文章编号:1007-5321(2025)05-0025-07 **DOI**:10.13190/j.jbupt.2025-062

基于 Linked-RAG 和大语言模型的电信客户投诉判责方法

李征仁[1], 黄佳宝[1], 陶昀昕[2], 吕廷杰[1], 陈 飞[1]
(1. 北京邮电大学 经济管理学院, 北京 100876; 2. 北京体育大学 体育工程学院, 北京 100084)

摘要: 随着电信业务规模的指数级增长,客户投诉判责成为通信运营商合规管理的关键环节。传统方法在处理复杂型投诉和非结构化历史案例时面临语义解耦困难、知识复用效率低等挑战。该研究面向现有业务流程基于大语言模型技术,提出一种链接式检索增强生成框架,通过层次化语义解耦与历史知识动态检索机制,实现投诉判责的精准化和高效化。该框架提出 2 级投诉点拆分机制:1 级采用混合式语义解析剥离复合诉求,2 级通过大模型提示技术从 1 级投诉列表中提取核心投诉点,同时动态处理历史文档适配 2 级投诉匹配,完成数据库的结构化改造。最后,基于某省级运营商 423 份实际投诉数据的实验表明,对比几种传统检索增强生成(RAG)模型,该方法在检索准确率方面提升 18.87%,在检索召回率方面提升 14.13%。该研究为强监管背景下的智能判责提供了高效的可复用的技术路径,具有重要的实践意义。

关 键 词: 检索增强; 投诉判责; 语义解耦; 知识融合; 大语言模型
中图分类号: TP393.0 **文献标志码**: A

Liability Judgment Method of Telecom Customer Complaint Based on Linked-RAG and Large Language Models

LI Zhengren[1], HUANG Jiabao[1], TAO Yunxin[2], LU Tingjie[1], CHEN Fei[1]
(1. School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. School of Sports Engineering, Beijing Sport University, Beijing 100084, China)

Abstract: With the exponential growth of telecommunications business scale, customer complaint responsibility determination has become a key link in the compliance management of communication operators. Traditional methods face challenges such as semantic decoupling difficulties and low efficiency of knowledge reuse when dealing with complex complaints and unstructured historical cases. This study, based on the existing business process and large language model technology, proposes a linked retrieval-enhanced generation framework. Through hierarchical semantic decoupling and dynamic retrieval of historical knowledge, it achieves the precision and efficiency of complaint responsibility determination. The framework proposes a two-level complaint point splitting mechanism: The first level uses a hybrid semantic parsing to strip off compound demands, and the second level extracts core complaint points from the first-level complaint list through large model prompt technology, while dynamically processing historical documents to adapt to the second-level complaint matching, completing the structural transformation of the database. Finally, experiments based on 423 actual complaint data from a provincial

收稿日期: 2025-06-21
基金项目: 国家重点研发计划雄安新区科技创新专项(2023XAGG009304)
作者简介: 李征仁(1987—), 男, 副教授, 硕士生导师, 邮箱:lizhengren@bupt.edu.cn。

operator show that compared with several traditional retrieval augmented generation (RAG) models, this method improves the retrieval accuracy by 18.87% and the retrieval recall rate by 14.13%. This research provides an efficient and reusable technical path for intelligent responsibility determination under a strong regulatory background and has important practical significance.

Key words: retrieval augmented generation; complaint judgment; semantic decoupling; knowledge integration; large language model

随着电信业务爆发式增长,客户投诉责任判定成为合规管理关键瓶颈。运营商需按要求对工信部投诉判责并生成合规报告,依赖多模态证据链验证,但现行机制存在2重问题:1)用户投诉内容模糊,情绪化,人工处理易误判;2)非结构化历史案例难复用,同类案件重复投入成本高,据工信部2024年数据,34.3%服务争议申诉因证据链不全引发2次复议,暴露传统模式语义解耦与知识复用缺陷。

基于检索增强生成的垂直大模型技术虽能结合领域知识库与大模型语义理解能力,提升判责准确性与效率,但传统检索增强生成处理复杂投诉时,用户情绪化表述与多诉求交织易产生检索噪声,静态知识融合难适配动态举证需求,需通过层次化语义解耦与动态知识链接突破瓶颈。

本研究提出链接式检索增强生成(Linked-RAG,linked retrieval augmented generation)框架,以2级投诉点拆分机制重构判责范式:1级通过动态检索引导与大模型少样本提示Few-shot拆分繁杂投诉,2级提炼标准化核心投诉内容并构建与判责逻辑的动态映射。某省级运营商423份真实数据验证显示,该方法检索准确率96.22%,召回率较传统RAG提升14.13%。笔者后续将梳理相关研究、阐述框架架构、验证模型有效性、解析模块必要性并探讨技术普适性与挑战。

1 文献综述

1.1 投诉判责领域相关研究

在投诉判责领域中,人工智能技术应用成效显著。自然语言处理(NLP, natural language processing)技术广泛用于投诉内容自动分类与责任判定[1],例如李芳等提出基于NLP的电信投诉智能定责框架,通过分析客户与客服通话文本识别投诉根本原因,显著提升判责效率[2]。

多模态技术进一步拓展判责数据维度。Singh等构建基于联邦元学习的多模态多任务框架,整合文本、图像和音频数据优化投诉识别与情感分析,其含3 928条标注数据的公开数据集,为多模态投诉研究提供基准[3],能在复杂投诉场景减少信息缺失导致的判责偏差。

1.2 RAG理论研究现状

RAG技术结合检索与生成,可提升自然语言处理任务准确性与相关性[4]。相较于传统大语言模型(LLM,large language model),RAG通过动态知识注入缓解模型幻觉,在需实时更新或领域特异性知识场景表现突出,同时保留LLM语言能力并赋予其动态知识获取与领域适配能力[5]。

架构创新上,混合检索结合密集检索(DPR[6], dense passage retrieval)与稀疏检索(最佳匹配(BM, best matching)25算法[7])优势,通过动态权重优化结果多样性;跨模态检索与多源知识融合技术成为热点,可提升复杂问题解答能力[8]。排序优化方面,自主检索增强生成(AutoRAG, autonomous retrieval augmented generation)通过贪婪算法自动选择最优模块组合,重排序混合搜索结合关键词与语义检索,通过多阶段过滤和查询扩展策略减少噪声干扰[9]。

1.3 大模型理论研究现状

大模型基于深度学习技术(尤其Transformer架构与预训练框架),通过词向量转化与自注意力机制实现自然语言理解与生成[10],经预训练学习语言知识,微调优化特定任务,还可通过人类反馈强化学习提升性能[11]。

生成优化聚焦知识整合与参数调控:自适应注意力机制平衡内部参数与外部知识;自反思检索增强生成(Self-RAG, self-reflective retrieval augmented generation)通过自省标记机制动态调节生成过程[12];注意力机制融合允许生成器动态加权检索片段[13];低秩适应(LoRA,low-rank adaptation)等参数高效微调技术,仅需更新0.1%参数即可适配新领域,节约大量计算资源[14]。

1.4 文献评述

现有研究虽推进客户投诉自动化判责,但存在

2 点关键问题：1）难以解析含多诉求与情绪化表述的文本，现有 RAG 框架无法细粒度拆分诉求（如资费争议与服务中断交叉场景）；2）非结构化历史案例与制度条款关联依赖人工，同类案件重复查证成本高。

为此，笔者提出 Linked-RAG 框架：在 RAG 基础上，通过 2 级投诉点拆分实现层次化语义解析，结合动态权重混合检索与覆盖性验证算法构建投诉点–制度条款动态映射，最终在判责准确率（96.22%）、检索召回率（提升 14.13%）上实现突破，为复杂投诉场景提供可审计的智能判责方案。

2　方法论

2.1　整体架构设计

笔者提出的 Linked-RAG 框架通过 3 级协同机制实现投诉判责的语义解耦与知识增强，系统架构由预处理层、解析层与决策层构成，如图 1 所示，依次完成多粒度知识结构化、层次化语义解析及判责推理，核心创新在于将复杂投诉的语义拆分与判责逻辑动态关联，通过 2 级投诉拆分策略突破传统 RAG 的语义边界约束与静态知识依赖。

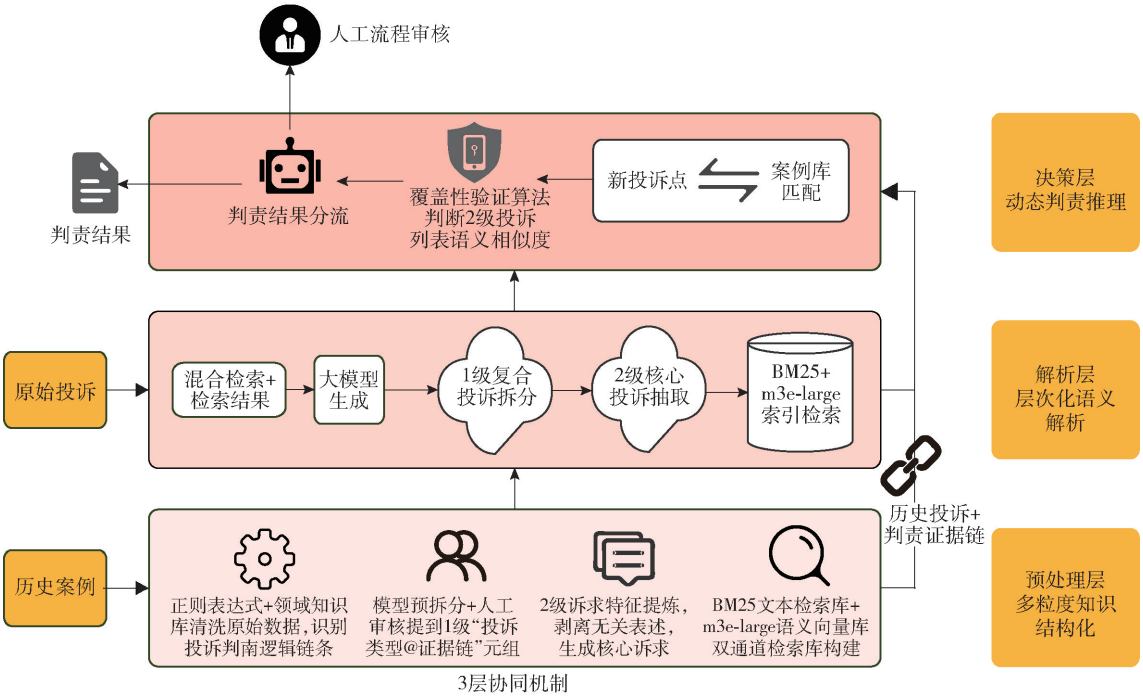

图 1　3 层协同架构

2.2　链接式 RAG 算法原理

预处理层通过结构化多粒度知识构建双通道检索库。给定历史案例库 D，首先应用正则表达式与领域知识库进行数据清洗：

$$(C_{\text{raw}}, L_{\text{raw}}) = R(d), d \in D \tag{1}$$

R 是对原始文档进行过滤，c_i 是提取出的原始投诉文本，l_i 是对应该条投诉的判定逻辑。通过预拆分模型和人工校验生成 1 级投诉–证据链元组：

$$C_1 = \{(c_{i1}, l_{i1}) \mid H(\text{Qwen}M_{\text{split}}(c_i, l_i)), c_i \in C_{\text{raw}}, l_i \in L_{\text{raw}}\} \tag{2}$$

其中：c_{i1} 是拆分后的 1 级原子投诉点，l_{i1} 是专属于该投诉点的判定逻辑片段，$c_i \in C_{\text{raw}}$ 为从原始投诉集合中取出单个投诉文本，$\text{Qwen}M_{\text{split}}(c_i, l_i)$ 为使用 Qwen 大语言模型处理 2 个输入。

H 人工校验环节的设置源于电信投诉领域的特殊性：一方面，其可确保初始知识库中“投诉类型@证据链”元组符合业务合规逻辑，规避 Qwen 模型因训练数据偏差产生“弃真错误”带来的合规风险与企业损失；另一方面，该环节仅作用于初始知识库构建阶段的 1 级投诉点校准，知识库稳定后新投诉判责全流程可脱离人工干预，实现自动化运行。利用大模型提炼 2 级投诉点：

$$C_2 = (c_{i2} \mid \text{Qwen}M_{\text{refine}}(c_{i1})) \tag{3}$$

$\text{Qwen}M_{\text{refine}}$ 是使用模型输出 2 级投诉点，且每个 1 级投诉点只产生 1 个最优 2 级投诉点。

解析层通过层次化语义解析与动态知识检索实

现复杂投诉的精准判责。给定投诉文本 C,1 级检索生成独立诉求集合,2 级检索对每个诉求执行双通道检索。稀疏检索采用 jieba 分词,BM25 算法计算关键词匹配得分:

$$s_{\mathrm{BM25}}(r_i,c_j)=\sum_{t\in r_i} I(t)\frac{f(t,c_j)(k_1+1)}{f(t,c_j)+k_1\left(1-b+b\dfrac{|c_j|}{a}\right)} \tag{4}$$

其中:r_i 是新用户投诉,c_j 是知识库中的用户投诉,t 是查询中的 1 个词项,$I(t)$ 是词项 t 的逆文档频率,$f(t,c_j)$ 是词项 t 在文档 c_j 中的词频,$|c_j|$ 是文档 D 的长度,a 是语料库中所有文档的平均长度,k_1 和 b 为调节参数。稠密检索通过 m3e-large 模型生成语义向量 $\boldsymbol{v}_{r_i}$ 和 $\boldsymbol{v}_{c_j}$,计算余弦相似度为

$$s_{\mathrm{vec}}(r_i,c_j)=\frac{\boldsymbol{v}_{r_i}\boldsymbol{v}_{c_j}}{\|\boldsymbol{v}_{r_i}\|\,\|\boldsymbol{v}_{c_j}\|} \tag{5}$$

最终检索得分通过动态权重融合:

$$s(r_i,c_j)=\alpha s_{\mathrm{BM25}}+(1-\alpha)s_{\mathrm{vec}} \tag{6}$$

动态权重 $\alpha=0.5$,经业务场景验证最优。

决策层通过大模型语义对齐实现判责逻辑映射。给定新用户投诉,通过 2 级投诉列表 $\{c_{i2}\}_{i=1}^{n}$ 与检索到的 Top-5 历史案例的 2 级投诉集合进行匹配。

$$M_{ijk}=[M_{\mathrm{match}}([c'_{i2}][c^*_{jk}])],\forall i\in[1,n],\ c^*_{jk}\in C^*_{i2} \tag{7}$$

其中:i 为新 2 级投诉的列表位数,j 和 k 为检索出的第 j 个 2 级投诉的第 k 个投诉点,M_{match} 为 Qwen-72B 模型,M_{match} 判断 2 投诉点是否为描述同一类问题(返回 True/False)。判责逻辑生成函数定义为

$$R(s_i^*)=\begin{cases}\mathscr{L}(s^*_{ijk}), & \exists i,j,k\\ \text{s.t. } M_{ijk}=1 & \\ \mathrm{HumanAudit}(s'_i), & \text{其他}\end{cases} \tag{8}$$

最终输出判责结果集 $R=\bigcup_{i=1}^{n}R(s_i^*)$,未覆盖投诉点 $\{s'_i\mid\forall j,M_{ijk}=0\}$,同步触发人工审核流程 HumanAudit,$\mathscr{L}(\cdot)$ 为历史案例的判责逻辑提取函数。数据处理流程如图 2 所示。

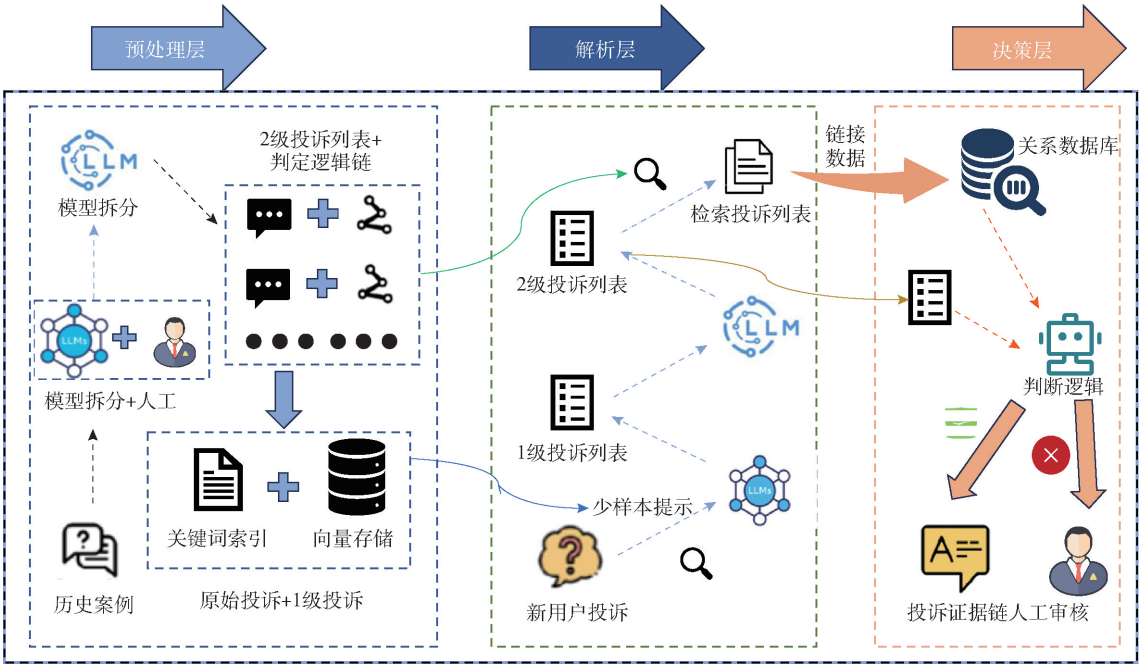

图 2 数据流转图

伪代码实现如下:

算法 链接式 RAG 判责流程

Input: 新用户投诉

1:检索结果 = 混合检索(新用户投诉)#BM25 + 向量检索

2: 1 级投诉列表 = Few_shot(检索结果)

3: 2 级投诉列表 = 提取特征(1 级投诉列表)

4:检索案例(2 级投诉列表)

5:匹配矩阵 = 生成匹配矩阵(2 级投诉列表)

6:loop

7: if 匹配存在(投诉点,匹配矩阵)

8: 判责结果.append(读取历史判责(投诉点))

9: else

10: 判责结果.append(人工判责(投诉点))

11：end loop
Output：判责结果

3　实验设计

3.1　数据集描述

本研究的电信投诉判责数据集，源于某省级运营商 2024 年 5～7 月脱敏的 423 份真实申诉文档，涵盖资费争议（34.7%）、服务中断（28.1%）、合约纠纷（22.5%）、隐私泄露（14.7%）4 类争议，含非结构化文本、业务截图、计费表格、通话录音等多模态信息，完整还原投诉证据链。

为平衡案例稀缺性与训练需求，数据集采用“随机划分 + 定向生成”策略：423 份真实数据中，370 份为基础训练集，53 份为初始测试集，对测试集案例通过替换场景、调整表述保持核心判责逻辑一致，确保投诉点分布均衡。知识建模采用半人工监督机制：10 名通信专家制定标注规范，Qwen-72B 预标注投诉点后经双盲校验修正，m3e-large 生成 768 维语义向量，结合 Milvus 构建多模态检索空间，最终形成含 528 条标准判责逻辑链的知识库，每条逻辑链对应具体 2 级投诉点。

3.2　基线模型介绍

本研究选取 4 类基线模型进行横向对比，系统验证链接式 RAG 框架的技术优势。

1）传统混合检索增强生成（HybridRAG）：作为工业界主流方案，由 BM25（jieba 分词构建词频–逆文档频率（TF-IDF，term frequency-inverse document frequency）倒排索引）与稠密向量检索（m3e-large 生成 768 维向量，Milvus 近邻搜索）按 1∶1 固定权重融合；生成模块采用 Qwen-72B，输入为检索结果与原始投诉文本拼接，复现工业界标准 RAG 判责流程。

2）模型优化类方案：包含 3 种细分配置，均围绕“模型适配性与语义拆分能力”验证。一是 LoRA 微调方案，含 BigScience 大型开放科学开放获取多语言模型（BLOOM，bigscience large open-science open-access multilingual）-1b4-zh（低秩适配器注入参数，探索小模型轻量化潜力）与 Qwen-7B（结合电信行业规则设计模板，验证中模型领域知识注入性能）2 种基座；二是基于 Transformer 的双向编码器表示（BERT，bidirectional encoder representations from transformers）分类引导方案，基于 bert-base-chinese 构建双阶段框架，先训练分类头输出投诉点数量，再指导大模型拆分，设早停机制防过拟合，验证层次化拆分必要性；三是纯大模型方案，含 Qwen-72B（测试零样本推理能力）与 Qwen-7B（作为轻量化对照）2 种规模，均采用动态上下文窗口与自洽校验，统一判责归属与依据判定模板，同时为 LoRA 微调提供同基座基准。

3）AutoRAG 方案：通过贪婪算法自动选择检索器、重排器和生成器参数，实现检索流程端到端自动化配置；与 Linked-RAG 共用检索库与输入数据，验证通用模块组合优化策略的效能。

4）InstructRAG 方案：利用大模型生成“问题–推理步骤–答案”结构化链条，基于链条与检索结果的语义对齐度过滤噪声；采用默认指令模板，与 Linked-RAG 共用检索库，验证指令引导式检索在复杂场景的适应性。

3.3　核心实验介绍

本研究的链接式 RAG 框架采用 2 级检索引导的层次化拆分机制：1 级投诉拆分阶段，通过混合检索（BM25 + m3e-large）召回知识库 Top-3 相似历史案例，以其拆解逻辑为动态提示输入 Qwen-72B，生成结构化 1 级投诉点集合；2 级投诉解析阶段，利用 Qwen-72B 对 1 级投诉点细粒度拆分，提取标准化 2 级投诉内容，并基于混合检索匹配历史案例库判责逻辑。覆盖性验证算法通过大模型判定 2 级投诉与历史案例的语义相似度，匹配失败触发制度条款泛化检索，未覆盖案例自动转入人工复核，形成闭环判责链路。

实验将数据集分为训练集（423 份）与测试集（53 份），采用分层采样保证类别分布一致。为验证框架有效性，设计 3 组消融实验：Abl-1 移除 2 级投诉拆分模块，对原始投诉文本单级检索；Abl-2 禁用混合检索中 BERT 重排序机制，仅保留 BM25 与向量检索的线性加权结果；Abl-3 关闭覆盖性匹配功能，以检索内容直接指导模型逻辑链输出。

3.4　评估指标介绍

本研究采用 4 维度量化指标体系验证链接式 RAG 框架的判责效能。

1）Top-5 检索精度（Precision@5）定义为在 Top-5 检索结果中正确匹配投诉点的比例 P，计算式为

$$P = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(d_j \in R_i)}{5N} \tag{9}$$

其中：d_j 表示第 j 位检索文档，R_i 为第 i 个投诉的真实相关文档集，$I(\cdot)$ 为指示函数。

2)投诉点召回率(Recall@5)衡量知识库中所有相关文档的覆盖程度 R,计算式为

$$R=\frac{\sum_{i=1}^{N}\left|R_i\cap D_i^{\text{Top-5}}\right|}{\sum_{i=1}^{N}\left|R_i\right|} \tag{10}$$

$D_i^{\text{Top-5}}$ 表示第 i 次检索的 Top-5 文档集合。

3)Top 覆盖值通过动态阈值法计算覆盖全部投诉点所需最小检索量 $C_{\min}$,计算式为

$$C_{\min}=\frac{1}{N}\sum_{i=1}^{N}\min k \text{ s.t. } U_{j=1}^{k}d_j\supseteq R_i \tag{11}$$

反映系统在最少检索量下实现全诉求覆盖的能力。

4)命中率量化检索结果对新投诉的决策支持价值 H,计算式为

$$H=\frac{\sum_{i=1}^{N}I(\exists d_j\in D_i^{\text{Top-5}},\text{Sim}(d_j,c_i))}{N} \tag{12}$$

其中 Sim(·)调用模型判断语义是否相似。测试集采用53份真实投诉样本进行端到端验证,通过分层交叉验证消除数据分布偏差,确保指标计算的鲁棒性。

4 实验结果

4.1 实验结果介绍与分析

表1展示综合性能对比。

表1 5种方案在电信投诉判责任务中的核心指标表现

评估方案	Precision@5	Recall@5	Top 覆盖	HitRate
混合检索	0.773 5	0.830 1	—	—
RAG_lora_1.4B	0.652 8	0.760 0	2.566 0	0.380 5
RAG_lora_7B	0.903 8	0.931 5	1.635 0	0.465 9
RAG_纯大模型_7B	0.881 5	0.902 3	1.886 8	0.512 0
RAG_BERT_72B	0.886 7	0.928 6	1.528 3	0.537 7
RAG_纯大模型_72B	0.924 5	0.961 9	1.886 8	0.478 3
AutoRAG	0.943 3	0.957 7	1.547 2	0.417 3
InstructRAG	0.927 5	0.929 6	1.622 6	0.410 1
Linked-RAG	0.962 2	0.971 4	1.377 4	0.521 4

Linked-RAG 在电信投诉判责任务中表现出全面优越性,如表1所示,其 Precision@5 达96.22%,较次优纯大模型方案提升3.77%,较传统混合检索方案提升18.87%,验证了动态权重混合检索对判责依据捕获的完整性;Recall@5 以97.14%领先所有基线模型,复杂投诉场景中2级投诉点召回率较混合检索方案提升14.13%;效率上,Top 覆盖值仅需1.38个检索结果即可覆盖全部投诉点,较纯大模型拆分方案(1.886 8)降低0.509 4,凸显检索引导机制对知识定位效率的提升,仅 HitRate(52.14%)略低于 BERT 方案(53.77%)。

对 Linked-RAG、AutoRAG、InstructRAG 和 RAG_lora_7B4 个模型进行 Kappa 系数分析,以评估模型之间的分类一致性。Kappa 系数矩阵如表2所示。

表2 4种 RAG 的 Kappa 一致性分析结果

模型	Linked-RAG	AutoRAG	InstructRAG	RAG_lora_7B
Linked-RAG	1.000 0	0.898 3	0.765 3	0.656 6
AutoRAG	0.898 3	1.000 0	0.682 1	0.572 6
InstructRAG	0.765 3	0.682 1	1.000 0	0.465 6
RAG_lora_7B	0.656 6	0.572 6	0.465 6	1.000 0

从表2可以看出,Linked-RAG 和 AutoRAG 模型之间的 Kappa 系数最高,达到0.898 3,表明这2个模型在分类任务上具有较高的一致性。相比之下,InstructRAG 和 RAG_lora_7B 模型之间的 Kappa 系数最低,为0.465 6,这表明它们在分类任务上的一致性相对较低。总体而言,Kappa 系数矩阵反映了不同模型之间的分类一致性,为模型选择和优化提供了有价值的参考。

4.2 消融实验结果分析

为验证链接式 RAG 框架各模块的贡献度,本研究设计3组消融实验(Abl-1 ~ Abl-3),依次移除关键组件后观察性能变化,性能影响分析如表3所示。

表3 展示消融实验对核心指标的影响

消融实验	Recall@5	前3正确率	判责逻辑数量
完整模型	0.962 2	0.547 2	3.207 5
Abl-1(无2级检索)	0.830 1	—	—
Abl-2(无 BERT)	—	0.452 8	—
Abl-3(静态匹配)	—	—	6.603 8

消融实验结果表明,链接式 RAG 框架的关键模块对系统性能具有决定性影响,如表3所示,移除2级投诉检索模块(Abl-1)后,Recall@5 从0.962 2降至0.830 1(下降13.72%);禁用 BERT 重排序机制(Abl-2),使前3正确率从0.547 2降至0.452 8(下降17.26%);关闭覆盖性匹配(Abl-3),则导致判责逻辑数量从3.207 5增至6.603 8(增长105.88%),印证了层次化语义解析在解耦复合诉求、避免单级

检索证据链断裂中的核心作用。

5　结束语

本研究提出的链接式 RAG 框架在电信客户投诉判责场景中技术优势显著：通过 2 级投诉点拆分机制实现复杂投诉层次化语义解耦，1 级混合式语义解析剥离情绪化表述与交织诉求，2 级核心投诉列表使相同投诉证据链匹配率提升 18.87%；动态证据链融合机制基于 2 级投诉列表并行检索，核心投诉点独立触发历史案例匹配并经大模型跨链推理，Recall@5 达 97.14%，显著优于传统 RAG；覆盖性验证算法通过双向匹配实现闭环判责，解决非结构化知识复用难题。

Linked-RAG 框架仍存在局限：1）依赖高质量标注数据构建知识库，小语种或新兴业务场景标注成本高，泛化能力受限；2）训练集以传统业务场景为主，新型投诉数据覆盖稀疏，模型在长尾场景鲁棒性不足，且增量学习存在灾难性遗忘风险。未来可从 3 方面突破：1）探索轻量化自监督学习策略，通过对比学习与领域自适应预训练减少标注依赖；2）构建弹性知识蒸馏框架，融合主动学习筛选高价值样本，设计特征解耦算法，分离通用规则与场景特异性知识；3）引入注意力可视化等可解释性增强技术，提升判责逻辑透明度，推动智能判责系统向“自适应、可解释、多模态协同”演进。

参考文献：

[1] GAMA F, MAGISTRETTI S. Artificial intelligence in innovation management: A review of innovation capabilities and a taxonomy of AI applications[J]. Journal of Product Innovation Management, 2025, 42(1): 76-111.

[2] 李芳, 陈震原, 肖军. 一种基于自然语言处理技术的智能定责应用研究[J]. 广东通信技术, 2023, 43(1): 8-12.
LI F, CHEN Z, XIAO J. Research on an intelligent responsibility determination application based on natural language processing technology[J]. Guangdong Communication Technology, 2023, 43(1): 8-12.

[3] SINGH A, CHANDRASSEKAR S, SAHA S, et al. Federated meta-learning for emotion and sentiment aware multi-modal complaint identification[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 16091-16103.

[4] FAN W, DING Y, NING L, et al. A survey on rag meeting LLMs: Towards retrieval-augmented large language models[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024: 6491-6501.

[5] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[6] KARPUKHIN V, OGUZ B, MIN S, et al. Dense passage retrieval for open-domain question answering[C]//EMNLP (1), 2020: 6769-6781.

[7] ROBERTSON S E, WALKER S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval[C]//SIGIR' 94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Organised by Dublin City University. London: Springer London, 1994: 232-241.

[8] HOU T D, JIN R, WANG Y Y, et al. A review of cross-modal retrieval research[J]. Journal of Computer Engineering and Applications, 2022, 58(24): 61-72.

[9] YU Y, PING W, LIU Z, et al. Rankrag: Unifying context ranking with retrieval-augmented generation in LLMs[J]. Advances in Neural Information Processing Systems, 2024, 37: 121156-121184.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[11] CHRISTIANO P F, LEIKE J, BROWN T, et al. Deep reinforcement learning from human preferences[J]. Advances in Neural Information Processing Systems, 2017, 30: 4299-4307.

[12] ASAI A, WU Z, WANG Y, et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection[C]//The Twelfth International Conference on Learning Representations, 2024.

[13] SHAHADE A K, DESHMUKH P V. Enhancing natural language processing: A comprehensive review of retrieval augmented generation[C]//2024 4th International Conference on Sustainable Expert Systems (ICSES), IEEE, 2024: 609-611.

[14] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. Qlora: Efficient finetuning of quantized LLMs[J]. Advances in Neural Information Processing Systems, 2023, 36: 10088-10115.