, 2129-2145. https://doi.org/10.18653/v1/2024.eacl-long.130

[14] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., *et al*. (2020) Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Online, 16-20 November 2020, 6769-6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[15] Izacard, G., Caron, M., Hosseini, L., *et al*. (2022) Unsupervised Dense Information Retrieval with Contrastive Learning. arXiv:2112.09118.

[16] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D. and Liu, Z. (2024) M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings through Self-Knowledge Distillation. *Findings of the Association for Computational Linguistics ACL* 2024, Bangkok, 11-16 August 2024, 2318-2335. https://doi.org/10.18653/v1/2024.findings-acl.137

[17] Zhang, Y., Li, M., Long, D., *et al*. (2025) Qwen3 Embedding: Advancing Text Embedding and Reranking through Foundation Models. arXiv:2506.05176.

[18] Gao, L., Ma, X., Lin, J. and Callan, J. (2023) Precise Zero-Shot Dense Retrieval without Relevance Labels. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Toronto, 9-14 July 2023, 1762-1777. https://doi.org/10.18653/v1/2023.acl-long.99

[19] Wang, L., Yang, N. and Wei, F. (2023) Query2doc: Query Expansion with Large Language Models. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 9414-9423. https://doi.org/10.18653/v1/2023.emnlp-main.585

[20] Raudaschl, A.H. (2024) RAG-Fusion: A New Take on Retrieval-Augmented Generation. arXiv:2402.03367.

[21] Edge, D., Trinh, H., Cheng, N., *et al*. (2024) From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130.

[22] Baek, J., Aji, A.F. and Saffari, A. (2023) Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. *Proceedings of the* 1*st Workshop on Natural Language Reasoning and Structured Explanations* (*NLRSE*), Toronto, 13 June 2023, 78-106. https://doi.org/10.18653/v1/2023.nlrse-1.7

[23] Sen, P., Mavadia, S. and Saffari, A. (2023) Knowledge Graph-Augmented Language Models for Complex Question Answering. *Proceedings of the* 1*st Workshop on Natural Language Reasoning and Structured Explanations* (*NLRSE*), Toronto, 13 June 2023, 1-8. https://doi.org/10.18653/v1/2023.nlrse-1.1

[24] Sun, J., Xu, C., Tang, L., *et al*. (2024) Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. arXiv:2307.07697.

[25] Goel, R., Kumar, S.P., Agrawal, A., Poddar, D., Narang, P. and Kumar, D. (2025) Domain-Partitioned Hybrid RAG for Legal Reasoning: Toward Modular and Explainable Legal AI for India. arXiv:2602.23371.

[26] Chae, K., Yeom, J., Park, J., Bae, S., Jang, I., Jin, H., et al. (2026) Beyond Case Law: Evaluating Structure-Aware Retrieval and Safety in Statute-Centric Legal QA. arXiv:2604.06173.

[27] Khattab, O. and Zaharia, M. (2020) ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *Proceedings of the* 43*rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event, 25-30 July 2020, 39-48. https://doi.org/10.1145/3397271.3401075

[28] Nogueira, R., Jiang, Z., Pradeep, R. and Lin, J. (2020) Document Ranking with a Pretrained Sequence-to-Sequence Model. *Findings of the Association for Computational Linguistics*: *EMNLP* 2020, Online, 16-20 November 2020, 708-718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[29] Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., *et al*. (2023) Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 14918-14937. https://doi.org/10.18653/v1/2023.emnlp-main.923

[30] Alonso, O., Strotgen, J., Baeza-Yates, R. and Gertz, M. (2011) Temporal Information Retrieval: Challenges and Opportunities. TWAW. https://ceur-ws.org/Vol-707/TWAW2011-paper1.pdf

[31] Kanhabua, N., Blanco, R. and NJrvęg, K. (2015) Temporal Information Retrieval. *Foundations and Trends® in Information Retrieval*, **9**, 91-208. https://doi.org/10.1561/1500000043

[32] Palmirani, M. and Vitali, F. (2011) Akoma-Ntoso for Legal Documents. In: Sartor, G., Palmirani, M., Francesconi, E. and Biasiotti, M., Eds., *Legislative XML for the Semantic Web*, Springer, 75-100. https://doi.org/10.1007/978-94-007-1887-6_6

[33] Athan, T., Governatori, G., Palmirani, M., Paschke, A. and Wyner, A. (2015) LegalRuleML: Design Principles and Foundations. In: Faber, W. and Paschke, A. Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 151-188. https://doi.org/10.1007/978-3-319-21768-0_6

[34] Publications Office of the European Union (2026) EUR-Lex: CELEX Numbering System. https://eur-lex.europa.eu/

[35] Fowler, J.H. and Jeon, S. (2008) The Authority of Supreme Court Precedent. *Social Networks*, **30**, 16-30. https://doi.org/10.1016/j.socnet.2007.05.001

[36] Sadeghian, A., Sundaram, L., Wang, D.Z., Hamilton, W.F., Branting, K. and Pfeifer, C. (2018) Automatic Semantic Edge Labeling over Legal Citation Graphs. *Artificial Intelligence and Law*, **26**, 127-144. https://doi.org/10.1007/s10506-018-9217-1

[37] Leblay, J. and Chekol, M.W. (2018) Deriving Validity Time in Knowledge Graph. *Companion Proceedings of the The Web Conference* 2018, Lyon, 31 23-27 April 2018, 1771-1776. https://doi.org/10.1145/3184558.3191639

[38] García-Durán, A., Dumančić, S. and Niepert, M. (2018) Learning Sequence Encoders for Temporal Knowledge Graph Completion. *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, Brussels, 31 October-4 November 2018, 4816-4821. https://doi.org/10.18653/v1/d18-1516

[39] Kanapala, A., Pal, S. and Pamula, R. (2017) Text Summarization from Legal Documents: A Survey. *Artificial Intelligence Review*, **51**, 371-402. https://doi.org/10.1007/s10462-017-9566-2