https://doi.org/10.3799/dqkx.2026.032

基于多源持续预训练与集成检索增强生成的矿产勘查大语言模型构建

张雨昂[1],谢 忠[2,3,4],田 苗[3],吴麒瑞[4],吴 亮[2],邱芹军[2*],陈建国[3,5]

1. 中国地质大学(武汉)地理与信息工程学院,湖北武汉 430078
2. 中国地质大学(武汉)计算机学院,湖北武汉 430078
3. 中国地质大学(武汉)地质探测与评估教育部重点实验室,湖北武汉 430074
4. 中国地质大学(武汉)未来技术学院,湖北武汉 430078
5. 中国地质大学(武汉)资源学院,湖北武汉 430074

摘 要: 为解决矿产勘查场景下通用大语言模型领域语料稀缺、领域术语覆盖与语体适配不足、事实性幻觉突出的问题,构建约2 500万token规模的领域语料库,在此基础上提出课程式持续预训练策略,按术语、机制、案例三阶段组织训练数据,并配合渐进式Transformer block解冻与学习率调度,对Qwen3-1.7B进行持续预训练以实现分阶段领域适配,得到面向矿产勘查场景的大语言模型Geo-MineLLM;推理阶段集成Hybrid RAG,以混合检索与证据约束生成提升事实一致性.人工评估表明,Geo-MineLLM相较基座模型与同系列更大参数规模的模型能显著提升领域问答表现;集成Hybrid RAG后,综合领域问答表现接近GPT-4.1.该训练、推理一体化方案为矿产勘查领域大模型构建与可靠问答提供了轻量化路径.

关键词: 大语言模型;持续预训练;检索增强生成;矿产勘查;人工智能.

中图分类号: P628 **文章编号:** 1000-2383(2026)03-1025-15 **收稿日期:** 2025-12-30

A Large Language Model for Mineral Exploration via Multi-Source Continual Pre-Training and Integrated Retrieval-Augmented Generation

Zhang Yuang[1], Xie Zhong[2,3,4], Tian Miao[3], Wu Qirui[4], Wu Liang[2], Qiu Qinjun[2*], Chen Jianguo[3,5]

1. *School of Geography and Information Engineering, China University of Geosciences (Wuhan), Wuhan* 430078, *China*
2. *School of Computer Science, China University of Geosciences (Wuhan), Wuhan* 430078, *China*
3. *Key Laboratory of Geological Survey and Evaluation of Ministry of Education, China University of Geosciences (Wuhan), Wuhan* 430074, *China*
4. *School of Future Technology, China University of Geosciences (Wuhan), Wuhan* 430078, *China*
5. *School of Earth Resources, China University of Geosciences (Wuhan), Wuhan* 430074, *China*

Abstract: To address the challenges faced by general-purpose large language models in mineral exploration, including scarcity of domain corpora, insufficient coverage of domain terminology and register adaptation, and pronounced factual hallucinations. We

基金项目: 国家自然科学基金项目(Nos. 42301492, 42571487);国家重点研发计划项目(Nos. 2023YFC2906404, 2023YFC2906400).

作者简介: 张雨昂(1997 —),男,博士研究生,从事地质知识图谱构建及领域大模型应用研究. ORCID:0009-0000-6213-9081. E-mail: zhangyuang@cug.edu.cn

* **通讯作者:** 邱芹军,ORCID:0000-0002-9850-3751. E-mail:qiuqinjun@cug.edu.cn

引用格式: 张雨昂,谢忠,田苗,吴麒瑞,吴亮,邱芹军,陈建国,2026.基于多源持续预训练与集成检索增强生成的矿产勘查大语言模型构建.地球科学,51(3):1025-1039.

Citation: Zhang Yuang, Xie Zhong, Tian Miao, Wu Qirui, Wu Liang, Qiu Qinjun, Chen Jianguo, 2026. A Large Language Model for Mineral Exploration via Multi-Source Continual Pre-Training and Integrated Retrieval-Augmented Generation. *Earth Science*, 51(3): 1025-1039.

constructed a mineral-exploration corpus of approximately 25 million tokens and, on this basis, proposed a curriculum-based continual pre-training strategy, which organizes training data into three stages: terminology, mechanisms, and cases. Coupled with gradual unfreezing of Transformer blocks and learning-rate scheduling, we conducted continual pre-training of Qwen3-1.7B to achieve stage-wise domain adaptation, resulting in a mineral-exploration-oriented LLM, Geo-MineLLM. During inference, we integrated a Hybrid RAG framework, leveraging hybrid retrieval and evidence-constrained generation to enhance factual consistency. Human evaluation indicates that Geo-MineLLM substantially improves domain question-answering performance relative to the base model and larger-parameter models within the same family. With Hybrid RAG enabled, overall domain QA performance approaches that of GPT-4.1. The proposed training-inference integrated framework provides a lightweight pathway for building mineral-exploration LLMs and enabling reliable domain-specific question answering.

Key words: large language models; continual pre‑training; retrieval‑augmented generation; mineral exploration; artificial intelligence.

0 引言

近年来,以大语言模型(Large Language Models, LLMs)为代表的新一代生成式人工智能技术推动多个行业发生范式变革(Lachowycz, 2024),如医学(Yang *et al.*, 2022)、农业(Zhang *et al.*, 2024b)、金融(Wu *et al.*, 2023)、工程(Hou *et al.*, 2024)等.与这些领域类似,LLMs的发展也改变了地球科学的研究范式(左仁广等, 2024),领域学者开展了一系列相关研究,如Deng *et al.*(2024)提出首个地球科学领域的LLM模型K2、Zhang *et al.*(2024a)将LLM固有的语义理解能力与GIS领域内成熟的工具相结合并开发GeoGPT框架.

人工智能是挖掘和提升大数据价值的关键(Qiu *et al.*, 2023a).当前,基于LLMs对地球科学中的地质资源、矿产勘查领域的探索仍十分有限,近期工作(成秋明, 2025; 师路易和左仁广, 2026)表明,构建矿产勘查领域大模型有望成为推动该领域智能化发展的新范式,可作为突破研究方法瓶颈的重要途径,具有明确的研究必要性.Wu *et al.*(2025)提出面向地质领域的GeoProspect,通过构建高质量地质语料并结合持续预训练与有监督指令微调实现通用模型的领域适配.Fu *et al.*(2025)提出一种融合私人地质文档和知识注入方法的混合方案,以开发大语言模型GeoMinLM用于地质调查和矿产勘探领域的智能问答应用.

与通用领域相比,矿产勘查场景具有更强的表达专业性,其知识高度依赖长期积累的专业术语、文献资料、调查报告,同时受数据隐私与获取成本因素影响,通用LLMs在该场景下面临领域语料供给不足、专业知识覆盖有限以及术语与语体适配能力不足等问题.持续预训练(Continual Pre‑Training, CPT)是一种在已有通用预训练语言模型的基础上,使用目标领域或目标任务分布的无标注文本,沿用原有自监督预训练目标进行下一阶段预训练(Gururangan *et al.*, 2020),可用于缓解通用语料与目标领域语料之间的分布偏移,从而提升模型的领域适配能力.然而,若在缺乏数据配比、选择与保留机制的情况下对目标域偏置的数据分布进行朴素持续预训练,LLMs模型可能出现既有能力退化,即灾难性遗忘等风险(Gupta *et al.*, 2023).

通用大语言模型在应对专业领域与知识密集型问题时,往往更易产生事实性幻觉(Ji *et al.*, 2023),是LLMs可靠性的重要障碍(Farquhar *et al.*, 2024).为缓解幻觉,先前研究(Lewis *et al.*, 2020)通常引入检索增强生成(Retrieval‑Augmented Generation, RAG)框架,通过在生成过程中检索外部证据并将其作为条件约束模型生成,以提升回答的可归因性与事实性.在地学语境下,已有工作通过将领域模型GeoGPT与LightRAG架构相结合以构建问答系统,从而加速矿产勘探知识发现与应用(Zhou and Li, 2025).

尽管上述相关研究已推出如K2、GeoGPT、GeoProspect等地球科学大模型或框架,推动了地学知识问答与文本理解能力的发展,但矿产勘查任务在语料形态、知识组织与推理链条上与通用地学场景存在显著差异.其一,矿产勘查语料以勘查报告、调查成果和专题评价为主,普遍呈现长篇幅、强流程叙事与高密度专业术语等特征;其二,矿产勘查知识具有更突出的层次化结构与证据链依赖,往往需要在术语定义、成矿机制、典型案例之间建立稳定衔接.受上述差异影响,直接沿用通用地学模型的训练范式可能容易出现领域术语覆盖不足、机制性表述不稳定以及在缺乏证据约束时的事实性幻

觉等问题(张宝一等, 2026).为此,本文面向多源地学数据,构建规模约 2 500 万 token 的矿产勘查领域语料库,包含辞典语料、文献语料与报告语料三类子集,以多维度覆盖领域专业知识与术语表达;依托语料库专业知识结构,提出面向矿产勘查场景的课程式持续预训练策略(Prospect-Curriculum CPT),即按术语、机制、案例分阶段组织训练数据,配合渐进式 Transformer block 解冻与学习率调度,实现递进式领域适配,据此开发领域大语言模型(Geological Mineral Exploration Large Language Model, Geo-MineLLM);针对领域专业知识问答任务,在推理阶段基于 Geo-MineLLM 集成混合检索生成(Hybrid RAG)框架,通过混合检索与证据约束生成机制增强回答的事实一致性;引入人工评估方法,多维度、多角度对模型输出质量进行评估与对比分析,实验结果表明,Geo-MineLLM 相较多种对比模型在矿产勘查问答任务中能取得更优表现,且在集成 Hybrid RAG 后整体性能可与 GPT-4.1 达到相近水平.

1　数据集与方法

1.1　语料库构建

地质知识是人类对地质对象或过程的空间分布、演化和相互作用模式的认知成果,矿产勘探数据作为地质知识具有显著层次性(Qiu *et al.*, 2023b).本文将其归纳为三层,一是术语与定义层面,包括词条、名词解释以及计量单位与年代范式等基础概念;二是机制与过程层面,涉及成矿作用链等推理性内容;三是案例与报告层面,侧重勘查线索的综合分析、证据链条的叙述及研究结论的归纳.本文以新疆东天山地区为主要研究区域,面向铜矿及多金属成矿规律认知建模与智能问答应用场景,基于领域知识层次性构建由辞典、文献、报告三类来源组成的多源中文语料库,用于后续 LLMs 持续预训练.

表 1　语料明细

Table 1　Corpus details

语料数据明细	辞典语料	文献语料	报告语料
文件数(个)	8	106	152
token 数(万)	1 314	72	1 135
样本平均 token 数(个)	124.36	417.76	413.49
样本数(个)	105 676	1 720	27 465

语料库构建过程具有较强的工程复杂性与时间成本,原因在于原始数据载体以 PDF、Word 等非结构化文档为主,且文本抽取与去噪是影响语料质量的关键环节.

为保证后续训练文本的语义完整性与结构一致性,本文首先明确抽取范围:保留题名、摘要、作者信息、章节标题及章节正文、结论等与知识表达直接相关的内容;而对目录、页眉页脚、页码、版权声明、水印以及参考文献列表等非正文信息进行剔除,以降低版式噪声对模型学习的干扰;对于图表信息,直接删除可能造成论述链条断裂与指代丢失,本文在不进行表格数值重建的前提下,选择在图表出现位置邻近段落的段尾保留对应图名与表名,以维持文本叙事的连贯性与必要的指代线索.

完成抽取后,本文以自然段落为基本切分单元,按照原文顺序将段落依次序列化为 JSONL 样本,每条样本仅包含 text 字段,用以承载该段文本内容.同时,为避免过短段落导致的语义碎片化与上下文割裂,本文对长度不足的段落与其后相邻段落进行合并,以提高样本的语义完整度与训练稳定性.语料库构建的整体流程如图 1 所示,分为三种子语料,明细见表 1.

现在对三种子语料进行详细介绍,分别对应的数据示例见表 2.

(1)辞典语料:语料组成分为四部分,首先是专

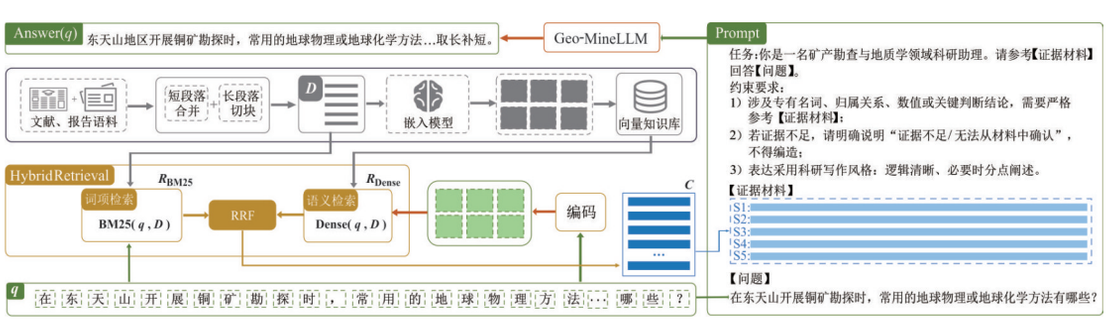

图 1　集成 Hybrid RAG 的智能问答框架

Fig.1　An intelligent question-answering framework integrating Hybrid RAG

表2 语料样本示例

Table 2 Example of corpus samples

数据类别	样本示例
辞典数据	黄铜矿,化学组成为$CuFeS_2$.有同质三象变体.常见的是四方晶系变体温度稳定在550～213 ℃.成分中常有机械混入物,银、金、铊、硒、碲、锗、镓、铟等.黄铜矿中还常见闪锌矿、黝锡矿等的小包体.晶体少见,呈四方四面体,主要呈致密块状或粒状集合体.黄铜黄色.表面常有蓝、紫红的斑状锖色.条痕绿黑色.金属光泽.不透明.硬度3～4.性脆.相对密度4.1～4.3.黄铜矿分布很广,可在各种条件下形成.主要有超基性岩铜镍硫化物矿床,也有接触交代矽卡岩矿床,与黄铁矿、磁铁矿、磁黄铁矿共生.
文献数据	已有研究表明,在所有的流体类型中,岩浆流体对金的迁移能力最强,而单纯的岩浆结晶作用并不能形成金矿,这是因为岩浆中大部分Au将随岩浆结晶分异和去气作用而被萃取至流体相,而且这些过程会产生使成矿流体沿通道发生迁移的驱动力,这种驱动作用对于金矿床的形成至关重要.马头滩矿区内广泛发育岩浆岩,且相邻矿区内含矿围岩与矿石的Sr、Nd、pb、S、Si同位素具有继承相似性,说明岩浆作用为Au的富集和运移提供了前提(刘重芃等, 2014).
报告数据	东天山成矿区包括有8个Ⅲ级构造-成矿单元,面积约18×10^4 km^2.区内已发现铜矿产地63处,其中有超大型资源远景的矿床1处(延东)、大型矿床1处(土屋)、中型矿床2处(黄山及黄山东)、小型矿床11处.已探明铜储量145.08×10^4 t,占新疆已知铜储量的29.30%;占天山已知铜储量的55.22%.

业领域辞典,主要涵盖《地球科学大辞典:基础学科卷》、《地球科学大辞典:应用学科卷》、《地球科学大辞典:地球化学》、《中国百科大辞典》第十四卷:地质学部分;其二是引入全国科学技术名词审定委员会公布的《地质学名词》等权威审定资源.其三,整理并引入地学领域本体知识,将地质要素如地质体、地质事件、地质年代等相关词汇对齐到可复用的标准语义框架.其四,通过网络爬虫程序爬取地质学家介绍、地质学著作等用于补全辞典覆盖之外的情况,仅选择来源清晰、可追溯且可核验的公开资源.

(2)文献语料:覆盖期刊论文与会议论文两类来源,其中期刊论文98篇,来自领域知名中文期刊如《地球科学》、《中国地质》、《岩石学报》、《地质学报》、《西北地质》、《矿床地质》、《地质通报》等,会议论文8篇,主要来自全国青年地质大会.数据聚焦我国东天山-北山及其邻区等典型有色、贵金属成矿带与关键成矿单元,覆盖天山中西段相关区段及新疆北部-甘肃西部结合部等重点研究区.矿种方面以铜、金、铅锌及多金属组合为主,并涉及镍等相关矿床研究.文献内容侧重矿床地质特征、控矿构造与岩浆作用、成矿机制与成因模型、找矿标志及预测意义等方向.

(3)报告语料:报告数据来自全国地质资料馆、地质云等线上开放平台.报告类型可归纳为矿产勘查与评价类报告、基础地质与区域地质调查类报告、专题研究与应用类报告三种,其中矿产勘查与评价类报告为最主要类型,共有95篇,包括普查报告、详查报告、资源储量核实报告、矿产调查与评价报告、区域典型矿种报告等;基础地质与区域地质调查类共有33篇,主要包括区域地质调查报告、图幅说明书、构造与成矿带研究等,其中包含丰富的数据和文本信息(He *et al.*, 2024);专题研究与应用类报告共有24篇,主要包括关键地质问题调查报告、成矿规律研究、新技术应用等.整合东天山-北山及其邻区、哈密周缘、天山成矿带等典型构造-成矿单元的地质调查与矿产勘查成果,报告多以金属矿产如金、铜、铁、钴及铅锌多金属组合为研究对象,空间范围延伸至安徽、甘肃、广东、贵州、湖北、辽宁、内蒙古、山东、山西、陕西、西藏、黑龙江、海南、河南等省、自治区,覆盖造山带、克拉通边缘、陆内盆地等多种构造背景下的不同类型工作区.语料涵盖传统矿产勘查流程,包括区域地质填图、矿产地质调查、资源潜力评价与规划、详查-勘探阶段钻探验证等,并收纳城市环境地质、水文地质及地质灾害(邱芹军等, 2023)调查与区划等应用领域成果.

1.2 课程式持续预训练方法

本文提出面向矿产勘查领域语料结构与知识形态的课程式持续预训练策略Prospect-Curriculum CPT.该策略旨在将术语、机制、案例的认知顺序工程化:在约2 500万token的专业语料规模下,通过三阶段课程编排即由基础概念到过程机理再到案例叙事,结合渐进式Transformer block解冻与分阶段学习率调度实施持续预训练,旨在强化领域术语覆盖与语体适配能力的同时,尽可能保持基座模型的通用能力,并降低领域漂移与过拟合风险.

Prospect-Curriculum CPT的设计理念受到Bengio *et al.*(2009)研究中的课程学习思想启发,按样本复杂程度进行数据分桶、以学习节奏控制促进稳定收敛与泛化.在参数更新策略上,本方法引入迁移学习中的渐进解冻范式(Howard and Ruder,

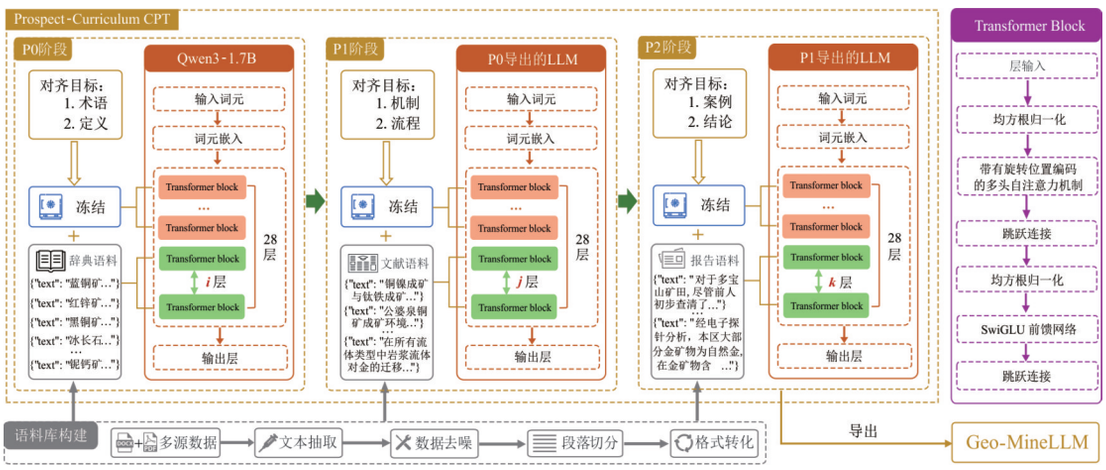

图 2　Prospect-Curriculum CPT 方法

Fig.2　Prospect-Curriculum CPT Method

2018),以缓解因过于激进的整体更新导致的能力退化与遗忘风险.在优化配置方面,本方法借鉴 Gupta *et al.* (2023)的研究,在各阶段切换时重新设置学习率,以兼顾阶段式领域适配的训练效率与稳定性.Prospect-Curriculum CPT 方法见图 2.

P0、P1、P2 三阶段分别以术语与定义、机制与流程、案例与结论为对齐目标.综合硬件条件与训练开销,本文选用包含 28 个 Transformer block 的 Qwen3-1.7B 作为 CPT 基座模型.已有研究表明(Jawahar *et al.*, 2019),Transformer 模型内部存在清晰的层级功能分工,底层主要承载通用词法与基础句法表示,中层负责跨句语义组合,而高层更侧重概念抽象与任务相关推理.因此,在 CPT 过程中仅解冻中高层并保持底层冻结,旨在注入领域知识的同时,较好地保留模型原有的通用语言能力.

本文对可训练 Transformer block 层数设置全局上限,基于渐进解冻范式以及 Raffel *et al.* (2020)研究中部分参数冻结以保持既有知识的做法,优先更新包含相对较少通用知识的顶层以缓解灾难性遗忘.同时,在不同阶段采用差异化的解冻深度设置.P0 阶段以大规模辞典与术语语料为主,侧重领域概念与术语语义空间的构建,解冻 i 层 Transformer block,以增强高层语义对领域概念的适配能力,同时避免对底层通用表示的扰动.P1 阶段语料规模较小,主要包含成矿机制与地质过程描述,为降低过拟合风险并保持前序阶段已形成的领域语义结构,解冻最顶层 j 层 Transformer block,对高层语义风格与机制性表达进行有限调整.P2 阶段以地质调查报告和找矿案例为主,文本篇幅长、逻辑结构复杂,对篇章级建模与推理能力要求更高,解冻最顶层 k 层 Transformer block,使模型中高层能够参与跨段语义融合与推理模式学习,同时继续冻结底层以控制领域漂移.i、j、k 取值见章节 2.2.

通过在不同知识层级与语义复杂度下动态调节模型可塑性,旨在实现在有限专业语料条件下稳定注入领域知识与保留通用能力之间的平衡.

1.3　集成 Hybrid RAG 的智能问答框架

矿产勘查场景的报告与调查成果通常篇幅较长、结构复杂且包含大量与问题无关的流程性叙述与版式噪声,同时受模型上下文窗口长度约束,长文直接拼入 Prompt 会带来输入截断风险,并显著增加推理开销,也容易引入无关信息干扰,削弱回答的论述聚焦.此外,矿产勘查问答往往对专业术语、细粒度事实高度敏感,仅依赖 CPT 的模型仍然容易产生事实性幻觉(Ji *et al.*, 2023),需要兼顾词项匹配召回的精确性与语义召回的语义泛化能力.

基于以上考虑,本文在推理阶段引入混合检索的检索增强生成框架 Hybrid RAG.该框架通过融合基于词项的检索方法 BM25(Robertson and Zaragoza, 2009)与基于语义向量的检索方法(Karpukhin *et al.*, 2020),结合排序融合与证据约束生成机制(Reciprocal Rank Fusion, RRF;Cormack *et al.*, 2009)整合多路检索结果,以提升检索召回率与系统鲁棒性.

Geo-MineLLM 在生成时以检索到的外部证据作为条件输入并进行约束生成,从而提升输出的事实一致性.框架工作流程见图 1.

本文基于第 1.1 节所述的文献与报告语料构建

外部向量化知识库.先对语料进行段落级预处理,通过短段落合并与长段落条件切块操作,形成文档块集合$D=\{d_1,d_2,\cdots,d_N\}$,其中,每个文档块d_i为段落级文本单元,是后续检索与证据引用的最小语义单位.在此基础上,采用bge-large-zh-v1.5中文嵌入模型对文档块进行向量化表示,并将生成的嵌入向量注入向量知识库,为后续语义检索与Hybrid RAG问答提供支撑.

Hybrid RAG兼顾关键词匹配能力与语义泛化能力,采用双通道检索机制,在不同表示空间中对同一文档块集合进行检索.现设检索通道集合为$R=\{\mathrm{BM25},\mathrm{Dense}\}$,每个$r\in R$,分别表示一种独立检索通道.

基于词项的检索BM25在文本段落层面对问题q与文档块$d\in D$进行词项匹配,得到按相关性排序的文档块列表$R_{\mathrm{BM25}}=\mathrm{BM25}(q,D)$,该检索过程完全基于文本倒排索引,不依赖向量表示,能够有效捕获专业术语、地名与地质单元等显式信息.

基于语义向量的检索通过将问题与文档块映射至同一向量空间,并基于向量相似度进行近邻搜索,得到语义相关的候选结果$R_{\mathrm{Dense}}=\mathrm{Dense}(q,D)$,在实现层面,该过程等价于在向量数据库中对问题向量执行近邻检索;在方法层面,其本质是在嵌入空间中寻找与问题语义最相近的文档块.

由于BM25与语义检索在相关性评分尺度上不可直接比较,Hybrid RAG采用基于排名的融合方法RRF,对双通道检索结果进行融合.对于每一个检索通道$r\in R$,记其返回的排序结果为R_r.对于任意后续按文档块$d\in\bigcup_{r\in R}R_r$,其RRF融合得分公式为:

$$\mathrm{RRF}(d)=\sum_{r\in R}\frac{1}{k+\mathrm{rank}_r(d)},\tag{1}$$

其中,$\mathrm{rank}_r(d)$表示文档块d在通道r的排序结果R_r中的名次;k为平滑常数,用于控制不同排名位置的贡献衰减.本文结合矿产勘查语料以长文档为主、段落冗余与模板性叙述较多的特点,为进一步降低排名候选对融合结果的边际影响、加强对尾部噪声的抑制,采用比默认值$k=60$略偏保守的$k=80$作为平滑常数.根据RRF得分进行排序,并选取得分最高的Top-K=5个文档块,构成最终证据材料集合C.

在生成阶段,本文采用证据约束的检索增强生成策略,将检索得到的证据集合显式注入Geo-MineLLM的Prompt中,以提升生成结果的可验证性.当检索阶段未能返回任何证据,系统直接输出“证据不足,无法回答”,以避免事实性幻觉.当证据集合非空时,Geo-MineLLM在证据条件约束下生成答案$\mathrm{Answer}(q)$.

2 模型训练

2.1 训练环境

所有实验均在统一的硬件与软件环境下完成.实验硬件平台配置为NVIDIA RTX 4090 D显卡、128 GB运行内存及Intel i9-14900KF处理器,操作系统为Windows 10.关键软件环境及其版本包括Python 3.11、Torch 2.5.1+cu124、Elasticsearch 9.0.1以及Transformers 4.51.3.模型CPT实验基于LLaMA Factory开展,LLaMA Factory是一个用于微调大型语言模型的强大工具,可以适应不同的模型架构和大小,支持多种微调与训练技术.

2.2 持续预训练

三阶段语料规模差异显著,本文基座模型Qwen3-1.7B包含28层Transformer block,基于训练成本与稳定性预实验对比5、7、9、11层不同解冻深度的训练、验证损失稳定性与收敛趋势,现将最大解冻深度限定为模型顶部7层Transformer block,并在各阶段依据语料规模与语义复杂度进行阶段化调整:P0阶段基于辞典语料集,因规模较大且为对齐术语、定义,使用上限解冻即$i=7$以建立地学概念适配空间;P1阶段基于文献语料集,需对齐机制、过程,因语料规模最小,应收缩至更浅解冻以保持前序语义框架稳定,经对比1、2、3层不同解冻深度预实验,确定P1阶段解冻深度为$j=3$;P2阶段基于报告语料集,规模较大,需适度加深解冻以增强跨段落语义融合能力,经对比4、5、6层不同解冻深度预实验,确定P2阶段解冻深度为$k=5$.接下来介绍每阶段具体训练,参数优化对于提高模型准确性和可重复性至关重要.在后续CPT实验中,峰值学习率、模型导出策略的设置对模型训练效果有显著影响,因此本文将着重论述三阶段CPT中对峰值学习率的确定,并详细规定模型导出策略.三阶段CPT通用超参数设置见表3.

2.2.1 辞典语料注入:P0阶段 P0阶段以术语与定义对齐为主要目标,训练语料主要由辞典条目、名词解释与规范化定义文本构成,整体呈现定义性强、句式相对稳定、概念边界清晰的语言特征.为在既定数据规模与训练预算约束下获得稳健的优化配方,本阶段在其余训练配置保持一致的条件下,

表 3　三阶段 CPT 通用超参数设置

Table 3　Three-stage CPT general hyperparameter settings

超参数	值
Compute type	fp16
Cutoff length	2 048
Batch size	8
Gradient accumulation	2
LR schedule	cosine
Pack sequences	true
Enable thinking	true

设置验证集为 5%,对峰值学习率进行对比实验:设置 1×10^{-5}、3×10^{-5}、5×10^{-5}、8×10^{-5} 四组候选值,并在统一训练轮次上限 Epoch=4 下开展持续预训练,以综合评估不同学习率对收敛速度、验证集表现与训练过程稳定性的影响.P0 阶段的训练损失与验证损失曲线如图 3 所示,四组峰值学习率由小至大对应图 3 中从左往右的四列图像,其中图 3a～3d 给出不同峰值学习率下的训练损失曲线,图 3e～3h 给出对应的验证损失曲线.

由图 3 可知,随着峰值学习率增大,训练损失下降速度整体加快,表明更大的步长能够提高短期收敛效率;但当学习率增至 8×10^{-5} 时,验证损失在训练后段出现更明显的平台波动与回升趋势,提示在该设置下继续迭代可能带来泛化收益递减,并伴随一定不稳定风险.相较之下,1×10^{-5} 与 3×10^{-5} 的验证损失下降过程更为平滑,但在相同训练预算内其验证损失水平整体高于更大学习率组,体现出收敛偏慢的特征.综合验证损失水平与后期稳定性,5×10^{-5} 在本阶段呈现更为均衡的表现:其验证损失整体低于 1×10^{-5} 与 3×10^{-5},同时相较 8×10^{-5} 具有更好的后期稳定性.因此,本文将 5×10^{-5} 作为 P0 阶段峰值学习率,并以验证损失最低点作为截断依据进行模型选择,即在当前 Epoch 上限约束下选取验证损失达到最小值时对应的全局训练步所保存的模型检查点,导出作为 P1 阶段初始化模型,本阶段取 Training Step=1 100.

2.2.2　文献语料注入:P1 阶段　P1 阶段面向机制、过程层面的知识对齐,且该阶段语料规模相对有限,若延续大范围参数更新,容易在小样本条件下放大过拟合风险并扰动 P0 阶段已建立的术语、概念语义空间.因此,本研究在 P1 阶段采用更保守的分层冻结策略.具体而言,本阶段将训练轮次上限设置为 10 epochs,并将验证集比例提高至 10%,以增强对泛化行为的过程性约束.同时,基于 cosine 学习率调度,使学习率在训练中后期逐步衰减,降低训练后期更新幅度与振荡,抑制小样本条件下的过拟合倾向,从而提升训练稳定性与泛化能力.在此基础上,为系统评估不同峰值学习率对收敛速度与泛化趋势的影响,本阶段设置 1×10^{-5}、3×10^{-5}、5×10^{-5}、1×10^{-4} 四组候选峰值学习率开展对比实验.P1 阶段的训练损失与验证损失曲线如图 4 所示,四组峰值学习率由小至大对应图 4 中从左往右的四列图像,其中图 4a～4d 给出不同峰值学习率下的训练损失曲线,图 4e～4h 给出对应的验证损失曲线.

结果表明,随着峰值学习率增大,训练损失下降速率加快,其中 1×10^{-4} 在训练端收敛最快.然而,从验证曲线可见,该组在训练中后段即 Training Step 为 60 之后验证损失出现持续且幅度较大的回升(图 4h),呈现先降后升的反弹形态.这表明在当前小样本语料条件下,过大的峰值学习率会使有效更新步长过强,即便采用 cosine 衰减,仍可能在中前期引入较大幅度的参数扰动并导致后期泛化退化,

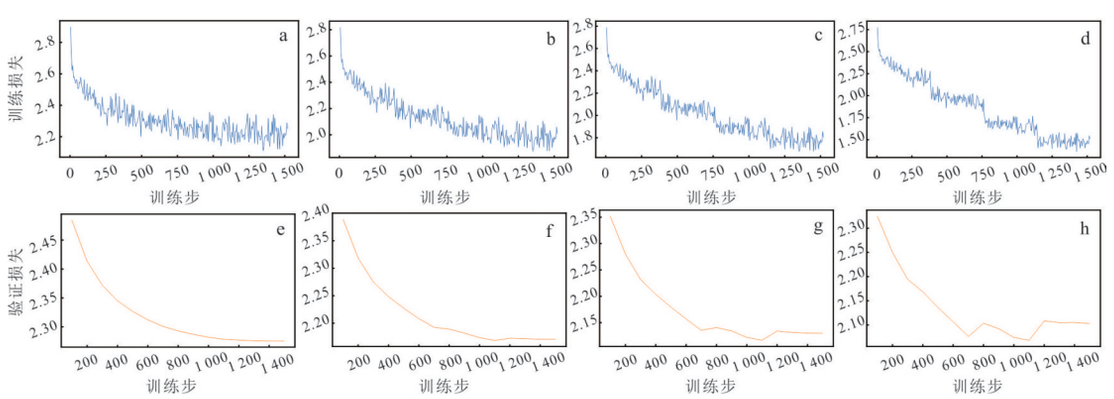

图 3　P0 阶段 CPT 的训练损失曲线与验证损失曲线

Fig.3　Training loss curve and validation loss curve for CPT during phase P0

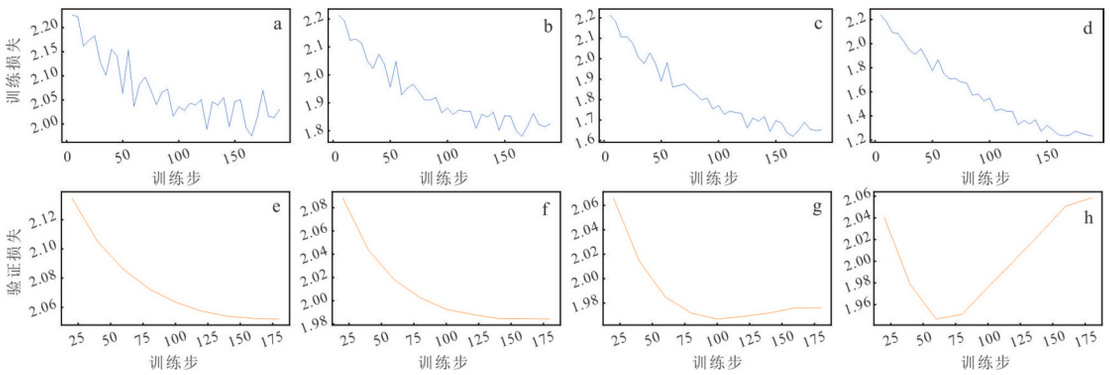

图 4 P1 阶段 CPT 的训练损失曲线与验证损失曲线

Fig.4 Training loss curve and validation loss curve for CPT during phase P1

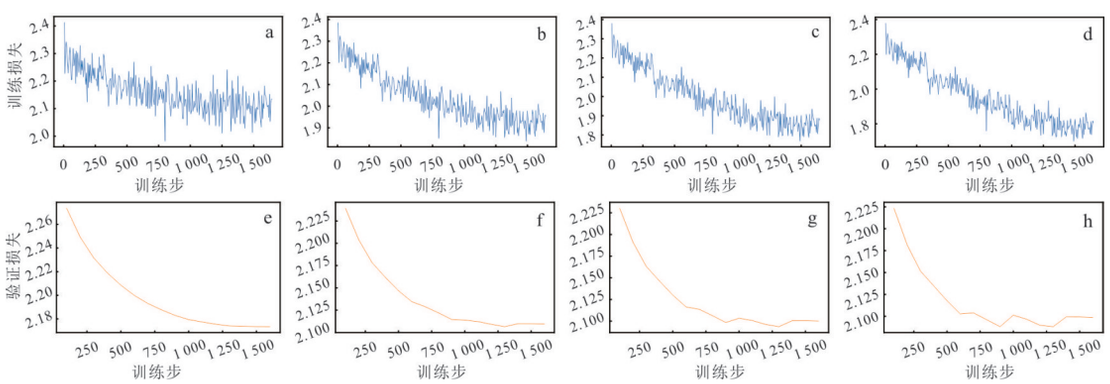

图 5 P2 阶段 CPT 的训练损失曲线与验证损失曲线

Fig.5 Training loss curve and validation loss curve for CPT during phase P2

表现为过拟合或不稳定优化.相对而言,1×10^{-5}的验证损失下降过程较为平滑,但整体下降幅度有限,且训练损失曲线波动相对更明显、最终验证损失水平高于中等学习率组(图 4e),说明该学习率设置下更新强度不足,存在一定欠拟合倾向.

综合训练与验证表现,峰值学习率为 3×10^{-5}与 5×10^{-5}均取得较好验证损失水平与收敛趋势(图 4f、4g),但二者在稳定性与扰动风险上存在差异:在小规模语料条件下,5×10^{-5}的更新幅度相对更大,潜在表示漂移风险更高;而 3×10^{-5}的验证损失下降更为单调、后期未出现反弹且趋于稳定,体现出更稳健的泛化趋势.为在有限语料上优先保证泛化稳定性、同时控制对前序阶段表征扰动,本阶段选择 3×10^{-5}作为 P1 阶段的峰值学习率.

模型导出遵循验证集最优原则,训练过程中以固定频率评估验证损失,并选择验证损失达到最小值时的模型检查点作为 P2 阶段初始化权重.在本实验的 10 epochs 范围内,峰值学习率为 3×10^{-5}的验证损失持续下降后趋于稳定且未出现回升,最优检查点位于训练末段,因此最终导出末轮对应的最优检查点用于后续阶段初始化.

2.2.3 报告语料注入:P2 阶段 在完成词汇与学术表达层面的适配后,P2 阶段进一步引入规模更大、结构更复杂的地质勘查报告语料,以增强模型对长篇技术文本、勘查流程描述及多要素综合论述的建模能力.本阶段训练语料为整理后的地质调查与勘查报告文本,总规模约 1 135 万 token.

考虑到语料规模显著扩大,为避免过度训练导致的表示漂移与计算成本上升,P2 阶段将训练轮次上限设置为 5 epochs,并固定验证集比例为 5%.在峰值学习率设置上,首先选取 1×10^{-5}、3×10^{-5}与 5×10^{-5}三组峰值学习率开展对比实验,综合分析训练损失与验证损失的收敛行为,以确定合理的参数更新区间.在此基础上,进一步补充 4×10^{-5}学习率实验,用于检验 5×10^{-5}条件下验证损失最低点是否源于过快学习带来的偶然性收敛谷底,以及在更温和更新强度下是否能够稳定逼近相同的验证性能.P2 阶段的训练损失与验证损失曲线如图 5 所示,四组峰值学习率由小至大对应图 5 中从左往右的四列图像,其中图 5a~5d 给出不同峰值学习率下的训

练损失曲线,图 5e～5h给出对应的验证损失曲线.

由图 5 可见,随着峰值学习率的逐步增大,训练损失整体呈现出更快的下降趋势,且训练曲线的局部波动逐渐减弱,表明模型在更高更新强度下能够更充分地拟合报告语料所包含的复杂表达结构.其中,峰值学习率为 5×10^{-5}时训练损失下降速率最快,显示出较强的参数更新能力;相比之下,1×10^{-5}虽然整体收敛过程较为平稳,但在限定训练轮次内损失下降幅度有限,且训练曲线存在较为明显的局部起伏,说明其更新幅度不足以有效驱动模型对报告语料中深层结构性与跨段落信息的充分建模.3×10^{-5}与 4×10^{-5}在收敛速度与训练稳定性之间表现出相对均衡的特征,能够在保证下降效率的同时维持较为平滑的训练轨迹.

进一步结合图 5e～5h的验证损失曲线分析,不同学习率在泛化表现上呈现出明显差异.峰值学习率为 1×10^{-5}时,验证损失随训练过程持续下降,但其最终水平显著高于其余配置,表明在当前语料规模与训练设置下,该学习率可能存在一定欠拟合倾向.5×10^{-5}在训练中前期达到最低验证损失,但随后出现一定程度的波动与回升,结合其训练端较大的更新幅度,可推断该最低点可能与较大的步长更新引发的偶然性收敛有关,存在一定稳定性风险.相比之下,4×10^{-5}的验证损失下降过程更为平滑,未出现明显反弹,在稳定收敛的前提下逐步逼近 5×10^{-5}所达到的最低验证损失水平,体现出更可靠的泛化趋势.

综合考虑训练损失的收敛效率、验证损失的稳定性以及大规模报告语料对参数更新安全边界的要求,本文在P2阶段最终选取 4×10^{-5}作为峰值学习率.选择验证损失进入稳定最低区间时对应的训练步所保存的模型检查点,作为后续实验阶段使用的最终模型Geo-MineLLM,本阶段取 Training Step=1 300.

3　评估与讨论

3.1　准备

本文引入研究(Fu *et al.*, 2025)提出的评估方法,并在其基础上扩展评估指标体系、细化评估粒度.围绕基础地质、地质矿产资源勘探、矿床类型与分布以及成矿规律四个维度,每个维度包含15个问题,构建包含共60个问题的评测集,用于系统评估不同LLMs在矿产勘查领域的知识理解与推理能力.对比模型涵盖:当前具有代表性的商业通用大模型GPT-4.1与DeepSeek-V3,均通过API调用;本地部署大模型 Qwen3-4B、Qwen3-1.7B、Geo-MineLLM,以及 Geo-MineLLM+Hybrid RAG.

为系统评估 Prospect-Curriculum CPT 方法及其集成 Hybrid RAG 智能问答框架的有效性,本文设计两阶段对照实验:

(1)Geo-MineLLM与轻量级本地大语言模型的对比,旨在评估基于Prospect-Curriculum CPT的领域知识注入在相同模型架构和近似规模条件下,对地质领域任务能力的提升效果.

(2)集成Hybrid RAG后的Geo-MineLLM与商用通用大语言模型的对比,旨在衡量结合检索增强生成策略后的Geo-MineLLM在专业问答任务中的竞争力,及其与当前业界通用模型基准之间的性能差异.

为保证评估结果的专业性与可靠性,本文邀请5位地质学家对各模型在相同问题集下生成的回答进行时间性能评估与独立人工评估.为降低主观偏差,所有模型回答均以匿名形式呈现,评估者未知模型来源.最终结果取三位评估者评分的平均值作为模型在各指标下的最终得分.

独立人工评估指标与评分规则包括四个方面:

(1)逻辑清晰度.评价模型回答在结构组织、论述连贯性及表达简洁性方面的表现.若模型回答逻辑关系明确、无多余重复且结构完整,记2分;若回答在逻辑组织上基本合理、但存在部分表达冗余或衔接欠佳,记1分;若回答逻辑混乱、无明显结构或难以理解,记0分.

(2)表达专业性.考察模型回答在专业术语使用、语言规范性及学科风格符合程度方面的表现.若回答全面采用地质学专有术语、表述规范且符合专业写作风格,记2分;若回答具有一定专业表达但术语使用不够准确或存在措辞偏差,记1分;若回答缺乏专业性、使用术语错误或表述模糊,记0分.

(3)回答全面性.评估模型是否覆盖问题所涉及的核心要点及相关背景信息.若回答全面且涉及问题的所有关键维度,记2分;若回答部分覆盖关键要点,但遗漏若干重要内容,记1分;若回答只涉及少量相关内容且无法满足问题要求,记0分.

(4)事实准确性.判断模型回答是否符合地学事实与问题语境,是否存在概念混淆、事实错误或不当推断.若回答在事实层面无明显错误且解释依据明确,记2分;若回答在大体事实框架上无严重错误,但存在细节不精确或轻微混淆,记1分;若回答

包含明显事实错误或误解基本概念，记 0 分.

3.2　评估结果

为消除不同解码策略对推理耗时的干扰、保证评估结果的公平性，本文对参与对比的本地模型统一采用确定性解码配置：设定 do_sample 为 False 且 num_beams 为 1，因此采样相关参数如 temperature、top_p 不参与解码过程且对输出与耗时不产生影响. 同时，统一设置 max_new_tokens 为 1 024 作为生成长度上限，从而减少由生成上限与解码策略差异引入的时延偏置. 本地模型推理均在同一硬件与软件环境下完成，统一设置 low_cpu_mem_usage 为 True 用于降低模型加载阶段的 CPU 内存峰值，统一采用 fp16 半精度以降低显存占用. 为避免通过改变数值精度或计算路径引入不可比性，所有本地模型均未启用 int8 或 int4 量化等额外加速策略.

时间性能评估采用模型从接收输入问题到生成完整回答的实际耗时作为衡量指标，以模型完成全部评测问题的总耗时的平均值表示.

由表 4 可得，六个模型在推理时延上存在显著差异：Geo-MineLLM 平均耗时最低，为整体最快；其后依次为 Qwen3-1.7B 与 Geo-MineLLM ＋ Hybrid RAG 等，Qwen3-4B 因在本地部署模型中的参数量最大，耗时最多. 整体上，Geo-MineLLM 体现出在小参数量模型与领域化知识注入条件下具有更高的推理效率；集成 Hybrid RAG 后，Geo-MineLLM 的耗时由 5.39 s 增至 8.22 s，反映出检索、重排与证据融合等环节带来的额外开销，但其端到端时延仍显著低于 GPT-4.1、DeepSeek-V3 等，揭示 Hybrid RAG 虽带来一定时延代价但可保持较高效率. 总之，在本文实验设置下，本地化领域模型及其 RAG 增强形态在推理时延方面具备较强的工程可用性与部署优势.

表 4　不同模型的时间性能

Table 4　Time performance of different models

模型名	时间性能(s)
DeepSeek-V3	15.86
GPT-4.1	14.31
Geo-MineLLM	**5.39**
Qwen3-1.7B	7.54
Qwen3-4B	22.95
Geo-MineLLM ＋ Hybrid RAG	8.22

3.2.1　Prospect-Curriculum CPT 有效性评估

现对 Geo-MineLLM 与本地 LLMs Qwen3-4B、Qwen3-1.7B 进行人工评估与对比，结果见图 6. 图中的每个柱形表示对应模型在该维度某一指标上的人工评分平均值之和.

由图 6a～6d 所示，三种模型在四项指标上呈现出较为一致的差异. Geo-MineLLM 在四个维度中均取得最高综合表现，优势主要集中于表达专业性与回答全面性：在图 6b～6d 三个维度中，Geo-MineLLM 在上述两项指标上显著高于 Qwen3-4B 与 Qwen3-1.7B，体现出 Prospect-Curriculum CPT 带来的领域术语掌握度、学科表述规范性与要点覆

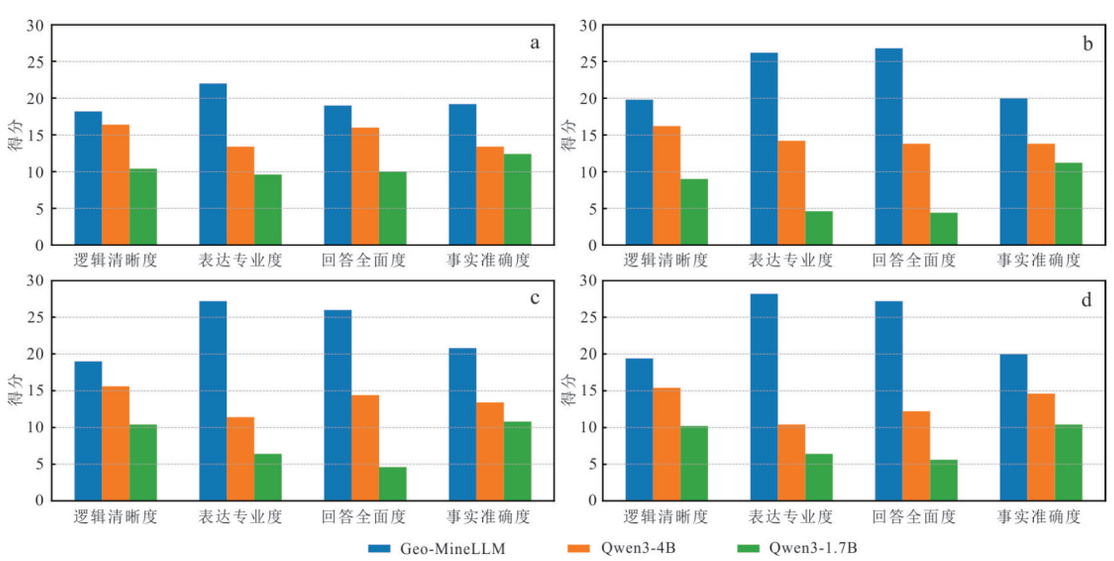

图 6　人工评估与对比

Fig.6　Human evaluation and comparison

a. 基础地质维度；b. 地质矿产资源勘探维度；c. 矿床类型与分布维度；d. 成矿规律维度

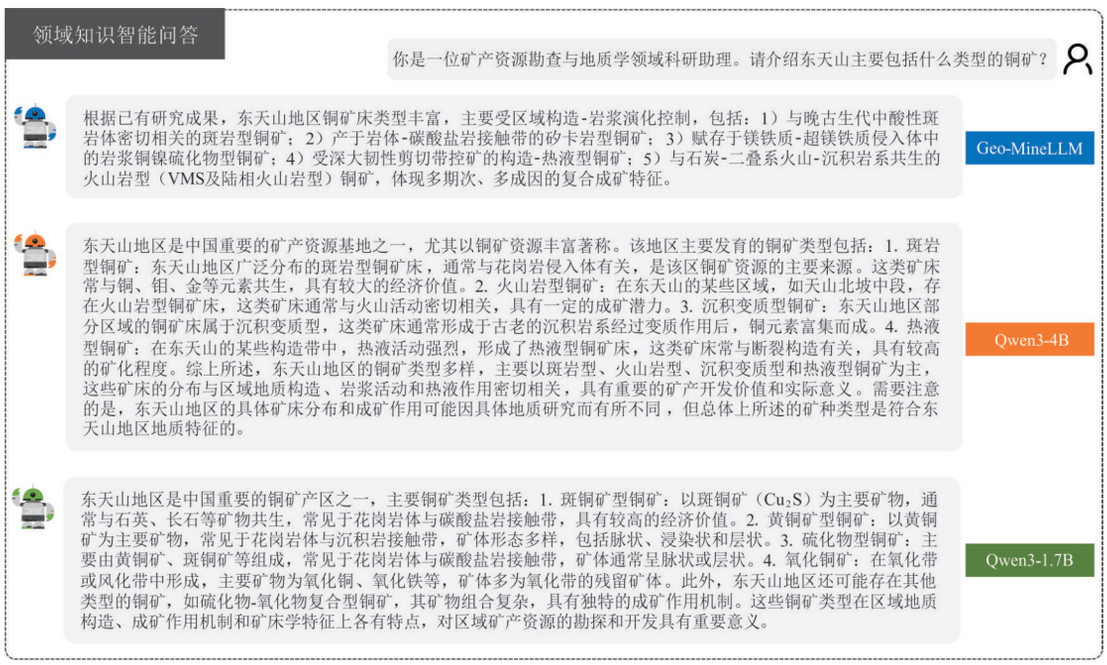

图 7　模型问答示例

Fig.7　Example of model-based question answering

盖能力提升;同时,Qwen3-4B 在表达专业性、回答全面性上略高于 Qwen3-1.7B,符合模型规模与通用能力差异的预期.就事实准确性而言,Geo-MineLLM 在四个维度上均表现为略优,其优势幅度相对小于前述两项指标;且 Qwen3-4B 与 Qwen3-1.7B 在事实准确性上的差距整体较小,呈现 Qwen3-4B 轻微占优但接近的态势.就逻辑清晰度而言,Geo-MineLLM 相较 Qwen3-4B 存在轻微优势,而 Qwen3-4B 相较 Qwen3-1.7B 优势更为稳定,说明在论述结构组织与连贯性方面,领域化知识注入与更强的基础生成能力均能带来正向增益.

跨维度对比进一步表明,Geo-MineLLM 在图 6b 勘探、图 6c 矿床类型与分布、图 6d 成矿规律三个维度的整体得分水平高于图 6a 基础地质维度,揭示其优势更集中地体现在需要综合术语、机制解释与多要点组织的专业问答场景中;而基础地质维度相对更接近背景性、概念性知识问答,其模型间差距相对收敛.

在相近参数规模的本地模型对比条件下,Geo-MineLLM 通过 Prospect-Curriculum CPT 实现的领域知识注入能够稳定、有效提升矿产勘查问答质量,且提升主要体现在专业表达、回答完整性等面向任务的生成质量维度,而事实准确性改善相对温和.不同模型问答示例见图 7.

3.2.2　集成 Hybrid RAG 智能问答框架有效性评估　现对集成 Hybrid RAG 后的 Geo-MineLLM 与商用通用大语言模型 GPT-4.1 与 DeepSeek-V3 进行对比,结果见图 8.

由图 8 可知,Geo-MineLLM + Hybrid RAG 在四个维度的事实准确度上均表现出显著优势,其柱状高度在各子图中明显高于 DeepSeek-V3、GPT-4.1,表明引入 Hybrid RAG 后,模型能够更有效地约束生成并提升事实一致性与错误抑制能力;同时,DeepSeek-V3 在事实准确度上整体略优于 GPT-4.1,但二者与 Geo-MineLLM + Hybrid RAG 之间仍存在清晰差距.就表达专业度而言,Geo-MineLLM+Hybrid RAG 在四个维度中普遍保持一定优势,说明领域化模型在术语使用、学科化表述风格与语言规范性方面仍具有相对优势,同时引入外部知识片段会一定程度提高回答专业性.相比之下,在逻辑清晰度维度上,Geo-MineLLM+Hybrid RAG 整体呈现一定劣势:多数子图中其得分略低于 DeepSeek-V3、GPT-4.1,但差距幅度较小,反映出在引入检索链路后,回答结构组织并未获得同等幅度增益,且可能受到证据整合与信息拼接带来的叙述连贯性波动影响.在回答全面度指标上,Geo-MineLLM+Hybrid RAG 在四个维度中均低于 DeepSeek-V3 与 GPT-4.1,且 DeepSeek-V3 整体表现最强、GPT-4.1 次之,说明商业通用大模型在开放式扩展与要点覆盖方面仍具优势;相较之下,Geo-MineLLM + Hybrid RAG 策略更突出事实对

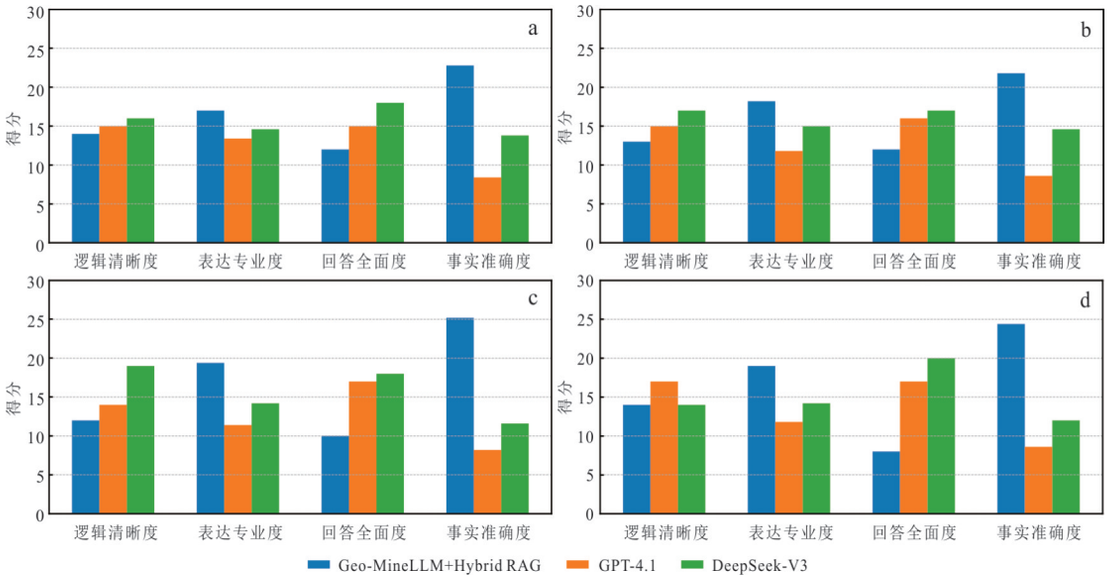

图 8 人工评估与对比

Fig.8 Human evaluation and comparison

a. 基础地质维度；b. 地质矿产资源勘探维度；c. 矿床类型与分布维度；d. 成矿规律维度

图 9 模型问答示例

Fig.9 Example of model-based question answering

齐而非内容扩展，从而形成准确性显著提升但全面性相对不足的现象 .

评估结果验证了集成 Hybrid RAG 智能问答框架对专业问答任务中事实可靠性的显著增益，表明轻量级模型在通过领域知识注入与检索增强生成策略可以达到综合表现不输于商业通用

大模型的表现.不同模型回答对比示例见图9.

3.3 讨论

本文方法仍存在一定局限.其一,当前语料主要集中于新疆东天山及相关成矿带,对其他地区与矿种覆盖不足,在超出主要语料分布范围且知识边界模糊或证据不足的情形下,模型仍可能产生一定程度的事实性幻觉,特别是在涉及具体数值、精细年代或局部地质细节时更为明显.其二,受模型参数规模限制,对于跨多个成矿阶段、需综合多源证据的复杂推理任务,1.7B的模型在此类复杂推理任务上的原生涌现能力仍存在上限,在推理深度与稳定性上仍弱于大型通用模型.而面向矿产勘查中高频的事实性、机制解释型与弱推理型问答,Geo-MineLLM + Hybrid RAG具有较为出色的性能.本文方法并不是对大型通用模型的全面替代,而是作为一种有限资源与条件下的领域适配轻量化方案.其三,本文对Hybrid RAG的评估主要面向检索返回证据整体可信且与问题相关的常规场景,并进一步通过Prompt规范证据不足的情况,但尚未在当前版本中系统构造并量化评测错误证据或矛盾证据对生成结果的影响.

4 结论

本文面向矿产勘查领域的专业知识问答需求,提出领域大语言模型Geo-MineLLM及其训练、推理一体化技术路线,旨在缓解领域语料稀缺与模型幻觉带来的专业性不足、事实一致性偏弱问题.(1)构建的高质量、多源融合的领域语料库,在公开资源相对匮乏的背景下兼顾数据的专业覆盖与内容完整性,可为领域模型研发提供专业的数据基础;(2)提出Prospect-Curriculum CPT策略,通过三阶段逐步解冻可训练层,实现递进式领域知识注入与领域表达结构学习,从而显著提升模型在矿产勘查问答任务中的术语规范性、论述专业度与回答完备性;(3)综合评估结果,Geo-MineLLM相较基座模型与同系列更大参数规模的模型具有显著优势,在推理侧引入Hybrid RAG后,能显著提升事实准确性与一致性,从系统层面降低幻觉风险,矿产勘查领域问答表现达到与主流通用大语言模型相当的水平.(4)Geo-MineLLM在硬件成本、响应速度具有独特优势,可显著降低部署与应用门槛,可为矿区现场、勘探团队等资源受限场景提供一定即时问答支持.

未来研究计划将围绕以下方向展开:

(1)在更大参数规模基座模型如7B、14B模型上验证方法可迁移性,以系统评估Prospect-Curriculum CPT方法在更高复杂度场景下的适用性与性能上限.

(2)构建规模更大、结构更丰富的领域语料库,拓展空间范围、矿种类型,引入更多区域地质报告与多尺度勘查数据,以提升模型泛化能力并缓解信息孤岛问题.

(3)后续工作将围绕噪声证据鲁棒性边界开展系统评测与方法增强.构建分层噪声注入基准,并探索证据质量控制与一致性约束策略如基于融合得分阈值的证据过滤、来源可信度加权、跨证据一致性检测与冲突重排等,以进一步提升Hybrid RAG在复杂证据环境下的稳定性.

(4)在Prospect-Curriculum CPT后续流程中引入针对性微调策略,动态调整Hybrid RAG策略以适应空间关系抽取、地名实体识别、矿床要素结构化提取等地学下游任务,推动模型由领域问答向地学智能分析工具拓展.

References

Bengio, Y., Louradour, J., Collobert, R., et al., 2009. Curriculum Learning. The 26th Annual International Conference on Machine Learning. Montreal. https://doi.org/10.1145/1553374.1553380

Cheng, Q. M., 2025. A New Paradigm for Mineral Resource Prediction Based on Human Intelligence-Artificial Intelligence Integration. *Earth Science Frontiers*, 32(4): 1−19 (in Chinese with English abstract).

Cormack, G. V., Clarke, C. L. A., Buettcher, S., 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. The 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston. https://doi.org/10.1145/1571941.1572114

Deng, C., Zhang, T. H., He, Z. M., et al., 2024. K2: A Foundation Language Model for Geoscience Knowledge Understanding and Utilization. The 17th ACM International Conference on Web Search and Data Mining. Merida. https://doi.org/10.1145/3616855.3635772

Farquhar, S., Kossen, J., Kuhn, L., et al., 2024. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature*, 630(8017): 625−630. https://doi.org/10.1038/s41586-024-07421-0

Fu, Y., Wang, M. G., Wang, C. B., et al., 2025. GeoMinLM: A Large Language Model in Geology and

Mineral Survey in Yunnan Province. *Ore Geology Reviews*, 182: 106638. https://doi.org/10.1016/j.oregeorev.2025.106638

Gupta, K., Thérien, B., Ibrahim, A., et al., 2023. Continual Pre-Training of Large Language Models: How to (Re) Warm Your Model? ICML2023, Hawaii. https://doi.org/10.48550/arXiv.2308.04014

Gururangan, S., Marasović, A., Swayamdipta, S., et al., 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. The 58th Annual Meeting of the Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.740

He, H., Ma, C., Ye, S., et al., 2024. Low Resource Chinese Geological Text Named Entity Recognition Based on Prompt Learning. *Journal of Earth Science*, 35(3): 1035－1043. https://doi. org/10.1007/s12583-023-1944-8

Hou, X. Y., Zhao, Y. J., Liu, Y., et al., 2024. Large Language Models for Software Engineering: A Systematic Literature Review. *ACM Transactions on Software Engineering and Methodology*, 33(8): 1－79. https://doi.org/10.1145/3695988

Howard, J., Ruder, S., 2018. Universal Language Model Fine-Tuning for Text Classification. The 56th Annual Meeting of the Association for Computational Linguistics. Melbourne. https://doi.org/10.18653/v1/p18-1031

Jawahar, G., Sagot, B., Seddah, D., 2019. What Does BERT Learn about the Structure of Language? The 57th Annual Meeting of the Association for Computational Linguistics, Florence. https://doi. org/10.18653/v1/P19-1356

Ji, Z. W., Lee, N., Frieske, R., et al., 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 1－38. https://doi.org/10.1145/3571730

Karpukhin, V., Oguz, B., Min, S., et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering. The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online. https://doi.org/10.18653/v1/2020.emnlp-main.550

Lachowycz, S., 2024. Utility of Artificial Intelligence in Geoscience. *Nature Geoscience*, 17(10): 953－955. https://doi.org/10.1038/s41561-024-01548-5

Lewis, P., Perez, E., Piktus, A., et al., 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv*, 2005.11401. https://arxiv. org/abs/2005.11401

Liu, C. P., Yang, H. M., Duan, R. C., et al., 2014. Metallogenic Age of the Matoutan Gold Deposit in East Tianshan and Its Geological Significance. *Geological Bulletin of China*, 33(6): 912－923 (in Chinese with English abstract).

Qiu, Q. J., Tian, M., Xie, Z., et al., 2023a. Extracting Named Entity Using Entity Labeling in Geological Text Using Deep Learning Approach. *Journal of Earth Science*, 34(5): 1406－1417. https://doi. org/10.1007/s12583-022-1789-8

Qiu, Q. J., Wang, B., Ma, K., et al., 2023b. A Practical Approach to Constructing a Geological Knowledge Graph: A Case Study of Mineral Exploration Data. *Journal of Earth Science*, 34(5): 1374－1389. https://doi.org/10.1007/s12583-023-1809-3

Qiu, Q. J., Wu, L., Ma, K., et al., 2023. A Knowledge Graph Construction Method for Geohazard Chain for Disaster Emergency Response. *Earth Science*, 48(5): 1875－1891 (in Chinese with English abstract).

Raffel, C., Shazeer, N., Roberts, A., et al., 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1－67.

Robertson, S., Zaragoza, H., 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333－389. https://doi.org/10.1561/1500000019

Shi, L. Y., Zuo, R. G., 2026. Foundation Model for Mineral Prospectivity Mapping. *Earth Science*, 53(3): 832－848 (in Chinese with English abstract).

Wu, G., Wang, H. T., Zhang, K. Y., et al., 2025. GeoProspect: A Domain-Specific Geological Large Language Model with Enhanced Continual Learning. *Neurocomputing*, 650: 130801. https://doi.org/10.1016/j.neucom.2025.130801

Wu, S. J., Irsoy, O., Lu, S., et al., 2023. BloombergGPT: A Large Language Model for Finance. *arXiv*, 2303.17564. https://arxiv.org/abs/2303.17564

Yang, X., Chen, A. K., PourNejatian, N., et al., 2022. A Large Language Model for Electronic Health Records. *NPJ Digital Medicine*, 5: 194. https://doi.org/10.1038/s41746-022-00742-2

Zhang, B. Y., Tang, J. C., Zhang, T. Y., et al., 2026. Knowledge Graph and Question-Answering Model for Geological Prospecting Empowered by Large Language Models. *Earth Science*, 53(3): 982－995 (in Chinese with English abstract).

Zhang, K. P., Ma, L., Cui, B. B., et al., 2024a. Visual Large Language Model for Wheat Disease Diagno-

sis in the Wild. *Computers and Electronics in Agriculture*, 227: 109587. https://doi.org/10.1016/j.compag.2024.109587

Zhang, Y. F., Wei, C., He, Z. T., et al., 2024b. GeoGPT: An Assistant for Understanding and Processing Geospatial Tasks. *International Journal of Applied Earth Observation and Geoinformation*, 131: 103976. https://doi.org/10.1016/j.jag.2024.103976

Zhou, B., Li, K., 2025. Fusing Geoscience Large Language Models and Lightweight RAG for Enhanced Geological Question Answering. *Geosciences*, 15(10): 382. https://doi.org/10.3390/geosciences15100382

Zuo, R. G., Cheng, Q. M., Xu, Y., et al., 2024. Explainable Artificial Intelligence Models for Mineral Prospectivity Mapping. *Scientia Sinica* (*Terrae*), 54(9): 2917—2928 (in Chinese with English abstract).

中文参考文献

成秋明, 2025. 面向人类智能与人工智能融合的矿产资源预测新范式. 地学前缘, 32(4): 1—19.

刘重芃, 杨红梅, 段瑞春, 等, 2014. 东天山马头滩金矿的成矿时代及其地质意义. 地质通报, 33(6): 912—923.

邱芹军, 吴亮, 马凯, 等, 2023. 面向灾害应急响应的地质灾害链知识图谱构建方法. 地球科学, 48(5): 1875—1891.

师路易, 左仁广, 2026. 矿产预测大模型. 地球科学, 53(3): 832—848.

张宝一, 唐嘉成, 张彤蕴, 等, 2026. 大语言模型赋能的地质找矿知识图谱与问答模型构建. 地球科学, 53(3): 982—995.

左仁广, 成秋明, 许莹, 等, 2024. 可解释性矿产预测人工智能模型. 中国科学: 地球科学, 54(9): 2917—2928.