刘浩然，苗润生，叶陈常. RAG 与微调驱动的大语言模型应急管理问答方法[J]. 智能计算机与应用，2025，15(12)：164-170. DOI：10.20169/j.issn.2095-2163.25092801

RAG 与微调驱动的大语言模型应急管理问答方法

刘浩然[1]，苗润生[1]，叶陈常[2]
（1 上海理工大学 出版学院，上海 200093；2 同济大学 数学科学学院，上海 200092）

摘 要：应急管理需要为作业人员与公众提供及时且可核查的答案。本文面向法规、指南与事件报告等权威语料，构建基于检索增强生成（Retrieval Augmented Generation，RAG）的应急问答系统，并通过参数高效微调（Parameter Efficient Fine Tuning，PEFT）含 LoRA/DoRA 对中文大语言模型进行领域适配；检索层融合倒排索引、密集向量与知识图谱，辅以时间感知重排与句级引用，生成端统一适配 Baichuan-13B、ChatGLM3-6B 与 LLaMA2-13B，并采用可复现实验流程（索引冻结、来源与版本追踪、固定随机种）。在保留集上，相对无 RAG 基线，Baichuan-13B 的 BLEU-4 由 9.86 提升至 35.38，ROUGE-1/ROUGE-L 由 32.38/24.67 提升至 59.52/52.79，P95 延迟由 110.38 s 降至 60.99 s；ChatGLM3-6B 在质量—效率上表现最佳（samples/s≈6.57，steps/s≈0.86）。结果表明，结合 PEFT 与 RAG 可将通用大模型转化为可验证的应急问答助手，在事实准确性、术语一致性与响应效率方面实现综合提升。

关键词：应急管理；问答系统；检索增强生成；参数高效微调；知识图谱；大型语言模型
中图分类号：TP391.41 **文献标志码**：A **文章编号**：2095-2163(2025)12-0164-07

RAG and Fine-Tuning-Driven LLM Method for Emergency Management QA

LIU Haoran[1], MIAO Runsheng[1], YE Chenchang[2]
(1 College of Publishing, University of Shanghai for Science and Technology, Shanghai 200093, China;
2 School of Mathematical Sciences, Tongji University, Shanghai 200092, China)

Abstract: Emergency management requires timely and verifiable answers for practitioners and the public. We develop a Retrieval-Augmented Generation (RAG) question-answering system grounded in authoritative regulations, guidelines, and incident reports, and adapt Chinese large language models via Parameter-Efficient Fine-Tuning (PEFT) including LoRA/DoRA. The retrieval layer integrates sparse inverted indexes, dense vector search, and a knowledge graph with time-aware reranking and sentence-level citations, while the generation layer uniformly supports Baichuan-13B, ChatGLM3-6B, and LLaMA2-13B. A reproducible protocol is adopted, including frozen indexes, source/version tracking, and fixed random seeds. On the held-out set, compared with a non-RAG baseline, Baichuan-13B improves BLEU-4 from 9.86 to 35.38, ROUGE-1/ROUGE-L from 32.38/24.67 to 59.52/52.79, and reduces P95 latency from 110.38 s to 60.99 s; ChatGLM3-6B achieves the best quality-efficiency trade-off (samples/s ≈ 6.57; steps/s ≈ 0.86). These results indicate that combining PEFT with RAG transforms general-purpose LLMs into verifiable emergency-QA assistants, delivering consistent gains in factual accuracy, terminological consistency, and response efficiency.

Key words: emergency management; question answering system; Retrieval-Augmented Generation; Parameter-Efficient Fine-Tuning; Knowledge Graph; Large Language Model

0 引 言

高效的应急管理通过协调各机构、关键基础设施运营者和社区的准备、响应与恢复来支撑社会韧性。当今风险格局快速演化，多灾种事件、级联故障和信息过载频发，这使得及时且权威的知识获取以及跨部门协调变得尤为重要[1]。实证研究表明，当政府机构与公民志愿者及社区网络合作时，应急工作的效果会显著改善，凸显出建立可靠、透明且可扩展沟通渠道的必要性[2]。在此背景下，应急管理问

基金项目：上海市哲学社会科学规划课题(2022ETQ004)。
作者简介：刘浩然(2003—)，男，硕士研究生，主要研究方向：自然语言处理，知识图谱；叶陈常(2003—)，女，硕士研究生，主要研究方向：自然语言处理，大语言模型。
通信作者：苗润生(1990—)男，博士，讲师，主要研究方向：社交媒体挖掘，机器学习与运筹优化算法。Email：miaors@usst.edu.cn。
收稿日期：2025-09-28

答系统旨在使规范性知识和经验性知识(如法律法规、行动指南、标准操作程序以及事件事后分析报告等)能被即时检索和解读,服务于政府用户(值班官员、现场指挥等)和公众(提供防护行动指导)[3]。已有实践证明,对话界面可以改善指挥中心、现场团队与公众之间的信息流,从而降低突发事件中的响应延迟和沟通歧义[3]。对于管理者,此类系统支持决策的一致性、可审计性和合规性;对于公众,其通过将回答锚定于权威来源来减少不确定性并遏制谣言传播。

传统的信息检索或常见问答(IR/FAQ)工具难以应对碎片化且频繁更新的应急指南,以及使用非专业语言表述的自由问句。大型语言模型(Large Language Model, LLM)具备跨语言、跨文体的强大语义理解与生成能力,但直接应用 LLM 回答专业应急问题会面临事实不准确和“幻觉”等问题。参数高效微调(PEFT)技术可以在有限的算力和数据条件下快速将大模型适配到应急领域特定术语和任务上[4]。为确保答案的事实一致性和可验证性,检索增强生成(RAG)通过将模型回答依据于外部权威文档,减少了胡编乱造现象,并支持在法规和 SOP 快速演变情况下持续更新模型知识[5]。近期有研究将 LLM 与结构化知识结合以提升应急决策相关性和可信度[6]。最新综述亦总结了灾害管理各阶段部署此类技术的新兴最佳实践[7]。除了问答对话应用,当 LLM 在领域约束下与 GIS 等系统交互时,如用于洪水制图、空间分析等任务也取得了可复现的性能提升,表明有望构建用于监测和决策支持的智能应急方案[8]。

基于上述需求,本文提出的应急问答框架建立在 3 个支柱之上:持续治理和维护最新的应急知识语料库;利用 RAG 使回答与可验证的证据对齐;以及应用 PEFT 使 LLM 适应应急领域的术语和任务。其目标是为专业用户和公众提供与现行法规和指南一致、可操作且可溯源的答案。

1　相关工作

1.1　应急管理问答的传统方法

应急管理问答领域的发展大致沿着两条路线演进:一是基于检索与知识库的传统系统,依托精心维护的结构化知识与规程,采用 BM25 词法匹配、模板问答、知识图谱规则推理,突出可追溯、可审计与合规优势,即便规则常变也能稳态运行。近期将知识图谱与案例检索结合的工作,能为决策提供先例支撑,在搜救与资源分配等场景提升精准率与召回率,并与专家意见一致[9-10]。二是基于 LLM 的对话系统;但已有实践表明,非 LLM 路线在部署成熟度上仍具优势,例如 2024 年一项多国验证的移动应急聊天机器人,可有效降低处置突发事件的通信延迟并澄清指导信息[3]。总体看,结构化知识与案例相似性保障了建议的可复现性,而 LLM 路线正在补齐语义理解与生成能力。

1.2　应急问答中的检索增强生成方法

RAG 在生成前后并入法规、报告与监测数据等权威外部知识,弥补单一 LLM 在知识覆盖与事实一致性上的不足,是降低灾害问答“幻觉”的关键路径[5]。实践表明,WildfireGPT 通过检索最新野火材料可使回答更有依据[5];Xia 等[11]在台风问答中引入文档检索+精整知识库,显著提升正确性,但也发现若仅依赖检索而不做模型适配,低质或弱相关证据会误导 LLM、反致准确率下降。因此,近期工作倾向将高质量检索与适度微调/提示工程联用,保证模型正确吸收证据[12]。He 等[13]提出“两阶段 RAG”:先用领域数据微调,再在查询时动态检索,相较直接使用未适配模型,该方法的准确性与领域相关性均显著提升。总体而言,RAG 已成为将应急问答锚定到可检索证据的主流方案。

1.3　知识图谱在应急问答中的应用

知识图谱(Knowledge Graph, KG)以结构化关系与时效元数据为 LLM 提供约束,可显著提升答案的可解释性与可核查性[6]。例如,E-KELL 将 LLM 推理与应急管理 KG 结合,通过本体化建模 + CoT 引导,专家评审认为回答明显优于未使用结构化知识的 LLM[6];在滑坡场景中,基于 ChatGLM2 构建的滑坡灾害 KG 与推理模型,较传统方法更系统地联结要素、揭示隐含关系并改进监测与回答质量[14]。KG 的局限在于构建/维护成本高且 KG 不完备会导致推理盲区[11]。实践中,KG 常与 RAG 互补:KG 提供可信关系与约束,文本检索提供覆盖度与新鲜度,共同支撑可验证生成,并提升决策支持能力。此外,面向多源知识图谱的事件级融合方法已在工程实践中获得验证,可为应急场景的事实整合与冲突消解提供可复用方案[15]。

1.4　大型语言模型的参数高效微调技术

在算力与标注受限的应急场景下,PEFT(如 LoRA/DoRA)以少量可训练参数实现快速领域适配,无需改动大多数预训练权重即可显著提升问答质量与效率[4,16]。典型做法是在注意力层插入低秩

适配器以更新少量参数;进一步的参数敏感 LoRA(SensiLoRA)按权重敏感度自适应秩值,在领域 QA 数据上较标准 LoRA 更高效、更有效[13]。同时,QLoRA 将低秩适配与低比特量化结合,使百亿级模型也能在中等资源下微调。多项结果表明,“PEFT + RAG”可同时获得术语/表达的领域化与事实的证据化两类收益;其中“SensiLoRA + RAG”的两阶段方案在专用评测中优于全参微调或仅检索基线[13]。需要关注的风险包括灾难性遗忘与小数据偏差。与此相关,行业实证也显示,基于 Prompt 的低成本域适配在医疗等垂直场景可显著提升回答质量,呼应本文对轻量适配路径的主张[16]。整体看,PEFT 负责“学会说专业话”,RAG 负责“有据可查”,二者协同已成为应急问答的主流技术路线[17]。

2 知识图谱增强的应急问答系统设计

2.1 系统架构

图 1 概述了系统总体架构。快速适应、可验证、可演进、可管理,在生成端采用 LoRA/DoRA 等 PEFT 技术进行领域化适配,并以小样本增量保持更新[4,16];在推理端统一调度稀疏检索、稠密检索与知识图谱三通道,在生成阶段落实句级证据对齐与显式引用,形成证据到答案的可追溯映射[18-19]。为保证可核查与复现,法规与指南等文档按版本化元数据治理,记录发布机构、条款编号、生效与废止、适用辖区、版本号;同时保留从查询、检索到生成的审计信息与谱系指标,采用索引冻结、来源与版本追踪、固定随机种子等设置以确保可复现与合规。整体流程为:治理后的知识库快照进入多通道检索与 KG-RAG 融合提供证据,再由生成模块在证据约束下产生回答并嵌入引用。其中,特别是 RAG 将答案锚定在可检索来源上,保证每个结论都有依据[5];PEFT 提供对领域术语与风格的快速吸收能力。已有研究提出多引擎知识问答的总体思路,本系统在应急场景中通过文本检索与知识图谱的融合,实现了准确性与可用性的统一[20]。

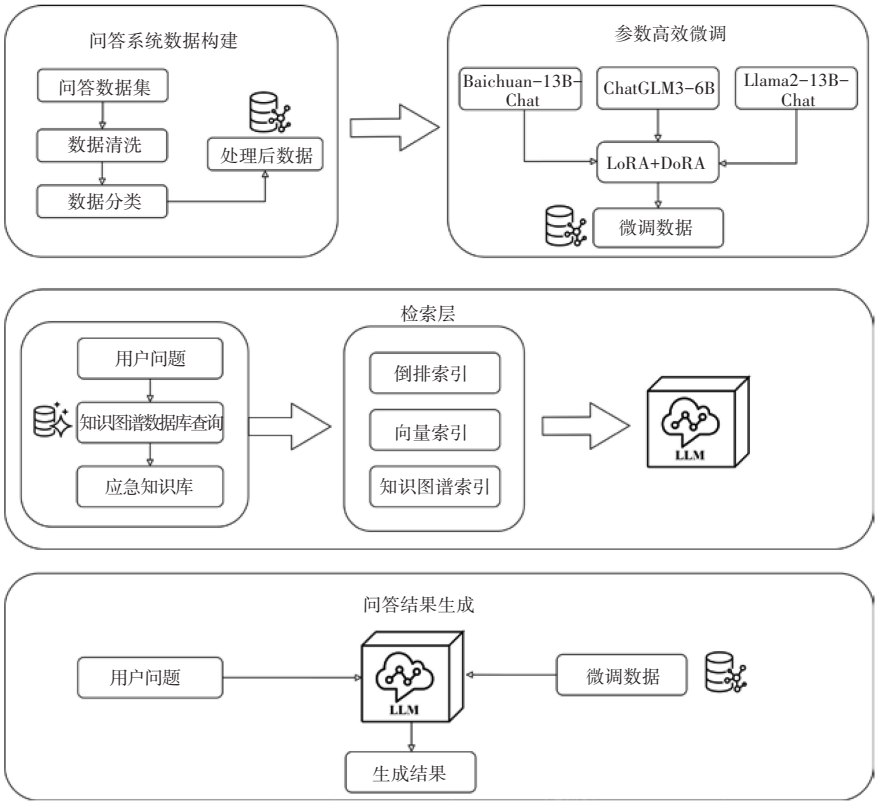


图 1 应急问答系统的方法论框架

Fig. 1 Methodological framework for the Emergency-QA system

2.2 应急领域知识图谱的构建与查询

按“法规—流程—角色—辖区—阈值—事件”构建本体;基于定期快照,以少样本/轻量微调的方式抽取实体、关系与流程槽位,生成包含来源锚点的三元组。同时,对每条三元组记录发布机构、条款编号、时效(生效/废止)、辖区与版本,保留历史快照并增量更新,线上索引指向当前有效。查询侧将自然语言解析为“实体—意图—约束”三元结构,遇授

权、阈值与关系判定自动生成参数化图查询(Cypher/SPARQL)，并裁剪 k-hop 子图；随后将图查询结果与稀疏/稠密文本命中合并，按词法、语义、KG 置信与时效重排，并在生成端插入句级引用。实践显示，KG+文本的双通道证据显著提升灾害问答的专业性与边界清晰度，且与密集检索的系统性发现相吻合[19]。

2.3　文档检索与索引模块

2.3.1　倒排索引

将文档切分为段落块并建立分字段 BM25/BM25F 索引(类型、机构、日期/版本、章节、辖区等)；查询重写对同义词、时态/地名变体、缩写与时间提示做规范化；结合短语/邻近匹配与字段加权优先命中标题、定义与规范条款；通过时间衰减机制与必要的类型过滤，控制过期及不相关内容的命中。对命中项记录术语及位置，便于后续审计与重排。

2.3.2　向量索引

使用领域自适应编码器+ANN 进行语义检索，支持增量并入新文档；通过分数校准/阈值调控缓解嵌入漂移问题，覆盖表述不一致与跨语种查询。每条命中保留来源指针(包含文档/段落或页码、版本标签、快照时间)，用于句级引用与回溯。为在词法精度与语义召回之间取得平衡，系统将稠密通道与倒排通道的排名进行互融(reciprocal-rank fusion)，并与知识图谱证据信号及时间新鲜度一并纳入统一打分。所有候选依据下式综合得分进行排序，词法相关度、语义相似度、图谱置信度与新近性共同决定进入生成端的证据顺序。

$$S(d\mid q)=(1-\lambda)\mathrm{BM25}(q,d)+\lambda\cos(e_q,e_d)+\mu\mathrm{KG}(q,d)+\nu e^{-\gamma\Delta t_d} \tag{1}$$

上式将词法、语义、图谱与新鲜度 4 类信号归一化后加权融合以排序证据；权重通过验证集调优，图谱信号内置冲突惩罚，新鲜度用指数衰减衡量文档时效。

2.3.3　知识图谱索引

KG 与应急本体对齐，覆盖灾害、资源、机构、地点、流程与阈值等核心概念；为图谱节点与类型化边绑定版本信息，每条三元组均关联文档/条款来源以便溯源。索引同时支持符号查询(模板/槽填充→裁剪目标关系子图)与语义扩展(KG 嵌入补全相关实体)；当问题涉及关系/授权/阈值时适当提升 KG 权重；若关系缺失或证据矛盾，系统发出证据不足信号以触发澄清/拒答策略。

2.4　大语言模型微调与回答生成

系统选用一款中等规模的 LLM 作为生成器，通过参数高效微调(PEFT)结合低秩适配器在生成表现与响应时延之间取得平衡，仅需更新极少量参数即可完成领域适配与增量更新[4,16]。适配器附着在各层注意力机制的 q_proj 和 v_proj 上，秩 $r=8$，alpha=16，dropout=0.10；对于 Baichuan 和 LLaMA 系列模型，结合使用 LoRA 与 DoRA，在相同显存预算下提升低秩容量，同时冻结基础模型参数，仅训练适配器权重(以及必要的层归一化系数)[16]。这种模型适配遵循标准的低秩参数化形式(见下式)，在微调过程中保持主干模型参数冻结，仅训练少量附加的适配器矩阵。

$$W'=W+\Delta W,\ \Delta W=B\cdot A,\ A\in R^{r\times k},\ B\in R^{d\times r} \tag{2}$$

由上式推出 $W'=W+B\cdot A$ 表示在冻结原始权重 W 的条件下，通过低秩增量 $B\cdot A$ 完成适配，仅训练小规模参数即可；适配器挂载位置与超参取值均采用上面的设置，用于平衡质量与时延。采用自回归语言建模目标(交叉熵)进行训练，训练阶段使用 Teacher Forcing 策略；共约 5 个 epoch，固定随机种子，并保存训练检查点与配置；训练集中交替加入“证据增强型问答/普通问答”样本，使模型在有检索支撑时插入句级引文，在证据不足或冲突时给出澄清或拒答。

3　实验结果与分析

3.1　实验数据集与实验设置

3.1.1　实验数据集

本研究构建的语料覆盖法规与应急预案、操作标准与技术手册、事故调查与评估报告及权威综述，面向定义/概念、流程/步骤、合规判定与行动要点等典型问答类型。共收集 86 份文档、121 921 条原始记录，清洗留存 75 443 条；每条均保留 DOI/URL 与页码或段落范围等出处信息。采用段落分块并适度重叠以兼顾召回与上下文；按来源与时间严格隔离训练/验证/测试，测试集未参与任何训练或微调。评估方案遵循近期关于 LLM 问答一致性与可复现性的指导原则[20]。

3.1.2　实验设置

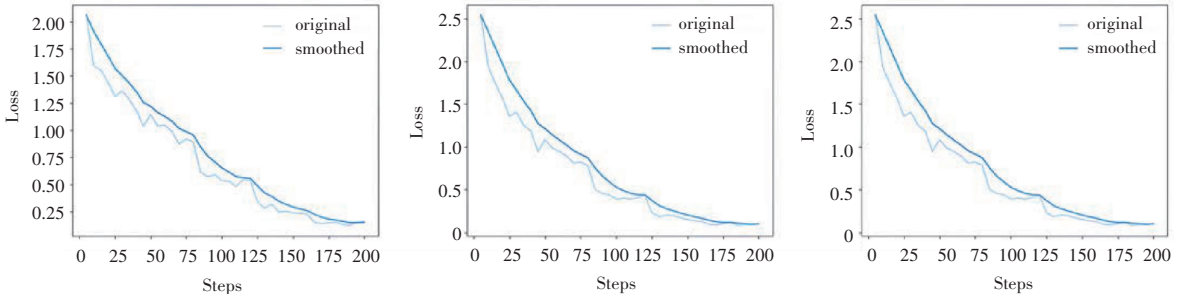
模型及微调：选用 Baichuan-13B-Chat、ChatGLM3-6B、LLaMA2-13B-Chat 进行参数高效微调。Baichuan 采用 LoRA+DoRA(lr=0.000 2，余弦调度，FP16)，ChatGLM3-6B 采用 LoRA ($r=8$, $\alpha=$

16,dropout=0.1)并结合 DoRA,LLaMA2-13B-Chat 沿用同一范式。统一固定随机种子、训练 5 个 epoch,配置与索引版本化存档;指令数据覆盖定义、流程、合规、操作等,含来源字段,按类型与来源分层采样并控制输入长度与证据范围,满足评测一致性与可复现性要求[21]。

RAG 流程:推理阶段在统一索引上定期更新法规、指南与报告,先做查询标准化与时间解析,再以 BM25 稀疏检索与向量检索联合召回,经过重排序融合后,将证据及来源、时间、页码或段落等元数据一并送入生成器,输出时插入句级引文并执行“无证据则拒答”。在统一提示与冻结索引下,同时运行直出与 RAG 两种模式,用以分离检索与生成带来的增益[21,22]。

评估指标与控制:文本质量报告 BLEU-4、ROUGE-1、ROUGE-L,效率报告响应延迟、samples/s、steps/s;检索侧采用 Recall@K、MRR、nDCG,并分析其与下游 EM/F1/ROUGE 的关联。对代表性问答进行句级审核与错误归因,对幻觉与不确定性按最新策略实施管控[22,23]。

图 2 显示了 Baichuan-13B-Chat (a)、ChatGLM3-6B (b) 和 LLaMA2-13B-Chat (c) 3 个模型在微调过程中损失值的变化曲线。可以看出,经过约 5 轮训练后,各模型的损失均已趋于收敛,其中 ChatGLM3-6B 的最终损失最低,Baichuan-13B-Chat 的损失在微调后降幅最大,而 LLaMA2-13B-Chat 的损失收敛速度相对较慢但最终值与 Baichuan 模型相近。



(a) Baichuan-13B-Chat (b) ChatGLM3-6B (c) Llama2-13B-Chat

图 2 模型训练损失曲线

Fig. 2 Model training loss curve

3.2 模型性能评估结果

收敛性及单模型提升:Baichuan-13B-Chat 在 5 个 epoch 后损失收敛至 0.685(图 2),采用 LoRA+DoRA 后质量与效率同步提升:BLEU-4 由 9.86 提升至 35.38,ROUGE-1 由 32.38 提升至 59.52,ROUGE-L 由 24.67 提升至 52.79;响应延迟由 110.38 s 降至 60.99 s。ChatGLM3-6B 损失收敛值约为 0.647 4,推理速度 ≈6.57 samples/s ≈0.86 steps/s;LLaMA2-13B-Chat 指标有增益,但吞吐与延迟不及前两者(见表 1)。结果表明 PEFT 能在应急问答中取得良好的质量-效率平衡[21]。

模型对比与任务适配:依据表 2,Baichuan-13B-Chat+RAG 更适合强调可验证性与内容完整性的场景;ChatGLM3-6B+RAG 适合优先低延迟与高吞吐的需求;追求综合折中时可选 LLaMA2-13B-Chat+RAG。该按任务约束进行模型选择的策略与最新评测与复现性建议一致。

表 1 模型微调前后数值对比

Table 1 Pre/Post Fine-tuning Metrics across Models

模型	阶段	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	预测运行时间	每秒处理样本数	每秒处理步骤数
Baichuan	微调前	9.86	32.38	10.71	24.67	110.38	2.075	0.263
Baichuan	微调后	35.38	59.52	32.54	52.79	60.99	2.263	0.295
GhatGLM	微调前	10.90	32.22	10.41	22.52	74.51	1.850	0.240
GhatGLM	微调后	36.91	58.05	30.53	50.87	20.99	6.570	0.860
LLaMA	微调前	33.19	44.38	15.88	3.20	114.61	1.204	0.157
LLaMA	微调后	32.94	57.57	30.40	51.20	79.91	1.727	0.225

表 2　跨模型比较指标

Table 2　Cross-model comparison metrics

模型	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	预测运行时间	每秒处理样本数	每秒处理步骤数
Baichuan2	35.38	59.52	32.54	52.79	60.99	2.263	0.295
ChatGLM3	36.91	58.05	30.53	50.87	20.99	6.570	0.860
LLaMA2	32.94	57.57	30.40	51.20	79.91	1.727	0.225

3.3　问答案例分析

检索增强效果及案例分析：对比不开启 RAG 和开启 RAG 两种模式下的模型回答可以发现，分析类/解释类查询在答案覆盖度、细粒度证据支撑和术语精确度方面有了显著提升，而定义类查询的改进相对有限——这与 RAG 主要用于补充背景信息和跨文档证据以提高答案可验证性的作用相符[21-22]。具体案例表明，经过检索增强优化后，模型能够在步骤、条件等方面提供条款级别的证据支撑，并通过跨文档综合减轻了引用过时信息的问题；这一表现与近期关于幻觉检测和不确定性控制的研究结论一致[22]（见表 3）。

表 3　RAG 融合与模型调参前后回答差异

Table 3　Answer differences before and after RAG integration and model tunin

问题	优化前	优化后（Fine-tuning+RAG）
广西省发生最频繁的自然灾害是什么？	广西省位于中国南部，地处桂西、桂中和桂东三个地貌区的交汇处，气候多样，自然灾害较为频繁. 根据历史数据统计，广西省发生最频繁的自然灾害是暴雨洪涝灾害. 由于广西省降水量较大，遇到强降水过程容易引发暴雨洪涝灾害，造成河流泛滥、农田浸泡、基础设施受损等影响.	广西省位于中国南部，气候多样，自然灾害较为频繁. 根据搜索结果，广西省发生较为频繁的自然灾害主要包括以下几种： 洪涝灾害：由于广西省地处南方，降水量较大，遇到强降雨时容易引发洪涝灾害，造成人员伤亡和财产损失 台风：广西省位于东亚季风区，每年夏秋季节容易受到台风的影响，台风可能导致强降雨、风暴等自然灾害，对当地生态环境和人类生活造成危害 地震：虽然广西省地震相对较少，但仍然存在一定的安全隐患 干旱灾害：广西省气候多样，干旱灾害也时有发生，特别是在北方地区，降水稀少的年份容易出现干旱灾害，影响农业生产和人类生活 综上所述，广西省发生较为频繁的自然灾害主要包括洪涝、台风、地震和干旱等. 在应对自然灾害方面，广西省政府和相关部门一直致力于加强防灾减灾工作，提高应对自然灾害的能力，以确保人民生命财产安全
水库对于应急防灾的重要作用是什么？	水库对于应急防灾的重要作用是防洪. 水库可以调节河流的水位和流量，降低洪水灾害的风险，减轻洪水对下游地区的压力，防止或减轻洪水灾害. 当发生洪水时，水库可以快速地调节水流，将洪水引导至水库，降低下游地区的风险. 因此，水库在防洪方面发挥着重要的作用，是应急防灾的关键设施	水库对于应急防灾具有重要作用，主要体现在以下几个方面： 控制洪水：水库可以有效地控制河流的洪水，防止洪水漫溢并造成周围地区的损失. 在发生洪水时，水库可以调节水流，降低上游地区的洪水压力，减轻下游地区的灾害影响 调节水位：水库可以调节上游和下游的水位，以应对不同的洪涝灾害. 当发生洪水时，通过调节水位，可以降低下游地区的风险；而在发生干旱时，可以通过调度水库来提供水源，保障水资源的供应 提供紧急水源：在地震、台风等自然灾害发生时，水库可以作为紧急水源，为受灾地区提供临时供水，保障人民生活必需品供应，降低灾害对当地居民生活的影响 改善生态环境：水库可以改善周围地区的生态环境，为水库周边提供水资源，促进植被恢复和水土保持，改善当地气候条件，减少自然灾害的发生 增强抗灾能力：水库可以增强抗震能力，减轻地震灾害对周边地区的破坏. 通过合理规划水库的位置和规模，可以在地震发生时调节水流，降低地震对水库周边地区的影响 总之，水库在应急防灾中发挥着重要作用，可以有效控制洪水、调节水位、提供紧急水源、改善生态环境和增强抗灾能力，为防灾减灾提供科学依据

4　结束语

本文在应急服务知识供给不足的现实下提出并实现了可部署的应急问答方案：生成端以 PEFT（LoRA/DoRA）完成低成本领域适配，推理端以 RAG 提供证据支撑与句级可验证，并配套法规/预案、技术手册与事故报告的版本化数据治理与受控评测。基于 Baichuan-13B-Chat、ChatGLM3-6B 与 LLaMA2-13B-Chat 的统一五轮微调与对照实验显示：ChatGLM3-6B 在 BLEU-4 与推理效率上更优，Baichuan-13B-Chat 在 ROUGE 与延迟优化上改善最显著，LLaMA2-13B-Chat 居中；RAG 显著提升分析/解释类问题的覆盖度、细粒度事实与证据对齐，对定义类问题提升有限。结合上游数据治理与句级引文机制，系统将回答从“可生成”提升为“可核验”，满足应急业务对可追溯与合规的要求。总体而言，PEFT+RAG 为核心的体系在证据约束下实现更广知识覆盖、更强可解释性与工程可落地性，为在更严格安全合规条件下的实际部署奠定了方法与实践基础。

参考文献

[1] YAZDANI M, LOOSEMORE M, MOJTAHEDI M, et al. Progress and landscape of disaster science: Insights from computational analyses[J]. International Journal of Disaster Risk Reduction, 2024, 108: 104536.

[2] KIRAC E, SHALTAYEV D, WOOD N. Evaluating the impact of citizen collaboration with government agencies in disaster response operations: An agent-based simulation study[J]. International Journal of Disaster Risk Reduction, 2024, 106: 104469.

[3] URBANELLI A, FRISIELLO A, BRUNO L, et al. The ERMES chatbot: A conversational communication tool for improved emergency management and disaster risk reduction [J]. International Journal of Disaster Risk Reduction, 2024, 112: 104792.

[4] DING N, QIN Y, YANG G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature Machine Intelligence, 2023, 5(3): 220-235.

[5] AMUGONGO L M, MASCHERONI P, BROOKS S G, et al. Retrieval augmented generation for large language models in healthcare: A systematic review[J]. PLoS Digital Health, 2025, 4(6): e0000877.

[6] CHEN M, TAO Z, TANG W, et al. Enhancing emergency decision-making with knowledge graphs and large language models[J]. International Journal of Disaster Risk Reduction, 2024, 113: 104804.

[7] XU F, MA J, LI N, et al. Large language model applications in disaster management: An interdisciplinary review [J]. International Journal of Disaster Risk Reduction, 2025, 127: 105642.

[8] ZHU J, DANG P, CAO Y, et al. A flood knowledge-constrained large language model interactable with GIS: Enhancing public risk perception of floods[J]. International Journal of Geographical Information Science, 2024, 38(4): 603-625.

[9] NASAR W, TORRES R D S, GUNDERSEN O E, et al. Improving search and rescue planning and resource allocation through case-based and concept-based retrieval[J]. Journal of Intelligent Information Systems, 2024, 62(5): 1431-1453.

[10] NASAR W, GUNDERSEN O E, DA SILVA TORRES R, et al. Knowledge graphs to accumulate and convey knowledge from past experiences in search and rescue planning and resource allocation [J]. Applied Artificial Intelligence, 2024, 38(1): 2434296.

[11] XIA Y, HUANG Y, QIU Q, et al. A question and answering service of typhoon disasters based on the T5 large language model [J]. ISPRS International Journal of Geo-Information, 2024, 13(5): 165.

[12] LIU S, MCCOY A B, WRIGHT A. Improving large language model applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines [J]. Journal of the American Medical Informatics Association, 2025, 32(4): 605-615.

[13] HE Y, ZHU X, LI D, et al. Enhancing large language models for specialized domains: A two-stage framework with parameter-sensitive LoRA fine-tuning and chain-of-thought RAG[J]. Electronics, 2025, 14(10): 1961.

[14] WU Z, YANG H, CAI Y, et al. Intelligent monitoring applications of landslide disaster knowledge graphs based on ChatGLM2[J]. Remote Sensing, 2024, 16(21): 4056.

[15] 王丹. 多源知识图谱事件知识融合方法研究[J]. 智能计算机与应用,2024,14(5):157-163.

[16] WANG L, CHEN S, JIANG L, et al. Parameter-efficient fine-tuning in large language models: A survey of methodologies[J]. Artificial Intelligence Review, 2025, 58(8): 227.

[17] 陆鑫涛,孙丽萍,童子龙,等. 基于 prompt 的医疗大语言模型自适应优化方法[J]. 智能计算机与应用,2025,15(8):190-196.

[18] ZHAO W X, LIU J, REN R, et al. Dense text retrieval based on pretrained language models: A survey[J]. ACM Transactions on Information Systems, 2024, 42(4): 1-60.

[19] SEQUEDA J, ALLEMANG D, JACOB B. Knowledge graphs as a source of trust for LLM-powered enterprise question answering [J]. Journal of Web Semantics, 2025, 85: 100858.

[20] 李春豹. 面向多类型知识库的多引擎智能问答系统设计[J]. 智能计算机与应用,2023,13(7):168-172.

[21] CHANG Y, WANG X, WANG J, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1-45.

[22] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[23] FARQUHAR S, KOSSEN J, KUHN L, et al. Detecting hallucinations in large language models using semantic entropy [J]. Nature, 2024, 630(8017): 625-630.