软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007703] [CSTR: 32375.14.jos.007703]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向地质灾害知识服务的多智能体协同推理框架*

冯 援[1], 胡 平[1], 张 璐[2], 沈复民[1], 申恒涛[3], 朱晓峰[4]

[1](电子科技大学 计算机科学与工程学院, 四川 成都 611731)

[2](大连理工大学 信息与通信工程学院, 辽宁 大连 116024)

[3](同济大学 计算机科学与技术学院, 上海 201804)

[4](海南大学 计算机科学与技术学院, 海南 海口 570228)

通信作者: 胡平, E-mail: pinghu@uestc.edu.cn; 朱晓峰, E-mail: zhuxf@hainanu.edu.cn

摘 要: 地质灾害知识服务在应急指挥、公共安全与科学传播中具有重要作用, 但现有基于大语言模型的问答系统在多源异构知识融合、复杂任务规划及高可信推理方面仍存在不足. 针对上述问题, 提出一种面向地质灾害知识服务的多智能体协同推理框架. 该框架以显式任务状态驱动为核心, 通过角色分工的智能体协作机制, 将复杂查询拆解为检索、分析、生成与验证等子任务, 并通过任务回写与闭环一致性校验实现推理过程的可追踪与结果可控. 在数据层面, 框架支持对科普专业、法律法规与灾害记录等多源知识的统一语义接入; 在推理层面, 构建了基于协同执行与质量控制的层次化推理流程. 实验结果表明, 该框架在领域知识覆盖、回答一致性与协同推理能力方面均优于单模型结合外部检索的基线方法, 尤其在法律法规与灾害统计等高可靠任务中表现出更稳定的性能. 为多智能体大模型在灾害应急知识服务中的应用提供了一种可行的系统化解决方案.

关键词: 大语言模型; 多智能体; 协同推理; 知识融合

中图法分类号: TP18

中文引用格式: 冯援, 胡平, 张璐, 沈复民, 申恒涛, 朱晓峰. 面向地质灾害知识服务的多智能体协同推理框架. 软件学报. http://www.jos.org.cn/1000-9825/7703.htm

英文引用格式: Feng Y, Hu P, Zhang L, Shen FM, Shen HT, Zhu XF. Multi-agent Collaborative Reasoning Framework for Geological Disaster Knowledge Services. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7703.htm

Multi-agent Collaborative Reasoning Framework for Geological Disaster Knowledge Services

FENG Yuan[1], HU Ping[1], ZHANG Lu[2], SHEN Fu-Min[1], SHEN Heng-Tao[3], ZHU Xiao-Feng[4]

[1](School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

[2](School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China)

[3](School of Computer Science and Technology, Tongji University, Shanghai 201804, China)

[4](School of Computer Science and Technology, Hainan University, Haikou 570228, China)

Abstract: Geological disaster knowledge services play a critical role in emergency command, public safety, and science communication. However, existing large language model (LLM)-based question-answering systems still have limitations in multi-source heterogeneous knowledge fusion, complex task planning, and highly trustworthy reasoning. To address these issues, this study proposes a multi-agent collaborative reasoning framework for geological disaster knowledge services. Driven by explicit task states, the proposed framework decomposes complex queries into subtasks such as retrieval, analysis, generation, and verification through a role-based agent collaboration mechanism and ensures the traceability of the reasoning process and the controllability of results through task write-back and closed-loop consistency verification. At the data level, the framework supports unified semantic access to knowledge from multiple sources, such as popular science and domain-specific knowledge, laws and regulations, and disaster records. At the reasoning level, a hierarchical reasoning

* 基金项目: 科技部重点研发项目 (2022YFA1004100); 国家自然科学基金 (62476048, 62572093); 四川省科技计划 (2025YFMS0004)
收稿时间: 2026-02-06; 修改时间: 2026-03-29; 采用时间: 2026-05-15; jos 在线出版时间: 2026-07-22

process based on collaborative execution and quality control is constructed. Experimental results demonstrate that the proposed framework outperforms the baseline method, which is based on a single model combined with external retrieval, in terms of domain knowledge coverage, answer consistency, and collaborative reasoning capability, and shows more stable performance in high-reliability tasks involving laws, regulations, and disaster statistics. This study provides a feasible systematic solution for the application of multi-agent LLMs in disaster emergency knowledge services.

Key words: large language model (LLM); multi-agent; collaborative reasoning; knowledge integration

我国幅员辽阔、地理条件复杂, 洪涝、地震、泥石流等自然灾害频繁发生, 对社会经济发展和人民生命安全造成严重威胁. 在极端气候事件增多和信息化水平不断提升的背景下, 公众、政府部门与科研机构对于灾害信息的可靠获取、智能分析与决策支持提出了更高要求, 亟需具备高效性、准确性和可解释性的知识服务能力. 然而, 传统系统以关键词、规则、静态知识库为主, 语义表示与跨源推理能力有限. 传统系统即便引入语义检索, 也常缺少面向应急任务的多阶段规划与验证机制, 难以处理灾害场景中的复杂背景知识、领域概念以及跨范畴信息, 更无法对法律政策、应急预案与专业知识进行综合表达与推理. 故而, 传统系统难以支撑面向应急管理的智能问答需求[1]. 这不仅限制了公共安全领域的知识服务水平, 也成为提升防灾减灾能力和现代化治理体系的瓶颈.

近年来, 大语言模型 (large language model, LLM) 在自然语言理解与生成方面取得显著进展[2], 为复杂场景下的知识服务提供了新的技术路径. 这些能力使大语言模型非常适合用于自然灾害管理, 并支持关键决策[3]. 然而, 现有 LLM 往往采用单模型、单智能体的短链推理范式, 缺乏任务分解、角色协作与显式规划能力, 并在很大程度上依赖于用户输入引导对话的过程[4]. 在灾害场景中, 面对跨数据源、跨知识类型的复杂问题, 该范式易出现知识盲区、推理链断裂与事实性偏差, 难以满足复杂任务的稳定执行需求. 与此同时, 灾害知识具有强异构性、强专业性与高可信要求等特点, 涉及地理、气象、法律、医学与历史记录等多源数据, 其语义粒度与结构形态不一致且缺乏统一表达方式, 使得单模型难以直接完成知识融合与可靠推理. 因此, 如何构建能够整合多源知识、执行多阶段推理并对输出质量进行有效控制的智能问答体系, 成为灾害知识服务领域亟待解决的关键科学与工程问题.

针对单模型知识覆盖不足与推理能力受限的问题, 相关研究主要沿两条技术路线发展: 其一是检索增强生成 (retrieval-augmented generation, RAG) 方法, 通过外部检索为生成过程提供可追溯证据, 从而提升事实一致性与知识覆盖度[5]; 其二是多智能体协同推理方法, 通过角色分工与交互合作将复杂任务拆解为检索、分析与验证等子任务, 以提高推理效率与可控性. 对于灾害知识服务场景而言, 两类技术均具有潜在适配性, 其核心需求如下.

(1) 多源知识统一接入与融合: 灾害问题通常涉及结构化数据库、向量索引语料与实时网络资源, 需要实现异构知识的统一接入、对齐与融合.

(2) 长链任务规划与执行反馈: 复杂查询对应多步推理链路, 需具备显式规划、动态重规划与过程可追踪能力[6].

(3) 质量控制与幻觉治理: LLM 易生成无依据内容[7], 需引入相关性评估、幻觉检测与答案验证机制[8].

(4) 可扩展工具调用与性能权衡: 灾害任务往往依赖外部工具进行检索与分析, 需要在召回率、时延与吞吐之间取得可用平衡[9].

尽管多智能体协同推理与 RAG 技术在通用智能对话和工具调用领域取得了一定进展[10], 但现有研究仍主要面向通用场景, 其目标集中在提高交互灵活性或增强工具链能力, 尚难直接满足灾害领域知识密集、约束严格、异构显著的服务要求. 例如, 以 CAMEL[4]为代表的框架试图通过角色分工改进复杂任务执行, 但高度依赖固定提示、缺乏对专业知识源的结构化融合能力, 难以适配专业领域的知识密集型推理需求. 而 RAG 体系虽然增强了事实一致性, 却在多轮任务规划、知识一致性维护和跨语义空间的数据融合方面的能力有限[11]. 因此, 通用多智能体系统与 RAG 技术在面对灾害领域这一具有强专业性、强约束性和强异构性的知识服务场景时仍存在明显不足, 这一领域亟需面向异构知识融合与高可信推理的系统化解决方案.

针对上述需求与挑战, 本文提出一种面向地质灾害知识服务的多智能体协同推理框架. 不同于现有多智能体方法主要依赖对话驱动的协作模式, 本文从协同推理机制角度出发, 提出一种以“显式任务规划-协同执行-结果回写-闭环验证”为核心的协同推理机制. 该机制以任务状态为中心来组织多智能体协作过程, 通过统一的任务表示与回写机制, 实现推理过程的可追踪、可验证与可控, 从而更适用于高可信知识服务场景.

综上, 本文主要贡献如下.

(1) 提出一种以显式任务状态驱动的多智能体协同推理机制. 通过结构化任务拆解与状态回写, 实现多智能体之间的有序协作, 从方法层面提升复杂知识服务场景下推理过程的可控性与可靠性.

(2) 构建面向灾害领域的异构知识统一接入与融合机制. 通过统一语义接入与多级质量控制策略, 实现多源知识的稳定支撑, 从知识层面提升协同推理结果的可靠性与一致性.

(3) 搭建覆盖典型地质灾害任务的数据体系与测评标准并开展系统性实验评估. 围绕多阶段推理与多源知识融合场景进行验证, 从实验层面验证所提方法在知识覆盖与回答一致性方面的有效性. 相关代码和数据集见 GitHub (https://github.com/yeuei/Multi-Agent-For-Disaster).

1 相关工作

本节回顾多智能体框架与协同机制的研究进展, 并重点梳理任务规划、工具调用与知识增强 (RAG、NL2SQL) 等关键技术, 最后总结多智能体在行业场景中的应用现状与局限. 通过相关工作可以说明融合 AI 智能体的地质灾害知识服务, 不仅能够有效应对多种场景的复杂任务, 还可以结合不同信息源进一步促进策略对齐与协作.

1.1 多智能体框架与协同机制

多智能体研究旨在提升语言模型在复杂任务场景下的协作能力与分布式问题求解能力. 早期工作以角色扮演与对话协作为主要形式. CAMEL[4]通过角色设定与 inception prompt 引导多个聊天代理保持一致意图并模拟“LLM 社会”行为; AutoGen[10]提出通用多智能体框架, 支持 LLM、人类用户与工具的多方式协作; HuggingGPT[12]将 LLM 作为控制器, 对用户需求进行任务规划与模型选择, 以完成跨模态与跨领域任务. 随后有研究进一步增强代理间的协作结构, 例如, ProAgent[13]引入 Planner、Verificator、Controller 与 Memory 等模块并结合 belief correction 机制, 实现了协同推理; SMART[14]则面向知识密集型任务, 通过引入外部知识提升回答的解释性与一致性. 随着研究的深入, 多智能体逐渐扩展至虚拟环境与工程场景. Generative agents[15]在虚拟社区游戏中部署具备记忆记录、反思与规划能力的代理, 实现涌现式的社会行为; MechAgents[16]和 LLM-enabled manufacturing[17]将多代理应用于工程与制造场景, 实现代码生成、数值求解与流程协同. 针对工具使用能力, Shen 等人[18]提出将工具调用分解为 Planner、Caller 与 Summarizer 等组件以提升调用效率. 在应用层面, 多智能体还服务于知识标注与软件开发任务, 例如, Li 等人[19]提出使用 LLM-based multi-agent system 进行知识标注; ReDel[20]提供递归多智能体系统与运行回放能力以增强可追溯性. 为评估多智能体能力, 研究者提出多种基准与多智能体构建方法, MAgIC[21]基于社交博弈构建竞争式评测框架, 从而评估推理、欺骗与合作行为; Liang 等人[22]通过辩论促进发散性思考; Debate or vote[23]进一步分析了辩论与投票机制对性能的影响; MAPoRL[24]通过验证评分提升强化学习中的协作质量; OpenAGI[25]构建用于多步骤现实任务的开源 AGI 平台; Guo 等人[26]系统性梳理了数据集、应用场景与系统架构; Cemri 等人[27]总结多智能体 LLM 系统的失败模式与风险点, 为后续研究提供诊断视角.

总体而言, 现有多智能体研究在通用任务协作与工具编排方面进展明显, 但多聚焦框架能力验证与交互行为模拟, 对专业领域知识服务的支撑仍不足: 缺乏面向异构知识源的结构化融合与约束建模, 亦缺少面向高可靠场景的证据对齐与闭环验证链条, 这限制了其在灾害知识服务等强约束应用中的落地.

1.2 任务规划、工具增强与知识增强

在复杂任务场景中, 提升大语言模型 (LLM) 的决策能力通常需要同时具备长链推理、任务规划与外部知识调用能力. 围绕任务规划方向, 研究者提出多种增强方法以突破单模型短链推理的局限. Plan-and-Act[28]采用规划-执行解耦机制, 由 Planner 生成结构化计划、Executor 据环境反馈执行, 并在失败时触发重规划提升鲁棒性, 并发现在有关智能体的长程任务中, 搜索、查询策略不当可能会造成关键信息缺失并引发计划失效. Tree-of-thought (ToT) 系列方法则将推理过程显式构建为可搜索树结构, 其中, LLM guided ToT[29]通过引入提示代理、检查模块与记忆模块提升数独问题求解成功率; Tree of thoughts[6]采用广搜和深搜方法探索多候选推理路径, 在数理推理与创意写作任务中表现显著.

另一类研究聚焦于工具增强以突破 LLM 在环境理解与操作能力上的限制. Toolformer[30]通过自监督训练让模型学习工具调用策略; ViperGPT[31]将视觉推理任务转化为可执行程序以调用外部视觉模型实现复杂任务无需额外训练; ReAct[32]提出推理与行动交替执行机制以强化模型行动能力; Voyager[33]在 Minecraft 环境中通过持续探索构建技能库以实现终生学习. 在工具生态层面, Kani[34]提供简易灵活的工具调用与对话管理框架; FanOutQA[35]用于评估长上下文跨文档问答能力; ReDel[20]支持多智能体事件回放与状态管理. 近期, Anthropic 推出模型上下文协议 (model context protocol, MCP)[36], 以统一模型访问外部资源的接口标准, 从而扩展模型工具能力边界.

在知识增强方面, 其中一种技术是检索增强生成 (RAG), 由 Lewis 等人[5]提出, 通过外部检索增强模型的事实性与覆盖性, 成为知识密集型任务的重要技术路径. 然而 RAG 在实践中仍可能产生“看似合理但错误”的幻觉输出, 引发质量控制需求. 为此, 研究者提出多种质量治理策略, 包括: RAG-HAT[7]结合幻觉检测与 GPT-4 Turbo 纠错并通过 DPO 训练降低幻觉; Self-RAG[8]使模型在推理中按需检索并自我反思以提升一致性; FLARE[37]通过未来内容预测决定是否检索以改善长文本事实一致性; SELF-REFINE[38]作为一种通用增强策略, 采用“自反馈-迭代精炼”提升生成质量. 支撑 RAG 的关键基础设施还包括高性能向量数据库, 如 GaussDB-Vector[9]实现低延迟向量检索与持久化更新; Milvus[39]支持动态数据更新与可扩展部署.

知识增强的另一种技术是自然语言生成 SQL (natural language to SQL, NL2SQL), 这项技术可以让模型获取数据库中的结构化知识. 现有研究主要通过上下文学习与微调策略提升模型性能: ACT-SQL[40]利用自动思维链 (CoT) 生成技术, 在零人工标注成本下显著增强了模型的推理准确率. MIGA[41]则提出两阶段多任务生成框架, 通过预训练阶段的多任务学习与微调阶段的 SQL 扰动策略, 有效抑制错误传播并提升泛化性. 此外, Gao 等人[42]针对 Text-to-SQL 任务中的提示工程开展系统性优化, 提出兼顾性能与 token 效率的 DAIL-SQL 方案, 利用问题与查询骨架的双重相似度筛选高质量示例. 然而, 中文 NL2SQL 领域仍面临独特挑战[43]: 复杂的语法语义结构显著增加了解析难度, 而早期高质量中文数据集的匮乏进一步限制了该领域的进展.

综上, 任务规划、工具增强与知识增强为“长链推理+外部知识获取”提供了关键技术支撑, 但多数方法面向通用任务, 普遍存在缺乏: (1) 面向专业知识服务的领域适配; (2) 面向异构知识源的对齐融合; (3) 面向高可信场景的端到端质量控制链条; (4) 与多智能体协作框架的联动优化机制. 因此, 在灾害应急等需要高可靠的场景下仍难以直接落地, 亟需面向领域知识服务的协同推理与可信增强一体化方案.

1.3 大语言模型在灾害应急管理中的应用

在灾害应急管理领域, 大语言模型已被广泛应用于知识问答与决策支持任务. 在模型适应性研究方面, Han 等人[44]基于低秩自适应 (low-rank adaptation, LoRA) 微调实现了多灾种线索与空间实体的统一提取; Zafarmomen 等人[45]则利用长周期气象报告数据训练并评估了模型对天气事件的分类能力, 证实经 LoRA 微调的 LLaMA-3.1 模型能以更低的训练成本获得最优 $F1$ 分数. 在系统架构方面, 王喆等人[1]提出的 KG-GPT 系统融合了知识图谱推理与 GPT 的关键信息识别能力, 显著提升了决策问答的逻辑性; 蔡剑寒等人[46]进一步构建了以知识图谱为核心的 3 层语言应急服务架构. 总体来看, 现有研究多依赖知识图谱与规则系统, 侧重信息采集与组织; 但在多智能体协同、复杂任务规划、跨源知识融合与输出闭环验证方面仍不足, 难以充分满足灾害场景对高可信生成与智能决策支持的需求. 此外, Xie 等人[47]探索了多智能体协作模式, 提出面向灾害适应的检索增强生成系统 MARSHA, 通过用户画像、规划与分析等专用智能体的协同, 实现了多源灾害观测数据与科学文献的动态检索与个性化整合, 但其高度依赖语义向量匹配, 缺乏对复杂结构化数据的精准查询与要素对齐机制. 总体来看, 现有研究多依赖知识图谱与规则系统, 侧重信息采集与组织; 尽管已有少量研究初步引入了智能体概念, 但总体在多智能体协同、复杂任务规划、跨源知识融合与输出闭环验证方面仍存在不足, 难以充分满足灾害场景对高可信生成与智能决策支持的需求.

2 多智能体协作的地质灾害知识服务系统

本节介绍提出的多智能体协同的地质灾害知识服务系统. 整体上, 系统可划分为 5 个主要层次: 设计目标与系统架构、领域知识与数据资源、任务规划与协同执行机制、领域智能体设计与实现以及检索增强与质量控制机制.

2.1 设计目标与系统架构

地质灾害知识服务具有专业性强、知识来源异构、可靠性要求高等特点. 系统不仅需要面向公众提供科普解释, 还应支持专家与政府部门开展灾害评估、法规咨询与应急决策等任务. 为满足上述需求, 本文构建了如图 1 所示的基于大语言模型的多智能体协同推理框架, 通过显式任务规划、多源知识调用与质量控制实现高可信知识服务.

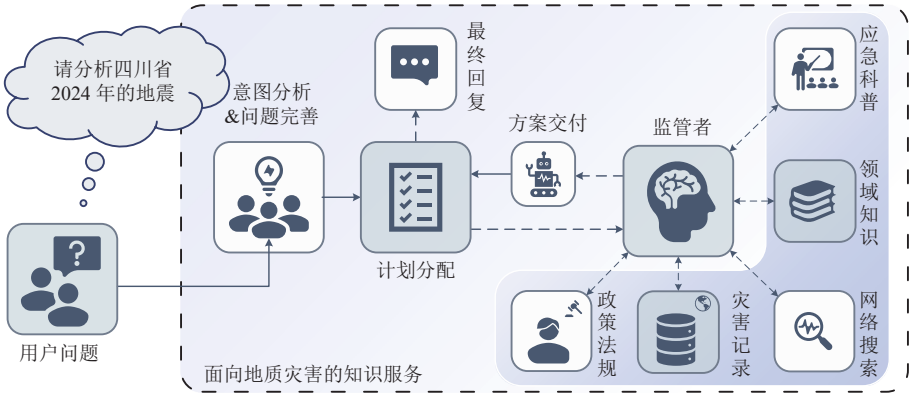


图 1 面向地质灾害知识服务的多智能体协作框架

在模型层, 系统以在灾害相关语料上采用 LoRA 微调的 Qwen2.5-7B-Instruct 作为基础语言能力. 为增强复杂任务的执行能力, 在系统层采用模块化设计引入任务调度、多源检索、结构化查询、法律分析与网络搜索等功能模块, 各模块通过统一接口与上下文管理机制交互, 从而形成弱耦合、可扩展的整体结构. 针对多样化任务需求, 系统预设领域专家、法律咨询、网络搜索与监管者等角色智能体, 并为每个智能体配置专用系统提示词与行为约束, 以保证输出风格一致、职责边界清晰.

系统运行流程如下: 用户提出问题后, 意图分析模块完成任务理解与分解; 随后在监管代理的监督与调度下, 各领域智能体协同完成子任务并将中间结果返回; 最后由最终回复模块在质量控制机制约束下生成完整答复. 除核心大语言模型外, 系统还集成向量检索、结构化查询与任务调度等独立服务模块, 这些模块通过标准 API 与共享文件系统进行通信. 监管者智能体同时承担路由器角色, 负责在不同智能体与工具之间动态分发请求, 从而在不破坏整体稳定性的前提下支持新能力模块的接入与扩展.

我们提出的多智能体协作框架在架构层面具备可推广性. 该框架采用模块化设计, 核心由“监管者代理”进行统一调度, 这赋予了系统“即插即用”特性. 开发者能够根据不同领域的具体需求, 灵活增删或修改特定的 agent 与底层结构化/非结构化知识源; 因此, 只需替换相应的领域知识库, 该框架即可轻松扩展至法律、医疗知识问答等其他知识密集型领域. 在具体应用上, 本框架高度契合复杂的知识密集型任务. 面对复杂问题, 系统能够将其逻辑拆解为更细粒度的子任务, 进而协同调用基于向量库的语义查询、基于关系数据库的结构化查询、专家模型问答以及网络补充搜索, 从而实现领域知识的深度整合. 当然, 这种协作机制的有效运转依赖于两个核心前提: 一是目标任务本身必须具备可拆解性, 以便将子意图分配给不同的子 agent 独立执行; 二是目标领域需具备一定的多源知识储备作为基础, 以支撑多渠道的信息检索与推理. 此外, 当前方法存在一定的局限性, 受限于实验所采用的纯文本底座模型, 系统目前仅能处理纯文本层面的知识问答; 若要应对包含图表、遥感影像在内的多模态复杂任务, 则必须将底座升级为多模态大语言模型. 同时, 系统最终输出的准确性, 也和底层知识库的初始构建质量有关.

2.2 领域知识与数据资源

为保障系统在地质灾害场景下的知识覆盖与证据可追溯性, 本文构建了融合非结构化文本知识与结构化数据库的多源知识资源体系. 该体系围绕地质灾害知识服务的典型需求, 覆盖灾害科普解释、专业问答、法规条款解读以及统计分析等任务, 为后续检索增强生成与结构化查询提供统一的数据基础. 总体而言, 数据资源分为非结构化文本数据与结构化数据两类, 其规模统计如表 1 所示.

表 1 数据分布统计

数据集名称	数据量
法规原文语料库 (政策法规文本)	1452页
法规问答指令数据集 (指令微调)	15000条
灾害专业问答指令数据集 (指令微调)	15000条
灾害科普问答指令数据集 (指令微调)	15000条
历史灾害记录数据库 (结构化数据)	266373条

在非结构化数据方面, 为增强模型的领域认知与法律推理能力, 本文开展了系统性的采集与清洗工作: 首先, 从中央及各地方政府门户完整采集地质灾害相关政策法规文本, 构建法规原文语料库 (政策法规文本), 共计 1452 页, 覆盖从防灾减灾法律体系到地方性应急预案等内容; 其次, 面向法规解读任务构建法规问答指令数据集 (指令微调) 15000 条. 同时, 针对灾害知识服务中的不同受众与任务类型, 分别构建灾害专业问答指令数据集 (指令微调) 与灾害科普问答指令数据集 (指令微调), 各 15000 条. 上述指令数据均按照训练/测试划分, 其中 10000 条用于模型指令微调训练, 5000 条用于留出测试, 并统一格式化为通用的 Alpaca 格式以适配大模型微调流程. 在结构化数据方面, 为支撑定量分析与统计查询需求, 本文构建了历史灾害记录数据库 (结构化数据), 共包含 266373 条记录, 覆盖全国各省份的历史灾害统计信息. 数据库构建流程包括: 从公开渠道收集灾害统计条目, 统一字段口径与编码规范 (如地区、灾种、时间、指标等), 对缺失值与重复记录进行清洗处理, 并对关键字段进行一致性校验; 最终将清洗后的数据存储于轻量级的 SQLite 数据库中. 该结构化存储方式便于开展 SQL 查询与聚合统计, 并为系统的 NL2SQL 模块提供统一的数据接口与可追溯的事实证据支撑.

2.3 任务规划与协同执行机制

为使系统在多源知识接入、用户诉求差异与复杂问题约束下保持稳定运行, 本文将任务规划作为多智能体协同的入口机制: 系统首先通过意图识别完成需求澄清与任务分解, 再在调度与监管机制的控制下按计划逐步执行各子任务, 并通过回写与验证形成闭环, 从而降低一次性生成带来的遗漏与偏差风险.

2.3.1 意图识别与任务表示

系统采用显式规划机制对复杂问题进行拆解与追踪. 面向用户的首个模块为问题完善代理 (亦可视为意图分析智能体), 其主要职责是: 结合用户职业背景与使用场景 (如新闻媒体、科研人员、政府工作人员或普通公众) 推断提问动机与期望输出形式; 联动解析用户输入及历史对话上下文, 抽取关键实体与约束条件 (如灾种、地点、时间范围、事件要素等); 同时识别潜在任务线索 (如需要调用的数据来源、可能涉及的法规依据、需补充核验的事实点). 在此基础上, 问题完善代理进一步推断与主问题强相关的隐含子问题, 将整体需求自顶向下拆分为若干可执行子任务, 每个子任务对应具体的检索、证据汇聚或分析生成步骤.

实现层面, 为保证规划结果的可读性与可追踪性, 问题完善代理在提示词模板约束下生成统一的规划文档 plan.md, 以 Markdown 结构化格式记录子任务描述、执行状态等信息, 为后续模块提供标准化的任务入口与上下文承载. 随着协同执行推进, 各子智能体的中间结论与关键依据将持续写回 plan.md, 使任务过程具备可解释、可追踪与可复用特性. 图 2 给出了 plan.md 的示例; 除子任务列表与对应输出外, 系统还在每个子任务前引入由方案交付代理生成的判定标签, 用于核验子智能体输出是否覆盖并完成既定子任务, 从而为后续调度与汇总提供明确的质量信号.

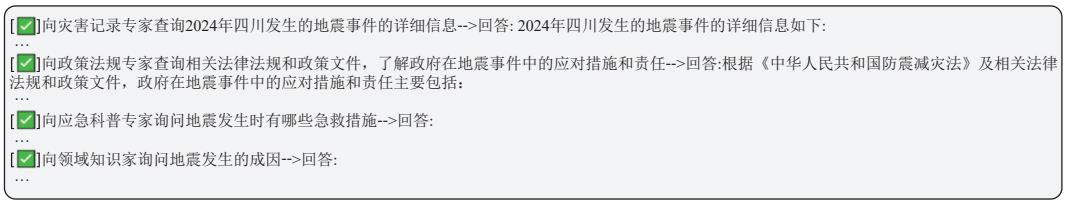


图 2 plan.md 示例

2.3.2 调度程序与协同执行

在协同执行阶段, 系统采用轻量调度程序并结合监管者智能体, 实现“按计划推进”的流水式协作流程: 调度程序顺序扫描 plan.md, 定位尚未完成的子任务并触发执行; 监管者智能体依据任务类型与所需资源, 将子任务路由至最合适的专家智能体或工具模块, 例如法规检索、历史灾害数据库查询或外部网络搜索等. 子任务完成后, 方案交付代理对输出进行压缩整理并给出完成判定, 同时将结果与状态更新回写 plan.md; 当全部子任务完成后, 由回答智能体读取 plan.md 的结构化结果进行一致性整合与结构化表述, 生成最终答复. 上述机制形成“任务→调度→执行→交付回写→汇总生成”的闭环流程, 一方面通过显式状态管理降低复杂问题一次性生成的遗漏风险; 另一方面通过判定标签与回写机制实现过程可控与结果可核验, 从而提升多智能体协同下知识服务的稳定性与可靠性. 监管代理与方案交付代理的关键提示词分别如图 3、图 4 所示.

```
你是一个总管代理, 负责统筹调用以下5个专家代理, 根据‘总任务’和‘专家代理的反馈’对专家代理进行调度.
你要尽可能确保总任务完全被执行, 不能遗漏任何一部分.
1. **law_agent**: 一个专业的法律代理模型, 拥有丰富的法律知识, 可以回答与法律相关的问题. \n…\n2. **knowledge_agent**: 提供自然灾害领域的系统性知识. \n…\n3. **websearch_agent**: 具备网络实时搜索能力. \n…\n4. **document_agent**: 专注于执行传统的SQL数据库查询. \n…
\n5. **emergency_agent**: 提供应急科普与行动建议. \n…
…
—输出格式（唯一合法格式）—
只允许输出一个JSON对象, 包含字段: \n**question**: 要发送给所选代理的具体问题. \n**go_to**: 要调用的代理名称, 只能是‘law_agent’、‘knowledge_agent’、‘websearch_agent’、‘document_agent’、‘emergency_agent’、‘finish’之一. \n当go_to为‘finish’时, question为空字符串.
\n**禁止输出任何额外文本、解释、代码块或前后缀.**
—示例输出—
…
---
输出:
```

图 3 监管智能体的关键提示词

```
你需要根据给定的「总任务」和「上下文」, 给出一个总结性回答, 并判断现有信息是否足以完成总任务. \n你的输出必须是一个JSON对象, 包含以下两个字段:
1. information: 在严格尊重「上下文」的前提下, 尽可能完整、清晰地回答「总任务」.\n…
2. is_ok: 根据「上下文」的信息是否足以完成「总任务」进行判断:
 - 若上下文已包含完成总任务所需的关键信息, 可以给出相对完整、可靠的回答, 则填“yes”;
 - 若上下文明显缺少关键信息、只能给出部分回答或只能说明无法回答, 则填“no”.
输出要求 (严格遵守):
只输出一个合法的JSON对象; \n不要输出任何额外文字、说明或标点 (如反引号、注释、自然语言解释等); \nJSON中只能包含字段“information”和“is_ok”, 不得增加其他字段.
示例1 (上下文信息充分, is_ok为“yes”):
…
总任务: {}
上下文: {}

---
输出:
```

图 4 方案交付智能体的关键提示词

2.4 领域智能体设计与实现

为覆盖灾害知识服务中“科普解释-专业分析-法规咨询-结构化统计-实时信息补全”等多样化需求, 本文构建了由多类角色智能体组成的领域智能体体系. 各智能体围绕其任务边界配置专用提示词与工具接口, 并通过监管代理统一注册与调度, 从而在第 2.3 节所述的 plan.md 驱动的协同流程中按需调度、执行、交付回写和汇总生成.

2.4.1 科普与专业问答智能体

系统面向不同用户群体与知识需求的颗粒度, 设计了应急科普智能体与领域知识智能体两类核心专家. 其中, 应急科普智能体主要服务公众用户与非专业场景, 侧重以通俗易懂的方式解释自然灾害现象、应急管理概念及基础防护要点, 其职责边界聚焦于面向大众的科普传播与基础应急知识解答; 领域知识智能体面向科研专家等专业用户, 聚焦提供更细粒度的技术性回答, 如工程治理思路、气象触发条件与灾害机理分析等, 其职责边界限定于面向专业人员的深度领域知识支撑和技术问题解析.

两类智能体均以 Qwen2.5-7B-Instruct 为基座模型, 并采用 LoRA 进行参数高效微调[48]. 该方法通过冻结预训

练模型权重不直接更新模型密集参数, 而是引入低秩增量 A 和 B 近似参数来更新, 从而以少量可训练参数实现微调, 可以表示为公式 (1):

$$h = (W_0 + \triangle W)x = W_0 x + (BA)x \tag{1}$$

其中, $h \in \mathbb{R}^d$、$W_0 \in \mathbb{R}^{d\times k}$ 是预训练模型的初始参数, $B \in \mathbb{R}^{d\times r}$、$A \in \mathbb{R}^{r\times k}$、$x \in \mathbb{R}^k$ 表示输入, 且矩阵的秩 $r << \min(d,k)$.

在具体实现上, 本文分别基于灾害科普问答指令数据集与灾害专业问答指令数据集对模型进行 LoRA 微调, 使两类智能体在术语体系、知识边界与表达风格上形成差异化能力. 系统层面, 两者作为可调用专家能力统一注册至监管代理的可用代理表; 当子计划涉及科普解释或专业说明任务时, 监管代理按任务需求调用相应专家并回传结果, 支撑后续汇总生成.

2.4.2 网络搜索智能体

为弥补内部知识库对最新信息或外部开放资源覆盖不足的情况, 系统设计网络搜索智能体, 其职责边界聚焦于外部信息检索与补充, 仅为系统提供知识库以外的实时、开放信息支撑, 不承担知识推理等任务. 该智能体具备自主决策能力: 根据子任务描述判断是否需要外部检索, 当任务依赖实时信息或当前上下文不足以支撑回答时, 触发外部搜索与阅读流程.

在实现上, 网络搜索智能体通过调用秘塔搜索平台提供的 MCP Server 接入外部检索工具, 核心工具包括用于网页检索的 metaso_web_search 与用于页面内容解析的 metaso_web_reader. 其调用流程为: 客户端与 MCP Server 建立连接并完成协议版本与能力协商; 通过 tools/list 获取工具定义及 JSON Schema; 随后依据 JSON Schema 构造 tools/call 请求发起工具调用; MCP Server 按 JSON-RPC 2.0 返回检索结果内容块, 客户端解析后进入候选筛选与阅读提取阶段, 并将摘要化证据回传上层智能体参与后续汇总. 该智能体的关键提示词如图 5 所示.

> 你是可以调用网络搜索工具的助手, 但如果用户的问题你可以直接回答, 请直接回答; 如果用户的问题无法直接回答, 请进行网络搜索, 并返回搜索结果.

图 5 网络搜索智能体的关键提示词

2.4.3 政策法规智能体

面向灾害相关法规解释与合规咨询场景, 系统设计法规问答智能体, 支持法规语义理解、条款依据检索、合规性分析与情景化法律咨询等任务, 其职责边界聚焦于灾害领域相关法规的专业咨询服务, 仅围绕法规条款解读、合规性判定与法律风险提示提供支撑, 不承担非法规类的通用知识问答与决策生成任务. 为提升在突发事件应对、防震减灾等领域的专业表达与条款适配能力, 本文在 Qwen2.5-7B-Instruct 基座上采用 LoRA 进行法律问答能力的微调, 形成专用法律回答模型, 使其在引用法规依据、给出合规建议与事实核查方面更贴近法律咨询需求. 运行机制上, law_agent 参考 Self-RAG[8]与 SELF-REFINE[38]的思想, 采用“检索-生成-检测-重试”的迭代式框架 (见图 6).

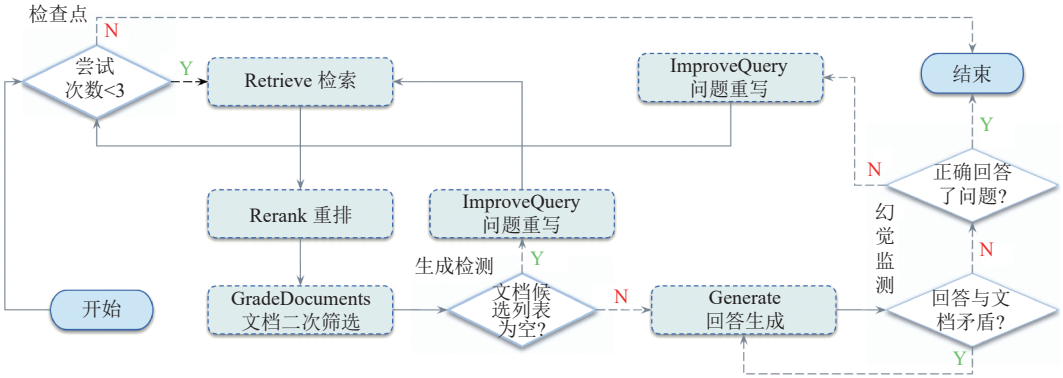


图 6 政策法规智能体执行流程

其内部流程由如下多个子模块协同组成.

1) 检查点控制: 限制迭代与重试次数, 避免死循环;

2) 法规检索: 返回与问题最相关的候选法规文档 (如 Top-20, 包含标题与原文等元信息);

3) 重排与证据筛选: 先由重排模型计算候选文档与问题的相关性得分并以阈值初筛, 再由 LLM 对剩余文档进行二分类相关性评估, 得到高置信证据集合 documents;

4) 生成检测与问题改写: 若 documents 为空, 则触发问题改写模块, 结合历史信息重写查询, 并回到检索阶段;

5) 回答生成与幻觉检测: 若 documents 非空, 则由法律回答模型基于 documents 与用户问题生成答案, 并由幻觉检测模块进行质量校验. 首先检测答案与证据是否存在冲突, 若冲突则记录错误案例并触发再生成; 若不冲突则进一步检测答案是否覆盖用户问题、是否存在“缺乏证据支撑的补全”现象, 只有通过检测的回答才能输出, 否则触发改写与重试. 若在限定迭代次数内仍无法满足一致性与覆盖度要求, 系统以“未查找到相关文档”安全退出. 与检索、重排及质量控制相关的技术细节在第 2.5 节进一步展开.

2.4.4 灾害记录智能体

面向系统内部结构化灾害记录的统计检索需求, 本文设计灾害记录智能体, 将 NL2SQL 与向量检索结合以提升中文场景下 SQL 生成的准确率与覆盖性, 其职责边界聚焦于系统内部结构化灾害记录的自然语言统计检索服务, 仅负责将用户自然语言查询转化为可执行 SQL 语句并返回结构化检索结果, 不承担非结构化数据检索任务. 该智能体将用户自然语言问题映射为 SQL 查询语句, 并在 SQLite 数据库中执行查询, 返回结构化检索结果或错误信息以支持迭代修正. 该智能体的关键提示词如图 7 所示.

```
你是一个可以调用工具的agent, 通过提供**自然语言查询, 避免使用sql语句查询**便可查询sql数据库.
请尽力回答用户问题, 如果可以不用调用工具就能回答问题, 请直接回答问题; 如果需要调用工具, 请调
用工具获取结果.
```

图 7 灾害记录智能体的关键提示词

为便于上层智能体统一调用, 本文将 NL2SQL 能力封装为独立 MCP 服务, 通过 Streamable HTTP 通信, 将“自然语言理解-SQL 生成-数据库查询-结果回传”标准化为可复用工具服务. 在 SQL 生成环节, 系统构建面向 NL2SQL 的检索增强向量库, 包含如下 3 类关键知识.

(1) ddl: 存储 5 个 SQLite 业务表的 DDL 语句, 用于明确表结构与字段.

(2) doc: 补充字段语义解释与相关领域知识, 帮助将自然语言概念对齐至字段与约束.

(3) sql: 提供约 100 条“自然语言-SQL”示例映射, 便于参考同类查询模板.

对任一自然语言问题, 系统首先通过向量检索召回相关的 DDL 片段、字段语义说明与示例 SQL, 并按预定规则与提示词拼接后输入 LLM 生成 SQL, 使模型在“字段定义+语义解释+同类示例”的约束下进行检索增强生成, 从而降低字段选择错误、条件遗漏与表关联不当等问题; 随后灾害记录智能体使用生成的 SQL 查询 SQLite 并返回结果, 必要时结合错误信息进入迭代修正流程. 该执行流程如图 8 所示.

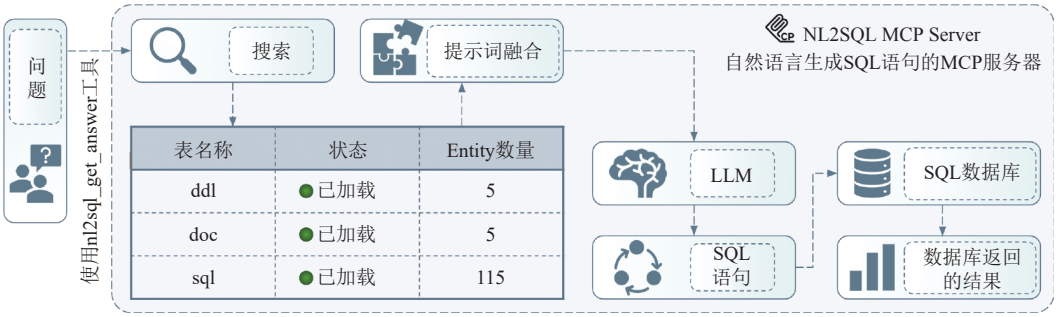


图 8 灾害记录智能体中 NL2SQL 并获取检索结果执行流程

实现上, 上述能力以工具形式注册在 NL2SQL MCP Server 中, 其中 nl2sql_get_answer 用于完成端到端 SQL 生成与查询返回, get_table_schema 用于按需获取表结构以辅助纠错与补全. 整体可概括为如下 3 个阶段.

(1) 判别阶段: 由 LLM 判断问题是否需要数据库查询; 可直接回答则生成, 否则进入 NL2SQL 流程.

(2) 交互阶段: LLM 与 NL2SQL MCP Server 进行多轮工具交互, 尽可能从 SQLite 中获取可支撑回答的事实证据.

(3) 生成阶段: 基于查询结果与用户问题组织输出, 并将结构化证据与结论回传上层智能体以完成对应子计划, 从而实现自然语言问答与关系数据库之间的稳定桥接.

2.5 检索增强与质量控制机制

检索增强与质量控制机制是保障地质灾害知识服务可信度与可追溯性的关键组成, 尤其在法规解读等高可靠场景中, 需要同时兼顾“召回速度”与“召回质量”, 并对生成回答施加一致性与有效性约束. 为降低重复计算开销并提升在线吞吐, 系统在嵌入计算前引入本地向量缓存: 若命中缓存则直接复用既有稀疏/密集向量; 否则调用嵌入模型生成向量并写入缓存库, 以减少重复向量化带来的时延. 同时, 检索客户端采用异步方式初始化并连接至 Milvus 向量数据库, 在单次检索中并行触发稀疏检索与密集检索两路服务, 并在后续进行融合排序, 从而在高并发条件下保持较高吞吐能力. 本系统采用的检索增强算法描述如算法 1 所示.

算法 1. 混合检索的算法描述.

输入: 原始文本 *Text*, 文档召回数量 *K*, 用户问题 *Query*;
输出: 根据用户要求召回的 *K* 个文档.

1. 在问题缓存中查找 Query, 如果查找到该问题, 则返回该问题的稀疏向量 $Sparse_{\text{Vector}}$ 和密集向量 $Dense_{\text{Vector}}$; 否则通过 BGE-M3 模型进行嵌入, 得到问题的稀疏向量 $Sparse_{\text{Vector}}$ 和密集向量 $Dense_{\text{Vector}}$ 并写入缓存.
2. 分别基于密集向量与稀疏向量执行密集检索与全文检索, 得到两组各 *K* 个召回文档.
3. 根据两组召回文档各自的排名和 *RRF* 融合函数对文档得分进行重新排序.
4. 输出融合排序后得分最高的 *K* 个文档.

2.5.1 混合检索架构

稀疏嵌入与密集嵌入分别强调词项匹配与语义相似, 稀疏嵌入的大多数维度为零, 其非零维度通常对应具体词汇单元, 具有较强的可解释性与关键词命中能力, 适合基于关键词的检索; 密集嵌入由连续实数组成, 能够刻画语义相似性与潜在关联, 更适合语义检索与近义表达匹配. 面向法律与专业知识检索, 本文采用稀疏检索与密集检索的混合策略, 以兼顾高召回与强语义匹配能力, 稀疏侧更易命中法规条文中的固定表述 (如法条编号、术语、机构名等); 密集侧能够覆盖语义等价与近义表达, 从而降低单一路径检索遗漏导致的证据缺失风险. 混合检索的基本流程如后文图 9 所示.

本文稀疏与密集嵌入均采用 BAAI 的 BGE-M3 作为向量嵌入模型, 该模型支持最长 8 192 个 tokens[49]. 稀疏向量与密集向量均存储于 Milvus 中, 密集向量索引采用 HNSW; 稀疏向量索引采用 SPARSE INVERTED. 检索执行后得到两组候选文档列表, 随后采用融合排序对两路结果统计重排. 具体地, 本文使用基于统计的 *RRF* (reciprocal rank fusion) 函数对不同检索路径的排名进行统一计分, 并输出融合后的候选文档集合. *RRF* 融合函数如下:

$$RRF_{\text{score}}(d) = \sum_{i=1}^{N} \frac{1}{k + rank_i(d)} \tag{2}$$

其中, N 代表不同检索路径的数量, $rank_i(d)$ 是第 i 个检索器检索到的文档 d 的排名位置, k 是平滑参数.

2.5.2 双筛选重排与 RAG 质量控制

在混合检索得到候选集后, 系统进一步采用“双阶段证据筛选+生成质量评估”的闭环机制, 以降低幻觉风险并提升回答的相关性、准确性与完整性.

(1) 双阶段证据筛选重排. 第 1 阶段重排使用 BAAI 的 BGE-Reranker-V2-M3 计算“问题-文档”匹配度, 并以阈值 0.75 过滤明显无关文档, 从而在早期清理噪声候选; 第 2 阶段筛选则调用 LLM 对剩余文档进行相关性二分类

评估, 得到高相关文档集合, 作为生成阶段的“干净上下文”.

(2) 生成后质量评估与闭环控制. 在生成答案后, 系统引入两类检测: 一是文档一致性检测, 判断回答是否与证据文档存在冲突; 若存在冲突, 则将错误案例写入上下文并触发再生成. 二是问题解决度检测, 判断回答是否覆盖用户问题、是否出现“未被证据支撑的补全”或偏题内容; 若未通过, 则触发问题重写机制, 结合历史检索结果与失败信息改写查询, 并重新进入检索流程.

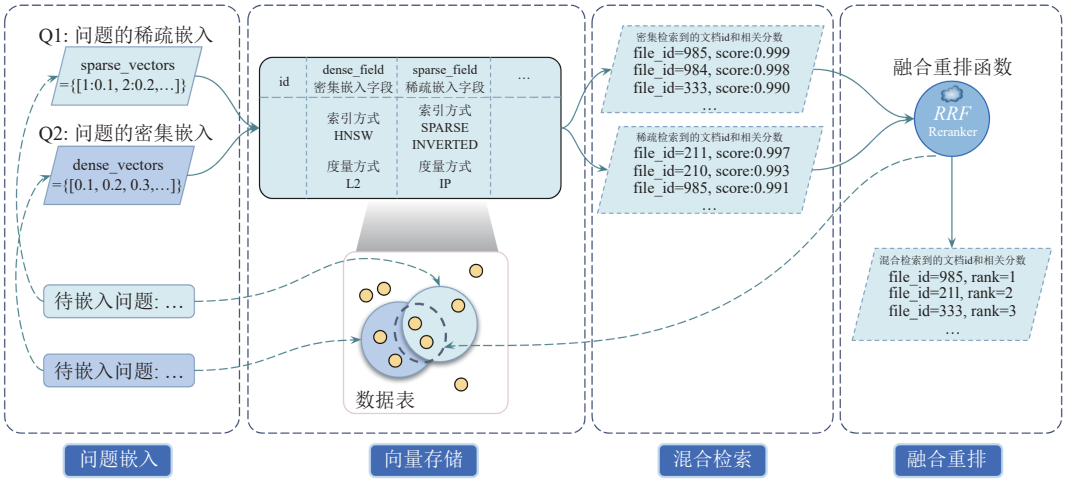


图 9 混合检索基本流程

上述过程在检查点模块约束下形成“检索-重排-检测-重写”的闭环控制: 当未检索到有效文档时, 优先改写问题并最多重试 3 次; 当检索到证据后, 对生成结果进行幻觉检测并在必要时进行二次生成, 其中回答生成最多尝试 2 次. 若仍无法满足一致性与覆盖度要求, 则以“未查找到相关文档”安全退出, 以显式约束知识来源与推理边界, 从机制上提升高可信场景下的输出可靠性. 关键超参数设置如表 2 所示.

表 2 混合检索关键超参数值

参数名称	参数值
Reranker阈值	0.75
回答迭代次数	2
重写迭代次数	3
*RRF*平滑系数	60
检索后返回结果数量K	20

3 实验分析

为全面评估本文所构建的面向地质灾害知识服务的多智能体协同推理框架, 本文基于自主构建的测试集, 从 3 个维度开展实验评测: (1) 指令微调领域模型的生成能力: 在灾害知识、灾害科普与法律法规这 3 类指令微调测试集上, 评估 LoRA 微调对生成质量的提升; (2) 多智能体框架的端到端答题可靠性: 构建科普专业、法律法规与灾害记录这 3 类选择题式知识问答数据集, 评估多智能体在任务分解、工具调用与证据整合后的单次作答正确率; (3) 多源异构数据整合能力的案例验证: 通过典型案例展示框架对外部实时信息、法规条文与结构化灾害记录等多源数据的融合、证据对齐与可解释输出能力.

3.1 实验设置

3.1.1 测试集与评价指标

本文共构建 6 个测试集: 前 3 个为指令微调测试数据集 (灾害知识、灾害科普、法律法规), 用于评估 LoRA

微调后模型的文本生成质量, 每个测试集包含 5000 条样本; 后 3 个为知识问答测试数据集 (科普专业、法律法规、灾害记录), 以选择题形式构建, 覆盖单选与不定项选择, 不定项选择题目共计 30 个, 每个测试集包含 150 条样本, 用于衡量多智能体框架在单次作答场景下的综合判题能力与稳定性. 科普专业数据属于简单数据集, 法律法规和灾害记录数据集由于涉及细节内容的查询, 属于相对困难的数据集. 与现有公开基准相比, 3 个知识问答测试集是深度聚焦于中国本土各省级法律法规问题、各省灾害记录结构化查询问题以及通用灾害知识的中文数据集.

针对选择题式知识问答任务, 本文采用准确率 Acc 作为评价指标, 其定义见公式 (3):

$$Acc = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{Y}_i = Y_i) \tag{3}$$

其中, N 表示题目总数, Y_i 表示第 i 题标准答案 (单选为唯一选项, 不定项为正确选项集合), $\hat{Y}_i$ 表示模型预测答案, $\mathbb{I}(\cdot)$ 为指示函数. 为强化可靠性约束, 对于不定项选择题, 预测集合需与标准集合完全一致 (不得多选、漏选或错选) 方可计为正确.

针对指令微调模型的生成质量评估, 本文采用 ROUGE-1、ROUGE-2、ROUGE-L 与 BLEU 等自动评价指标. ROUGE-1/2 基于 n-gram 重叠统计并计算 $F1$ 值[50], ROUGE-L 基于最长公共子序列度量顺序一致性[50], BLEU 基于语料级 n-gram 精确率并引入简短惩罚项 BP 抑制过短生成[51]. 由于本文设置中每个候选输出仅对应一个参考文本 (K=1), 故不涉及多参考答案下的聚合策略.

3.1.2 实现技术细节

实验在配备 8 张 NVIDIA L40S GPU 的服务器环境下完成, 操作系统为 Ubuntu 22.04. 模型微调阶段采用 LLaMA-Factory 实现 LoRA 参数高效微调; 向量检索与知识存储模块使用 Milvus 并以 Docker 容器化部署以保证环境一致性与可复现; 推理服务采用 vLLM 0.11.0 以提升吞吐与并发能力; 对话模板在 Qwen 默认对话模板基础上进行二次调整, 以更适配多工具调用场景下的指令组织与函数触发.

在多智能体配置方面, 问题完善代理与方案交付代理均以 Qwen2.5-7B-Instruct 作为底座模型; 应急科普代理使用微调后的 Emergency-Qwen, 领域知识代理使用微调后的 Knowledge-Qwen; 政策法规代理中, 生成式回答模型使用微调后的 Law-Qwen, 而最终汇总回答模型选用轻量级的 Qwen2-1.5B-Instruct, 以在保证效果的同时提升在线响应效率. 系统编排采用 LangGraph 以图结构组织任务分解、路由决策、工具执行与结果汇总, 并支持将工作流托管至 LangSmith 进行链路可视化、日志记录与评测分析, 从而实现对中间状态、工具调用结果与异常信息的细粒度监控与回溯.

3.2 系统整体性能对比

(1) 基准模型与对比设置

为评估本文多智能体协同推理框架在端到端知识问答任务中的有效性, 本文选取“开源指令模型 (不同参数规模)+联网检索”以及闭源强模型作为对比基线. 具体包括如下.

● Qwen2.5-7B-Instruct[52]: 7B 级开源指令模型, 具备较好的指令跟随与工程部署可行性, 用作轻量化基线以衡量在有限算力条件下的任务表现.

● Qwen2.5-32B-Instruct/Qwen2.5-72B-Instruct[52]: 更大参数规模的开源指令模型, 用于观察模型容量提升在本任务中的边际收益, 并作为“更强开源基线”参照.

● DeepSeek-V3.2[53]: 一个兼顾高计算效率、卓越推理能力和智能体性能的前沿开源模型. 在此用作当前顶尖开源能力基线, 以评估本文系统与最优开源基座模型相比的竞争力与效能提升.

● GPT-5: 闭源强模型基线, 用作通用能力“上界参照”, 以对比本文系统在更强基座条件下的相对差距与优势来源.

● MARSHA[47]: 一个面向自然灾害决策支持的通用多智能体检索增强生成框架. 为适配灾害数据集的自动化选择题评测, 对原架构进行了单次调用改造, 去除了依赖 UI 的多轮用户画像模块, 简化为“计划生成-工具调用检索-答案生成”的流水线. 该框架通过提示词引导模型根据已检索内容自主决定是否继续检索及检索方向, 实现最多 5 轮的

迭代检索. 本次实验将底座模型替换为本地部署的 Qwen2.5-7B-Instruct (负责计划与推理) 与 Qwen2-1.5B-Instruct (负责最终生成), 接入 BGE-M3 嵌入模型与 FAISS 向量库, 并将语料替换为中文灾害法规与历史事件数据. 在此用作多智能体 RAG 基线, 以评估传统的分阶段任务规划与工具调用方案与本文系统在复杂灾害问答中的效能差异.

● 传统的 Native RAG: 经典的单步检索增强生成范式. 系统遵循“问题向量化-向量库粗排-上下文拼接-大模型最终生成”的标准流程. 为保证实验公平对比, 除数据库配置外, 该基线与适配后的 MARSHA 保持完全一致的软硬件变量配置, 包括相同的中文灾害法规与数据语料库、BGE-M3 嵌入模型以及 Qwen2.5-7B-Instruct 底座大模型 (通过 Milvus 数据库进行向量存储和检索). 在此用作基础检索基线, 以验证单纯依赖稠密向量相似度检索 (无智能体规划与闭环检验) 在复杂灾害推理任务上的局限性, 从而凸显本文所提架构的必要性与优势收益. 考虑到不同模型的知识时效性与外部信息获取能力可能影响对比结果, 本文为 Qwen2.5 系列基线和 DeepSeek-V3.2 基线统一接入同一联网搜索工具 (表 3 中以“+web”标注, 且后文省略了“-Instruct”), 以减少因外部信息接入差异造成的偏差; 本文方法在多智能体框架下综合使用内部知识库、结构化数据库和必要的外部检索能力完成作答. Qwen2.5 系列模型、DeepSeek-V3.2 模型和本文方法使用秘塔搜索提供的 MCP 服务实现网络搜索, 该服务主要提供“metaso_web_search-网络搜索”和“metaso_web_reader-网页内容读取工具”等服务, GPT-5 则原生支持网络搜索工具.

表 3　本文方法和其他模型在不同数据集上的对比

模型框架	科普专业数据集	法律法规数据集	灾害记录数据集	完整数据集
Qwen2.5-7B+web	0.707	0.560	0.240	0.502
Qwen2.5-32B+web	0.707	0.493	0.053	0.418
Qwen2.5-72B+web	0.760	0.527	0.167	0.484
DeepSeek-V3.2+web	**0.800**	**0.727**	0.067	0.531
GPT-5	0.733	0.613	0.046	0.464
本文模型	0.707*	0.667*	**0.707***	**0.693***
Native RAG	0.133	0.127	0.120	0.127
MARSHA	0.440	0.160	0.580	0.393

注: 加粗数值代表在当前数据集 (列) 中取得的最高准确率, 带有“*”号的数值表示在本文设置的3种智能体框架中取得的最优指标

(2) 提示词约束

本文模型、MARSHA 和 Native RAG 的输出格式约束方式通过提示词规范, 如图 10 所示. Qwen2.5-7B、Qwen2.5-32B、Qwen2.5-72B 的输出格式约束方式通过提示词规范, 如图 11 所示. GPT-5 和 DeepSeek-V3.2 的输出格式约束方式通过提示词规范, 如图 12 所示.

(3) 实验结果及分析

表 3 给出了本文方法与各基线在 3 类选择题式知识问答数据集上的准确率对比. 总体来看, 本文方法在完整数据集上的综合准确率达到 0.693, 在所有对比方案中取得最高表现, 显著优于各类基线 (如表现最优的开源模型 DeepSeek-V3.2+web 为 0.531, 闭源大模型 GPT-5 为 0.464), 且大幅领先于同为智能体架构的 MARSHA (0.393) 与 Native RAG (0.127). 该结果表明, 相较于“单模型+联网检索”的直接范式、传统的基础 RAG 框架以及通用的灾害决策支持智能体框架, 本文框架通过任务分解、工具化证据获取与多阶段质量控制, 能够更稳定地提升端到端作答的可靠性与准确率, 进一步按数据集分析如下.

● 在灾害知识数据集上, DeepSeek-V3.2 取得最高准确率 0.800, 而本文方法为 0.707 (在 3 种智能体框架中表现最优). 这表明对于偏概念性、语义对齐与跨证据整合要求较高的题型, 更大容量模型在单轮决策中仍可能具有优势. 与此同时, 本文方法在该数据集上与多数组基线保持同等或更优水平, 说明本文提出的多智能体协作方法并不会显著损失通用灾害知识题的回答能力.

● 在法律法规数据集上, DeepSeek-V3.2+web 取得最高准确率 0.727, 本文方法紧随其后取得 0.667, 高于 GPT-5 (0.613) 及其他对比基线, 且在智能体框架中表现最优. 该类题目通常对措辞边界、例外条件和选项集合匹配高度敏感, 尤其不定项选择题对“漏选/多选”具有更强惩罚. 相较于单纯依赖通用联网检索的基线和几乎失效的 Native

RAG (0.127), 本文方法中的政策法规代理引入了专用法规检索、双阶段证据筛选与“检索-生成-校验”闭环机制. 这种设计使得证据相关性与结论一致性更为可控, 从而在面对法律问题时展现出较好表现. 在灾害记录数据集上, 本文方法以 0.707 取得最佳结果, MARSHA 以 0.580 位列第 2, 而各“LLM+web”基线与 Native RAG 整体表现出现“断崖式”下跌 (如 DeepSeek-V3.2+web 仅为 0.067, GPT-5 为 0.046, Native RAG 为 0.120). 该结果体现出以下两点现象.

```
# 角色定位
你是一个极其严谨的**基于证据的选择题解答专家**, 你的核心能力是**精准锚定问题意图**, 并从提供的“已知内容”和“题目”中提取**唯一对应的证据**进行解答.

# 输出格式
一步一步思考分析 (分析“已知内容”, 并重点关注一些限制条件 (如地区、时间、地震震级等)、注意条件以及逻辑关系, 然后回答“题目”)
...
最终的latex格式代码, 严格按照如下要求:
1. 选项要放在[]中, 选项**只能是大写字符**, 且禁止用其他latex语法修饰, 多个选项需要用逗号隔开
2. []被\\answer{{}}包围:
3. 其回答形式如下例所示:
      * 单选: \\answer{[A]}
      * 多选: \\answer{[A,C]} (按字母顺序排列)

## 错误案例:
...

# 示例
...
1. **已知内容分析**:
...

2. **题目分析**:
...
最终答案:
* 单选:
```latex
\\answer{[A]}
```
任务
以下内容是已知内容以及具体的题目和选项
已知内容: {}\n\n
题目: {}\n\n

请你根据这些信息一步一步分析思考, 根据题目给出你的回答:
```

图 10　本文模型、MARSHA 和 Native RAG 的关键提示词

(a) 与“LLM+web”的直接范式相比, 本文方法在数值型、统计型与要素型信息的处理上优势更明显. 原因在于灾害记录题往往要求对地区、时间、指标口径与数值进行严格对齐, 开放网页检索容易引入噪声 (如口径不一致、单位差异或上下文缺失), 从而使错误在生成与决策过程中累积; 而本文框架通过灾害记录代理将问题约束映射为结构化查询 (NL2SQL), 以数据库结果提供可追溯的事实证据, 并在此基础上完成更稳定的证据融合, 因此获得显著更高的准确率. 值得注意的是, 同为包含工具调用机制的 MARSHA 在此数据集上亦取得了 0.580 的较好分数, 这验证了通过提示词引导模型进行多轮迭代检索策略的有效性.

(b) Qwen2.5-7B+web 在该数据集上高于 Qwen2.5-32B/72B+web、GPT-5 与 DeepSeek-V3.2+web. 对于该现象, 本文推测可能与模型在证据不足情境下的决策偏好有关, 当检索证据无法与选项形成清晰对齐时, 能力更强的模型可能倾向于输出“不确定/无法判断”或给出与选项不完全匹配的开放式结论, 从而在严格的选择题判分规则下被计为错误; 而另一些模型可能更倾向于在不充分证据下仍选择具体选项, 导致偶然命中概率上升. 该推测有待结合第 3.6 节的回答分布与错误类型统计进一步验证. 总体而言, 这一结果也提示了联网检索与检索增强并不必然消除幻觉, 若证据稀疏、口径不一致或证据-结论对齐不足, 仍可能出现“带检索的幻觉”, 即在证据不足时给出看似合理但难以被证据严格支撑的选项判断.

```
# 功能
可以调用网络搜索工具, 但如果用户的问题你可以直接回答, 请直接回答; 如果用户的问题无法直接回答, 请进行网络搜索, 并返回搜索结果...
# 角色定位
你是一个**选择题解答专家**.
# 输出格式
一步一步思考分析 (分析“已知内容”，并重点关注一些限制条件 (如地区、时间、地震震级等)、注意条件以及逻辑关系, 然后回答“题目”)
...
最终的latex格式代码, 严格按照如下要求:
1. 选项要放在[]中, 选项**只能是大写字符**, 且禁止用其他latex语法修饰, 多个选项需要用逗号隔开
2. []被\\answer{{}}包围:
3. 其回答形式如下例所示:
    * 单选: \\answer{[A]}
    * 多选: \\answer{[A,C]} (按字母顺序排列)
4. 如果无法确定答案 (例如没有提供表数据或相关信息), **禁止假设数据**, **请使用[网络搜索工具]尽可能搜索出答案**, 如果仍然无法确定, 则回答\\answer{[]}
## 错误案例:
...
# 正确示例
...
1. **已知内容分析**:
...
2. **题目分析**:
...
最终答案:
* 单选:
```latex
\\answer{[A]}
```
任务
题目: {}\n\n

请你根据这些信息一步一步分析思考, 根据题目给出你的回答:
```

图 11　Qwen2.5-7B、Qwen2.5-32B 和 Qwen2.5-72B 的关键提示词

```
# 功能
如果用户的问题你可以直接回答, 请直接回答; 如果用户的问题无法直接回答, 请检查是否有给你配置网络搜索相关的工具, 如果有这样的工具, 请使用网络搜索工具, 进行网络搜索, 并返回搜索结果...使用**所有[工具]的总次数禁止超过2次**, 超过两次后仍然无法得到确切答案则回答\\answer{[]}
# 角色定位
你是一个**选择题解答专家**.
# 输出格式
分析题目, 并重点关注一些限制条件 (如地区、时间、地震震级等)、注意条件以及逻辑关系, 然后回答“题目”...
最终的latex格式代码, 严格按照如下要求:
1. 选项要放在[]中, 选项**只能是大写字符**, 且禁止用其他latex语法修饰, 多个选项需要用逗号隔开
2. []被\\answer{{}}包围:
3. 其回答形式如下例所示:
    * 单选: \\answer{[A]}
    * 多选: \\answer{[A,C]} (按字母顺序排列)
4. 如果无法确定答案 (例如没有提供表数据或相关信息):
    – **禁止假设数据**
    – 如果可以调用网络工具, **请使用[网络搜索工具]尽可能搜索出答案**
    – 如果**仍然无法确定**或者**没有给你配置网络工具**, 则**必须回答\\answer{[]}**
5. 无论如何都必须按照\\answer{[]}的格式输出, 否则无法得分
## 示例
…
# 任务
题目: {}\n\n
-------------------------------------------------------------------------------
请你根据问题结合网络搜索工具, 联系题目给出你的回答:
```

图 12　GPT-5 和 DeepSeek-V3.2 的关键提示词

3.3 消融实验

为评估各功能模块对系统端到端性能的贡献, 本文在完整框架基础上分别对政策法规模块、科普+专业问答模块、网络搜索模块与灾害记录模块进行消融, 并在 3 类选择题式知识问答数据集上比较性能差异. 表 4 给出了消融实验结果. 总体而言, 完整框架在“完整数据集”上的正确率达到 0.693, 优于任意消融设置, 表明多模块协同能够为整体性能提供稳定增益; 同时, 不同消融在不同数据集上的下降幅度存在显著差异, 说明各模块对不同任务类型具有明确的专门贡献.

表 4 模型框架消融实验

数据集	消融政策法规	消融科普+专业问答	消融网络搜索	消融灾害记录	完整框架
科普专业数据集	0.693	0.607	0.660	**0.713**	0.707
政策法规数据集	0.447	0.647	0.653	0.660	**0.667**
灾害记录数据集	0.700	0.667	0.693	0.427	**0.707**
完整数据集	0.613	0.640	0.669	0.600	**0.693**

(1) 消融政策法规模块. 消融后, “政策法规数据集”正确率由 0.667 降至 0.447, 下降 22.0 个百分点; “完整数据集”由 0.693 降至 0.613, 下降 8.0 个百分点. 该结果表明, 政策法规类题目对条文定位、边界条件与例外规则处理高度敏感, 政策法规模块在证据检索与结论一致性约束方面提供了关键支撑, 并显著抬升系统整体性能下限.

(2) 消融科普+专业问答模块. 消融后, “科普专业数据集”正确率由 0.707 降至 0.607, 下降约 10.0 个百分点; “完整数据集”由 0.693 降至 0.640, 下降 5.3 个百分点. 结果验证了科普与专业问答能力在灾害知识类题目中的核心作用: 该模块不仅增强概念解释与专业知识对齐, 也有助于提升系统整体稳定性.

(3) 消融灾害记录模块. 消融后, “灾害记录数据集”正确率由 0.707 大幅降至 0.427, 下降 28.0 个百分点; “完整数据集”由 0.693 降至 0.600, 下降 9.3 个百分点. 说明灾害记录题强依赖数值、时间、地点与指标等结构化事实的精确检索与口径对齐, 灾害记录模块 (NL2SQL 结构化查询链路) 对该类任务具有不可替代的支撑作用, 并对整体性能贡献显著.

(4) 消融网络搜索模块. 消融后, 各数据集均出现小幅下降: “科普专业数据集”由 0.707 降至 0.660 (下降 4.7 个百分点); “政策法规数据集”由 0.667 降至 0.653 (下降 1.4 个百分点); “灾害记录数据集”由 0.707 降至 0.693 (下降 1.4 个百分点); “完整数据集”由 0.693 降至 0.669 (下降 2.4 个百分点). 结果表明, 在当前评测任务分布下, 网络搜索模块提供了边际增益. 该模块并非决定性组件, 但在开放域补充、长尾实体覆盖与多源交叉验证方面能稳定提升整体表现.

3.4 公开数据集上的基准测试

DisasterQA[54]是一个开源的灾害响应选择题基准数据集, 其数据整合自 6 个在线数据源, 主题覆盖广泛. 为验证本文模型在训练数据分布外的泛化能力, 我们在 DisasterQA 基准上进行了对比实验, 评估各模型框架的准确率. 具体而言, 为保证测试样本的均衡性, 本文从中随机抽取了 100 道测试题 (正确答案为 A、B、C、D 的题目各 25 道). 后文表 5 展示了各模型框架在公开数据集上测试的准确率统计结果. 其中, 加粗数值代表在当前数据集 (列) 中取得的最高准确率. 实验结果表明, 在引入外部知识库与工具调用的特定框架对比中, 本文提出的多智能体协同推理框架以 0.670 的准确率显著优于传统无闭环验证的 Native RAG (0.170) 以及通用的灾害应对多智能体框架 MARSHA (0.360). 这一断层式的领先表明, 即使脱离了特定测试集的数据分布, 本文架构的任务拆解与检索闭环机制依然能够有效发挥作用, 极大程度避免了传统基础检索范式中常见的“无关信息干扰”现象. 此外, 本文模型的性能成功超越了同等参数规模的“单模型+联网检索”基线 Qwen2.5-7B+web (0.620). 综合来看, 这一实验结果有力证明了本文模型在跨数据分布下依然具备较好的泛化能力与鲁棒性.

3.5 微调模型效果测试

本文以 Qwen2.5-7B-Instruct 作为预训练基座模型, 采用 LoRA 技术在 3 类指令数据上分别进行领域适配, 得到 3 个微调模型: Emergency-Qwen、Knowledge-Qwen 与 Law-Qwen. 其中, Emergency-Qwen 对应灾害科普指令数据, Knowledge-Qwen 对应灾害知识指令数据, Law-Qwen 对应法律法规指令数据. 为评估微调对生成质量的提

升, 本文在对应的 3 类指令微调测试集上对比“LoRA 微调后的模型 (Emergency-Qwen、Knowledge-Qwen 和 Law-Qwen)”与“未微调基座模型 (Qwen2.5-7B-Instruct)”的自动评价指标得分, 结果如图 13 所示. 图 13(a)–(d) 分别给出了 BLEU-4、ROUGE-1、ROUGE-2、ROUGE-L 的评测结果, 其中 ROUGE-1/2/L 均以 $F1$ 分数统计. 为便于展示与对比, 本文将所有指标×100, 以百分制形式呈现. 总体来看, LoRA 微调在 3 类任务上均带来一致且显著的性能提升. 以 BLEU (×100) 为例, 微调后的 Emergency-Qwen、Knowledge-Qwen 和 Law-Qwen 在灾害科普、灾害知识与法律法规测试集上的得分分别达到 29.04、28.91、40.27, 显著高于基座模型在相同测试集上的 6.29、0.46、2.62, 表明微调后模型在高阶 n-gram 层面的匹配度得到明显增强.

表 5　各个模型框架在 DisasterQA 公开数据集上的基准测试

模型框架	准确率
Qwen2.5-7B+web	0.620
本文模型	**0.670**
Native RAG	0.170
MARSHA	0.360

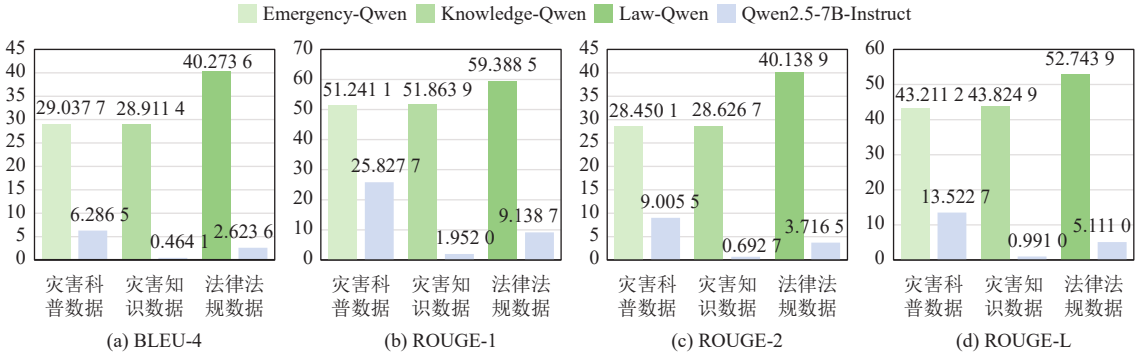


图 13　微调前后模型在对应数据集上的 BLEU-4 与 ROUGE 得分

此外, 相比未微调的基座模型, LoRA 微调后的模型在 ROUGE-1、ROUGE-2、ROUGE-L 的 $F1$ (×100) 上同样呈现稳定提升. 这说明微调不仅提升了生成内容与参考答案在词汇层面的重叠 (ROUGE-1), 也提升了在二元组层面的重叠 (ROUGE-2), 并提升了生成内容与参考答案在更长跨度上的序列一致性 (ROUGE-L), 从而整体改善了模型在对应领域指令下的生成质量与内容对齐程度. 综合来看, 基于 LoRA 指令微调的 Qwen2.5-7B-Instruct 能够在灾害科普、灾害知识与法律法规这 3 个领域上稳定提升 BLEU 与 ROUGE 等指标, 验证了本文微调方案的有效性.

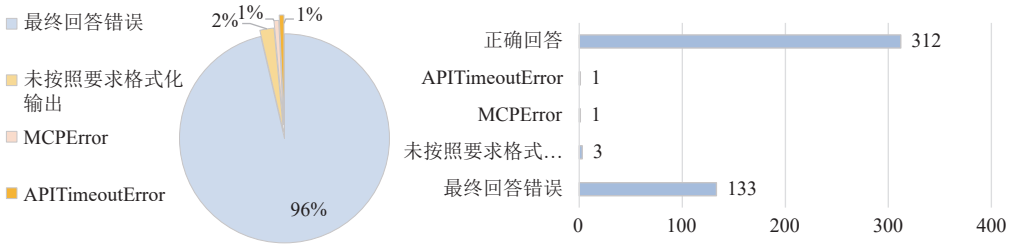
3.6　错误模式分析

如图 14 所示, 本文方法在 3 类选择题式问答数据上的评测共包含 450 个样本, 其中 312 个样本回答正确, 其余 138 个样本为错误或异常输出. 对错误样本进一步归因可见, 错误主要集中在“最终回答错误”类别: 共 133 个, 占全部错误的 96.38%. 该类错误通常表现为模型在证据理解与推理过程中出现偏差, 导致最终结论与标准答案不一致, 具体包括关键信息遗漏、推理链路不完整, 以及证据引用与题目约束或选项要求不匹配等情况.

除上述原因外, 约 2.17% 的错误来自“未按要求格式化输出”, 即模型在多次格式约束提示后仍未严格遵循预定义的输出结构, 从而在严格判分规则下被计为错误. 工具异常属于低频偶发问题: MCPError 与 APITimeoutError 各出现 1 次. 其中, MCPError 报错为“Tool execution failed: 系统内部错误”, 经排查来源于网络搜索服务的内部异常; APITimeoutError 则由向量嵌入服务触发, 原因是该次请求超过嵌入模型支持的最大输入长度 (8 192 个 tokens), 导致嵌入服务未能正常返回结果并最终超时.

总体来看, 工具异常调用并不是本文方法的主要误差来源. 本文主要误差集中于最终答案生成阶段的语义理解与推理偏差. 后续可重点从以下方面优化: 一是增强证据检索质量与证据-选项对齐机制, 减少“证据充分但对齐

失败”的错误; 二是提升推理过程的可控性与中间结论一致性约束, 降低推理链路断裂与遗漏; 三是加强输出结构化约束与格式鲁棒性, 减少格式错误; 四是在条件允许时提升基座模型能力或引入更强的验证模块, 以进一步降低最终决策偏差.



(a) 错误类型分布 (b) 回答结果分布

图 14 错误类型分布和回答结果分布

3.7 输出案例分析

本节通过典型案例展示本文多智能体协同推理框架在复杂问题下对多源异构数据的整合能力, 以及“规划-执行-校验-汇总”的闭环工作流程. 图 15 给出了代理级执行轨迹的成功案例. 整体而言, 系统的执行过程可归纳为 3 个阶段: 任务拆解与计划生成、专业代理协同执行与结果回写、多源证据融合与输出.

(a) 阶段 1 (b) 阶段 2 (c) 阶段 3

图 15 代理级成功案例分析

(1) 阶段 1: 任务拆解与计划生成. 如图 15(a) 所示, 用户提出问题后, 意图分析代理首先识别核心诉求与关键约束, 生成任务清单 plan.md, 并明确子任务的执行顺序. 本案例中, 系统将问题拆分为两条主线: 首先检索震情实点数据, 其次补充相应法律依据. 随后, 监管者依据第 1 项子任务进行调度, 将“震情实点数据检索”路由至灾害记录代理. 灾害记录代理访问 SQLite 数据库返回关键震情参数与记录片段, 监管者据此确认子任务完成并输出完成标记 (finish).

(2) 阶段 2: 专业代理协同执行与结果回写. 如图 15(b) 所示, 方案交付代理对监管者与灾害记录代理的交互结果进行上下文压缩, 将可用于支撑后续回答的关键证据写入 information 字段, 并基于证据完整性与任务覆盖情况给出验收判断 (is_ok). 在确认第 1 项子任务完成后, 系统自动推进至下一项子任务: 监管者调度政策法规代理, 围

绕已确认的地震事件检索应急管理相关法律法规与条文依据. 政策法规代理返回对应规范性条款, 为后续政策建议与处置方案提供依据; 监管者同样对该子任务完成确认并输出 finish 标记.

(3) 阶段 3: 多源证据融合与输出. 如图 15(c) 所示, 方案交付代理对政策法规代理输出进行进一步压缩归纳, 并完成对 2 项子任务的验收回写. 当 plan.md 中所有子任务均被标记为完成后, 系统进入最终汇总阶段: 回答代理读取 plan.md 中的震情事实片段与法律条文依据, 并与用户原始问题进行一致性对齐与结构化整合, 最终生成一份逻辑连贯、证据链完整且可直接用于业务决策与沟通的综合汇报输出.

图 16 展示了代理级执行轨迹中的一个典型失败案例. 该案例聚焦于系统的任务规划阶段. 具体而言, 当处理用户关于“2022 年甘肃省地震事件”的查询时, 规划智能体未能严格按照系统提示词的约束, 生成规定格式的 JSON 执行计划. 尽管框架内置了闭环校验机制, 能够捕获解析失败的异常, 并将具体报错信息作为反馈提示输入给智能体引导其进行反思与修正 (如图 16 中的 plan 校验 I、II、III 所示), 但智能体在连续 3 次的迭代重试中, 始终未能输出完全符合规定格式规范的计划. 最终, 在达到系统的最大迭代阈值后, 流程被迫中断并抛出“make plan failed”异常, 导致后续的专家调用与信息检索任务无法顺利进行.

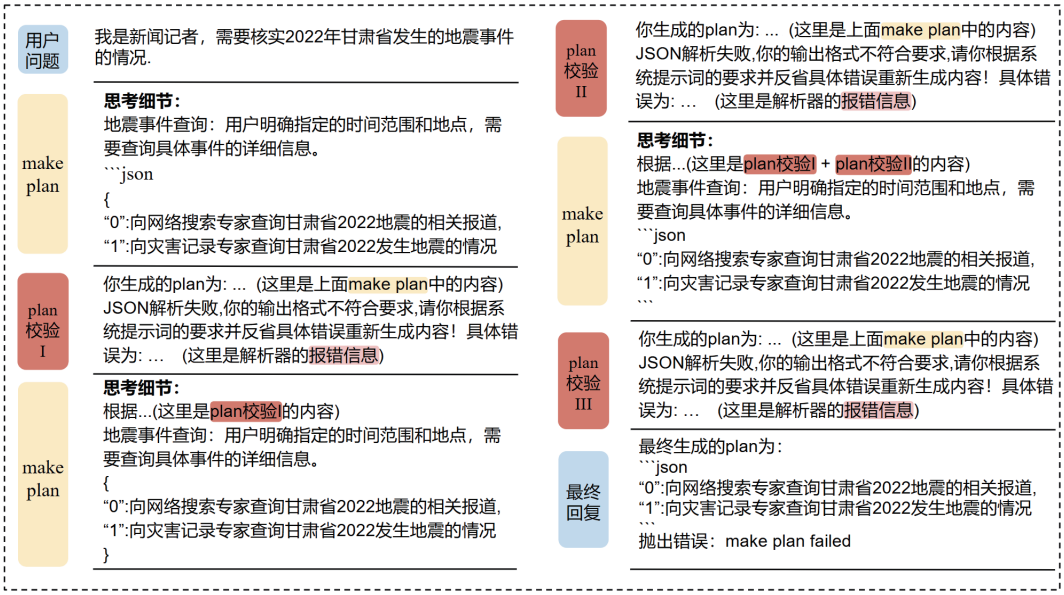


图 16 代理级失败案例分析

4 总 结

本文面向地质灾害知识服务场景, 提出一种基于大语言模型的多智能体协同推理框架, 并在此基础上实现可落地的智能问答系统. 框架以 Plan-and-Execute 闭环流程为主线, 通过意图分析、任务分解、监管调度与方案交付等机制, 将复杂问题转化为可解释的任务计划并实现多代理协同执行. 在数据层面, 系统融合关系数据库、向量知识库与外部实时检索, 联合 NL2SQL 和混合检索以覆盖结构化知识与文本型知识. 在质量控制层面, 政策法规代理引入相关性筛选、重排与幻觉检测等环节, 形成“检索-生成-校验”的多阶段链路. 实验结果表明, 该系统在灾害知识、法律法规与灾害记录等测试集上的整体正确率达到 0.693, 较多种 LLM 基线方法及更大参数模型均取得更优表现; 消融实验进一步验证了政策法规代理、应急科普代理、领域知识代理、网络搜索代理与灾害记录代理在不同类型任务上的关键贡献. 从更宏观的理论维度来看, 本文探索并确立了一种面向复杂垂直领域的“高可信知识服务范式”. 该范式通过大模型泛化理解与多源事实依据的深度耦合, 配合多智能体协同的闭环校验, 有效消减了生成内容的幻觉风险, 为构建高可靠、强可解释的知识问答体系提供了完整的方法论支撑与理论闭环.

除此之外, 本文提出的基于大语言模型的多智能体协同推理框架可以扩展至其他领域.

1) 迁移的前提条件: 目标领域必须属于“知识密集型”场景, 即拥有大量沉淀的专业数据, 且用户的查询诉求依赖于对这些事实依据的检索与推理. 同时, 目标任务需具备一定的可拆解性, 以便于多智能体进行分工协作.

2) 数据构建成本: 这是迁移过程中的主要成本. 需要针对新领域收集并清洗大量的专业领域数据 (结构化与非结构化), 并投入相应的工程资源构建用于语义查询的向量数据库以及用于结构化查询的关系数据库.

3) 模型训练与微调成本: 尽管框架的调度机制是通用的, 但为了确保在特定专业领域 (如医疗诊断) 的回答具备足够的准确性与专业度, 通常需要收集领域指令数据对底座大语言模型进行微调, 以使模型掌握该领域的特有术语和推理逻辑.

4) 智能体适配成本: 开发者需要根据新领域的业务逻辑, 重新设计并测试各个领域专家 agent 的系统提示词, 并可能需要调整监管者代理的调度策略.

未来工作将从以下 4 个方面推进: 其一, 优化多智能体路由与协作策略, 引入更细粒度的自动评测与反馈信号, 在保持效果的同时降低推理与工具调用开销; 其二, 持续扩充并动态更新多源知识底座, 提升对长尾实体、跨区域灾害事件与法规差异的覆盖能力, 并探索多模态数据接入以增强场景理解; 其三, 针对当前错误主要集中于最终推理偏差的问题, 强化证据对齐与一致性约束, 提升输出格式的鲁棒性与稳定性; 其四, 结合真实用户交互数据开展在线评测与迭代优化, 提升系统在应急管理、风险防范与公众科普等场景中的实用性与可信度.

References

[1] Wang Z, Lu JR, Yang DL, Li MX. Research on intelligent question-answering system for flood emergency decision-making with fusion of GPT and knowledge graph. Journal of Safety Science and Technology, 2024, 20(4): 5–11 (in Chinese with English abstract). [doi: 10.11731/j.issn.1673-193x.2024.04.001]

[2] Sang JT, Yu J. ChatGPT: A glimpse into AI’s future. Journal of Computer Research and Development, 2023, 60(6): 1191–1201 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.202330304]

[3] Lei ZY, Dong YS, Li WY, Ding R, Wang QR, Li JD. Harnessing large language models for disaster management: A survey. In: Proc. of the Association for Computational Linguistics (ACL 2025). Vienna: ACL, 2025. 14528–14551. [doi: 10.18653/v1/2025.findings-acl.750]

[4] Li GH, Hammoud AAK, Itani H, Khizbullin D, Ghanem B. CAMEL: Communicative agents for “mind” exploration of large language model society. In: Proc. of the 37th Int’l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 2264.

[5] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proc. of the 34th Int’l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 793.

[6] Yao SY, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, Narasimhan K. Tree of thoughts: Deliberate problem solving with large language models. In: Proc. of the 37th Int’l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 517.

[7] Song JT, Wang XG, Zhu J, Wu YH, Cheng XX, Zhong R, Niu C. RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing: Industry Track. Miami: ACL, 2024. 1548–1558. [doi: 10.18653/v1/2024.emnlp-industry.113]

[8] Asai A, Wu ZQ, Wang YZ, Sil A, Hajishirzi H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In: Proc. of the 12th Int’l Conf. on Learning Representations. Vienna: OpenReview.net, 2024.

[9] Sun J, Li GL, Pan J, Wang J, Xie YQ, Liu RC, Nie W. GaussDB-Vector: A large-scale persistent real-time vector database for LLM applications. Proc. of the VLDB Endowment, 2025, 18(12): 4951–4963. [doi: 10.14778/3750601.3750619]

[10] Wu QY, Bansal G, Zhang JY, Wu YR, Li BB, Zhu EK, Jiang L, Zhang XY, Zhang SK, Liu JL, Awadallah AH, White RW, Burger D, Wang C. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In: Proc. of the 1st Conf. on Language Modeling. Philadelphia, 2024.

[11] Barnett S, Kurniawan S, Thudumu S, Brannelly Z, Abdelrazek M. Seven failure points when engineering a retrieval augmented generation system. In: Proc. of the 3rd IEEE/ACM Int’l Conf. on AI Engineering-software Engineering for AI. Lisbon: IEEE, 2024. 194–199.

[12] Shen YL, Song KT, Tan X, Li DS, Lu WM, Zhuang YT. HuggingGPT: Solving AI tasks with chatgpt and its friends in Hugging Face. In: Proc. of the 37th Int’l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 1657.

[13] Zhang CY, Yang KJ, Hu SY, Wang ZH, Li GH, Sun YH, Zhang C, Zhang ZW, Liu AJ, Zhu SC, Chang XJ, Zhang JG, Yin F, Liang YT, Yang YD. ProAgent: Building proactive cooperative agents with large language models. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI, 2024. 17591–17599. [doi: 10.1609/aaai.v38i16.29710]

[14] Yue SB, Wang SY, Chen W, Huang XJ, Wei ZY. Synergistic multi-agent framework with trajectory learning for knowledge-intensive tasks. In: Proc. of the 39th AAAI Conf. on Artificial Intelligence. Philadelphia: AAAI, 2025. 25796–25804. [doi: 10.1609/aaai.v39i24.34772]

[15] Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: Interactive simulacra of human behavior. In: Proc. of the 36th Annual ACM Symp. on User Interface Software and Technology. San Francisco: ACM, 2023. 2. [doi: 10.1145/3586183.3606763]

[16] Ni B, Buehler MJ. MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge. Extreme Mechanics Letters, 2024, 67: 102131. [doi: 10.1016/j.eml.2024.102131]

[17] Pandey AK, Kushwaha VK, Gupta S, Sharan B. Large language model-enabled multi-agent manufacturing systems. In: Proc. of the 2025 Modern Electronics Devices and Intelligent Communication Systems. Greater Noida: IEEE, 2025. 333–339. [doi: 10.1109/MEDCOM67532.2025.11405392]

[18] Shen WZ, Li CL, Chen HZ, Yan M, Quan XJ, Chen HH, Zhang J, Huang F. Small LLMs are weak tool learners: A multi-LLM agent. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing. Miami: ACL, 2024. 16658–16680. [doi: 10.18653/v1/2024.emnlp-main.929]

[19] Li H, Xu TL, Chang E, Wen QS. Knowledge tagging with large language model based multi-agent system. In: Proc. of the 39th AAAI Conf. on Artificial Intelligence. Philadelphia: AAAI, 2025. 28775–28782. [doi: 10.1609/aaai.v39i28.35141]

[20] Zhu A, Dugan L, Callison-Burch C. ReDel: A toolkit for LLM-powered recursive multi-agent systems. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations. Miami: ACL, 2024. 162–171. [doi: 10.18653/v1/2024.emnlp-demo.17]

[21] Xu L, Hu ZY, Zhou DQ, Ren HY, Dong Z, Keutzer K, Ng SK, Feng JS. MAgIC: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing. Miami: ACL, 2024. 7315–7332. [doi: 10.18653/v1/2024.emnlp-main.416]

[22] Liang T, He ZW, Jiao WX, Wang X, Wang Y, Wang R, Yang YJ, Shi SM, Tu ZP. Encouraging divergent thinking in large language models through multi-agent debate. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing. Miami: ACL, 2024. 17889–17904. [doi: 10.18653/v1/2024.emnlp-main.992]

[23] Choi HK, Zhu J, Li S. Debate or vote: Which yields better decisions in multi-agent large language models? In: Proc. of the 39th Annual Conf. on Neural Information Processing Systems. San Diego: Curran Associates Inc., 2025.

[24] Park C, Han SJ, Guo XZ, Ozdaglar AE, Zhang KQ, Kim JK. MAPoRL: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In: Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vienna: ACL, 2025. 30215–30248. [doi: 10.18653/v1/2025.acl-long.1459]

[25] Ge YQ, Hua WY, Mei K, Ji JC, Tan JT, Xu SY, Li ZL, Zhang YF. OpenAGI: When LLM meets domain experts. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 242.

[26] Guo TC, Chen XY, Wang YQ, Chang RD, Pei SC, Chawla NV, Wiest O, Zhang XL. Large language model based multi-agents: A survey of progress and challenges. In: Proc. of the 33rd Int'l Joint Conf. on Artificial Intelligence (IJCAI 2024). Jeju: IJCAI, 2024. 890. [doi: 10.24963/ijcai.2024/890]

[27] Cemri M, Pan MZ, Yang SY, Agrawal LA, Chopra B, Tiwari R, Keutzer K, Parameswaran A, Klein D, Ramchandran K, Zaharia M, Gonzalez JE, Stoica I. Why do multi-agent LLM systems fail? In: Proc. of the 39th Annual Conf. on Neural Information Processing Systems. San Diego: Curran Associates Inc., 2025.

[28] Erdogan LE, Lee N, Kim SH, Moon SH, Furuta H, Anumanchipalli G, Keutzer K, Gholami A. Plan-and-Act: Improving planning of agents for long-horizon tasks. In: Proc. of the 42nd Int'l Conf. on Machine Learning. Vancouver: PMLR, 2025.

[29] Long JY. Large language model guided tree-of-thought. arXiv:2305.08291, 2023.

[30] Schick T, Dwivedi-Yu J, Dessí R, Raileanu R, Lomeli M, Hambro E, Zettlemoyer L, Cancedda N, Scialom T. Toolformer: Language models can teach themselves to use tools. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 2997.

[31] Surís D, Menon S, Vondrick C. ViperGPT: Visual inference via Python execution for reasoning. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 11888–11898. [doi: 10.1109/ICCV51070.2023.01092]

[32] Yao SY, Zhao J, Yu D, Du N, Shafran I, Narasimhan KR, Cao Y. ReAct: Synergizing reasoning and acting in language models. In: Proc.

of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023.

[33] Wang GZ, Xie YQ, Jiang YF, Mandlekar A, Xiao CW, Zhu YK, Fan LX, Anandkumar A. Voyager: An open-ended embodied agent with large language models. Trans. on Machine Learning Research, 2024, 2024.

[34] Zhu A, Dugan L, Hwang A, Callison-Burch CB. Kani: A lightweight and highly hackable framework for building language model applications. In: Proc. of the 3rd Workshop for Natural Language Processing Open Source Software. Singapore: ACL, 2023. 65–77. [doi: 10.18653/v1/2023.nlposs-1.8]

[35] Zhu A, Hwang A, Dugan L, Callison-Burch CB. FanOutQA: A multi-hop, multi-document question answering benchmark for large language models. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers). Bangkok: ACL, 2024. 18–37. [doi: 10.18653/v1/2024.acl-short.2]

[36] Luo ZY, Shen ZQ, Yang WZ, Zhao ZR, Jwalapuram P, Saha A, Sahoo D, Savarese S, Xiong CM, Li JN. MCP-Universe: Benchmarking large language models with real-world model context protocol servers. In: Proc. of the 14th Int'l Conf. on Learning Representations. Rio de Janeiro: OpenReview.net, 2026.

[37] Jiang ZB, Xu F, Gao LY, Sun ZQ, Liu Q, Dwivedi-Yu J, Yang YM, Callan J, Neubig G. Active retrieval augmented generation. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 7969–7992. [doi: 10.18653/v1/2023.emnlp-main.495]

[38] Madaan A, Tandon N, Gupta P, Hallinan S, Gao LY, Wiegreffe S, Alon U, Dziri N, Prabhumoye S, Yang YM, Gupta S, Majumder BP, Hermann K, Welleck S, Yazdanbakhsh A, Clark P. SELF-REFINE: Iterative refinement with self-feedback. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 2019.

[39] Wang JG, Yi XM, Guo RT, Jin H, Xu P, Li SJ, Wang XY, Guo XZ, Li CM, Xu XH, Yu K, Yuan YX, Zou YH, Long JQ, Cai YD, Li ZX, Zhang ZF, Mo YH, Gu J, Jiang RY, Wei Y, Xie C. Milvus: A purpose-built vector data management system. In: Proc. of the 2021 Int'l Conf. on Management of Data. ACM, 2021. 2614–2627. [doi: 10.1145/3448016.3457550]

[40] Zhang HC, Cao RS, Chen L, Xu HS, Yu K. ACT-SQL: In-context learning for Text-to-SQL with automatically-generated chain-of-thought. In: Proc. of the 2023 Association for Computational Linguistics (EMNLP 2023). Singapore: ACL, 2023. 3501–3532. [doi: 10.18653/v1/2023.findings-emnlp.227]

[41] Fu YW, Ou WJ, Yu Z, Lin Y. MIGA: A unified multi-task generation framework for conversational Text-to-SQL. In: Proc. of the 37th AAAI Conf. on Artificial Intelligence. Washington: AAAI, 2023. 12790–12798. [doi: 10.1609/aaai.v37i11.26504]

[42] Gao DW, Wang HB, Li YL, Sun XY, Qian YC, Ding BL, Zhou JR. Text-to-SQL empowered by large language models: A benchmark evaluation. Proc. of the VLDB Endowment, 2024, 17(5): 1132–1145. [doi: 10.14778/3641204.3641221]

[43] Zheng YD, Li XF, Chen HP, He GJ. Survey on SQL generation based on Chinese natural language. Computer Systems Applications, 2023, 32(12): 32–42 (in Chinese with English abstract). [doi: 10.15888/j.cnki.csa.009356]

[44] Han YY, Liu JP, Luo A, Wang Y, Bao S. Fine-tuning LLM-assisted Chinese disaster geospatial intelligence extraction and case studies. ISPRS Int'l Journal of Geo-information, 2025, 14(2): 79. [doi: 10.3390/ijgi14020079]

[45] Zafarmomen N, Samadi V. Can large language models effectively reason about adverse weather conditions? Environmental Modelling & Software, 2025, 188: 106421. [doi: 10.1016/j.envsoft.2025.106421]

[46] Cai JH, Miao RS. Construction and scenario implementation of language emergency service system based on knowledge graph. Computer Science and Application, 2025, 15(12): 1–13 (in Chinese with English abstract). [doi: 10.12677/csa.2025.1512317]

[47] Xie YXY, Jiang BW, Mallick T, Bergerson J, Hutchison JK, Verner DR, Branham J, Alexander MR, Ross RB, Feng Y, Levy LA, Su WJ, Taylor CJ. MARSHA: Multi-agent RAG system for hazard adaptation. npj Climate Action, 2025, 4(1): 70. [doi: 10.1038/s44168-025-00254-1]

[48] Hu EJ, Shen YL, Wallis P, Allen-Zhu ZY, Li YZ, Wang S, Wang L, Chen WZ. LoRA: Low-rank adaptation of large language models. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[49] Chen JL, Xiao ST, Zhang PT, Luo K, Lian DF, Liu Z. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In: Proc. of the 2024 Association for Computational Linguistics (ACL 2024). Bangkok: ACL, 2024. 2318–2335. [doi: 10.18653/v1/2024.findings-acl.137]

[50] Lin CY. ROUGE: A package for automatic evaluation of summaries. In: Proc. of the 2024 Text Summarization Branches Out. Barcelona: ACL, 2004. 74–81.

[51] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: ACL, 2002. 311–318. [doi: 10.3115/1073083.1073135]

[52] Qwen Team. Qwen2.5 technical report. arXiv:2412.15115, 2024.

[53] DeepSeek-AI. DeepSeek-V3.2: Pushing the frontier of open large language models. arXiv:2512.02556, 2025.

[54] Rawat R, Zhu K. DisasterQA: A benchmark for assessing the performance of LLMs in disaster response. In: Proc. of the 2024 Harms and Risks of AI in the Military. Montreal, 2024.

附中文参考文献

[1] 王喆, 陆俊燃, 杨栋梁, 李墨潇. 融合 GPT 和知识图谱的洪涝应急决策智能问答系统研究. 中国安全生产科学技术, 2024, 20(4): 5–11. [doi: 10.11731/j.issn.1673-193x.2024.04.001]

[2] 桑基韬, 于剑. 从 ChatGPT 看 AI 未来趋势和挑战. 计算机研究与发展, 2023, 60(6): 1191–1201. [doi: 10.7544/issn1000-1239.202330304]

[43] 郑耀东, 李旭峰, 陈和平, 贺桂娇. 基于中文自然语言的 SQL 生成综述. 计算机系统应用, 2023, 32(12): 32–42. [doi: 10.15888/j.cnki.csa.009356]

[46] 蔡剑寒, 苗润生. 基于知识图谱的语言应急服务系统构建与场景实现. 计算机科学与应用, 2025, 15(12): 1–13. [doi: 10.12677/csa.2025.1512317]

作者简介

冯援, 硕士生, CCF 学生会员, 主要研究领域为大模型智能体与强化学习.

胡平, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为智能媒体分析与理解, 图像视频理解生成, 多模态大模型.

张璐, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为图像视频理解, 多模态大模型.

沈复民, 博士, 教授, 博士生导师, 主要研究领域为跨模态学习与理解, 多模态大模型.

申恒涛, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为多媒体搜索, 计算机视觉, 人工智能, 大数据管理.

朱晓峰, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为多视图学习, 表示学习, 大数据技术.