分析与研究

文章编号：1671−251X(2025)08−0034−10 DOI：10.13272/j.issn.1671-251x.2025030031

基于 LoRA 微调与 RAG 融合的煤矿专业大模型应用关键技术

秦一凡[1]， 付翔[1,2,3]， 张智星[1]， 贾一帆[1]， 孙岩[1]

（1. 太原理工大学 矿业工程学院，山西 太原 030024；2. 智能采矿装备技术全国重点实验室，山西 太原 030024；3. 山西焦煤集团有限责任公司 博士后工作站，山西 太原 030024）

摘要：目前煤矿行业大模型仅对用户的提问进行知识问答，未与现场实时数据相关联，无法对煤矿生产运行状况进行实时分析与指导。针对这些问题，提出一种基于大语言模型的低阶适应（LoRA）微调和检索增强生成（RAG）融合的煤矿专业大模型。该模型先运用 LoRA 技术从历史文本语料中抽取出知识实体并定义知识结构输入大模型进行微调，使大模型能够深入理解领域知识，再将实时产生的生产数据、实时更新的作业规程、法规条例等数据经过向量化清洗输入向量数据库，并与 RAG 的检索机制相结合，确保数据信息的实时性和准确性。实验结果表明：① 经 LoRA 微调后，模型回答可以精准契合某煤矿“一通三防”管理制度汇编，不仅详细阐述了控制瓦斯排放的增阻限风、分风限风、逐段排放等具体方法，还对排放时间计算、传感器设置、图纸绘制及断电撤人等操作细则进行说明，实现了从泛泛而谈到精准定位具体煤矿特定文件内容的跨越。② 选取现场 143 万条液压支架时序数据，分别存入 Milvus 向量数据库与 MySQL 关系型数据库，从写入效率与查询性能 2 个维度进行对比，结果表明：Milvus 向量数据库写入速度为 MySQL 关系型数据库的 2.4 倍；在向量检索场景方面，Milvus 的向量相似度检索延迟稳定在 20 ms，在混合查询场景效率方面，MySQL 需全表扫描后排序，143 万条数据延迟超 100 ms，而 Milvus 将设备 ID 过滤后的子集输入分层可导航小世界图（HNSW）层，仅读取查询涉及的向量字段，避免了全表扫描。③ 将本地基于 LoRA 微调与 RAG 融合的煤矿专业大模型与本地离线 deepseekR1−7b 模型进行部署，对多项指标进行测试，结果表明：基于 LoRA 微调与 RAG 融合的煤矿专业大模型在煤矿专业领域知识学习性、知识动态化更新时效性、模型泛化与回答精确度方面具有显著优势，为工业级 AI 落地提供了可行路径。

关键词：人工智能；煤矿专业大模型；大语言模型的低阶适应微调；检索增强生成；分层可导航小世界图；LoRA 微调；RAG；HNSW

中图分类号：TD67 文献标志码：A

Key technologies for application of coal mine domain large model based on LoRA fine-tuning and RAG fusion

QIN Yifan[1]， FU Xiang[1,2,3]， ZHANG Zhixing[1]， JIA Yifan[1]， SUN Yan[1]

(1. College of Mining Engineering, Taiyuan University of Technology, Taiyuan 030024, China;
2. State Key Laboratory of Intelligent Mining Equipment Technology, Taiyuan 030024, China;
3. Post-doctoral workstation, Shanxi Coking Coal Group Co., Ltd., Taiyuan 030024, China)

Abstract: At present, large models in the coal mine industry only perform knowledge question answering for

收稿日期：2025-03-07；**修回日期**：2025-08-08；**责任编辑**：王晖,郑海霞。

基金项目：国家自然科学基金项目（52274157）；山西省基础研究计划联合资助项目（202403011241002）；“科技兴蒙”行动重点专项项目（2022EEDSKJXM010）。

作者简介：秦一凡（1999—），男，山西太原人，博士研究生，研究方向为智能采矿数智赋能技术，E-mail：838501076@qq.com。通信作者：付翔（1986—），男，山西长治人，副教授，博士（后），博士研究生导师，研究方向为智能采矿数智赋能技术，E-mail：14632235@qq.com。

引用格式：秦一凡，付翔，张智星，等. 基于 LoRA 微调与 RAG 融合的煤矿专业大模型应用关键技术[J]. 工矿自动化，2025，51(8)：34-42，50.
QIN Yifan, FU Xiang, ZHANG Zhixing, et al. Key technologies for application of coal mine domain large model based on LoRA fine-tuning and RAG fusion[J]. Journal of Mine Automation，2025，51(8)：34-42，50.

扫码移动阅读

users' queries, without being linked to real-time on-site data, and therefore cannot conduct real-time analysis and guidance on coal mine production and operation conditions. To address these problems, a coal mine domain large model based on Low-Rank Adaptation (LoRA) fine-tuning and Retrieval-Augmented Generation (RAG) fusion was proposed. The model first used LoRA technology to extract knowledge entities from historical text corpora and define knowledge structures, which were then input into the large model for fine-tuning, enabling the fine-tuned large model to deeply understand domain knowledge. Then, real-time production data, updated operating procedures, and regulations were vectorized, cleaned, and input into a vector database, and combined with the retrieval mechanism of RAG to ensure the timeliness and accuracy of the information. Experimental results showed that: ① after LoRA fine-tuning, the model's answers precisely matched a certain coal mine's "One Ventilation and Three Prevention" management regulations compilation, not only elaborating specific methods such as increasing resistance to limit airflow, branch airflow limiting, and section-by-section discharge for controlling gas emissions, but also explaining operational details such as discharge time calculation, sensor setting, drawing preparation, and power-cut evacuation, thus achieving a leap from general discussion to precisely locating the content of specific coal mine documents. ② A total of 1.43 million items of hydraulic support time-series data from the site were stored separately in the Milvus vector database and the MySQL relational database. A comparison was made in two dimensions: write efficiency and query performance. The results showed that the write speed of the Milvus vector database was 2.4 times that of MySQL. In vector retrieval scenarios, the vector similarity retrieval latency of Milvus was stable at the 20 ms level; in hybrid query scenarios, MySQL needed to perform a full table scan followed by sorting, with a latency exceeding 100 ms for 1.43 million data entries, whereas Milvus filtered the subset by equipment ID and then input it into the Hierarchical Navigable Small World (HNSW) graph, reading only the vector fields involved in the query, thereby avoiding a full table scan. ③ The locally deployed coal mine domain large model based on LoRA fine-tuning and RAG fusion, along with and the offline DeepSeekR1-7b model, were tested on multiple indicators. The results show that the coal mine domain large model based on LoRA fine-tuning and RAG fusion has significant advantages in domain knowledge learning, timeliness of dynamic knowledge updates, model generalization, and answer accuracy, providing a feasible path for the industrial application of AI.

Key words: Artificial Intelligence; coal mine domain large model; Low-Rank Adaptation fine-tuning; Retrieval-Augmented Generation; Hierarchical Navigable Small World; LoRA fine-tuning; RAG; HNSW

0　引言

煤炭行业具有专业知识复杂、数据类型多样、现场环境复杂多变等特点，传统的信息处理方式不能有效融合行业知识与现场数据，难以满足对实时洞察和智能决策的需求。为深入探究并获取更有价值的煤矿专业知识，必须采取一种高效的方法来系统地描述、整理煤矿相关信息[1-5]。

随着人工智能技术的迅猛发展，在大模型的垂域应用方面[6-8]，现有通用生成式人工智能通过引入更多的专业领域知识和数据来增强其跨领域的能力，能够自适应多种任务和场景，将这些前沿技术应用于煤矿行业，不仅有助于提升行业的智能化水平，更能显著提高生产效率和安全性，推动传统产业的数字化转型[1,9]。

大模型首先将煤矿领域的历史文本、操作规程、安全条例等专业数据作为微调的核心数据[10]，进行清洗、标注和结构化处理，然后进行领域自适应微调，进而实现知识关联与知识推理，完成对于煤矿领域知识的适配。目前，大模型构建技术已应用于煤矿领域。江耀森等[11]提出了一种基于 ChatGLM3−6B 模型的煤矿事故知识图谱构建方法，该方法通过构建提示模板，利用 ChatGLM3−6B 的上下文学习能力进行知识抽取，将结果存储于 Neo4j 图数据库中，并结合 LangChain 框架，实现针对煤矿事故知识图谱的智能查询问答功能，但提示模板的设计需要人工干预，且可能无法覆盖所有复杂的煤矿事故场景，易导致信息抽取不完整或错误。许洁等[12]提出了一种基于煤矿安全生产知识库的煤炭行业大模型生成方法，构建行业特定数据集，通过多样化的指令增强大模型对非正式查询的理解能力，但该方法对复杂或模糊的非正式查询的响应不准确，且未充分解决知识库和模型的动态更新问题，难以适应煤炭行业快速变化的安全规范和标准。王海玲等[13]设计了提示工程方法，利用 ChatGLM大语言模型解决煤矿安全

预警知识的自动提取问题，但其知识抽取的准确性依赖于输入数据的质量，如果原始文本数据存在噪声或不完整，可能导致知识图谱构建不准确。

当前煤矿行业大模型仅对用户的提问进行知识问答[14-15]，并未与现场实时数据相关联，现有大模型不具备对现场数据进行清洗、特征提取、分析的功能，无法对煤矿生产运行状况进行实时分析与指导，并无法对操作人员生成建议提供决策支持[1,16-18]。检索增强生成（Retrieval-Augmented Generation，RAG）技术可以连接到一个专门构建的向量数据库，该数据库存储了煤矿相关的实时生产数据、最新法规、技术文档等信息的向量化表示[19]。这些数据会随着时间推移不断更新，确保数据库中的知识始终与行业最新发展同步，从而使微调后的大模型能够从外部知识源中动态获取信息来辅助生成决策[20]。鉴此，本文提出一种基于大语言模型的低阶适应（Low-Rank Adaptation of Large Language Models，LoRA）微调和 RAG 融合的煤矿专业大模型，先运用 LoRA 技术从历史文本语料中抽取出知识实体并定义知识结构输入大模型进行微调，使大模型能够深入理解领域知识，再将实时产生的生产数据、实时更新的作业规程、法规条例等数据经过向量化清洗输入向量数据库，并与 RAG 的检索机制相结合，确保数据信息的实时性和准确性[21-23]。

1　煤矿专业大模型垂域应用技术架构

煤矿专业大模型垂域应用技术架构包括多模态数据库层、大模型主动洞察业务逻辑层和自然语言交互应用层，如图 1 所示。

多模态数据库层分为指令微调知识库与向量数据库。指令微调知识库将煤矿历史专业文献、设计资料、技术资料、煤矿规程等数据进行清洗、标注和格式化，构建高质量的模型微调训练与推理特征数据集，并输入大模型进行煤矿专业微调训练。向量数据库将现场实时产生的设备数据、生产数据、环境数据、文本数据等结构化和非结构化文本数据进行清洗与分块，再通过嵌入模型（Embedding Model）将其转换为高维向量，完成向量化存储，为大模型主动洞察业务逻辑层提供高效语义检索基础。

大模型主动洞察业务逻辑层首先通过煤矿专业大模型 LoRA 微调训练逻辑后台将指令微调知识库输入大模型，完成大模型在煤矿领域的高效适配，得到煤炭行业知识与现场数据融合分析大模型结构体，即微调后大模型。然后引入 RAG 技术，通过大模型专业数据向量化 RAG 后台将向量数据库中的

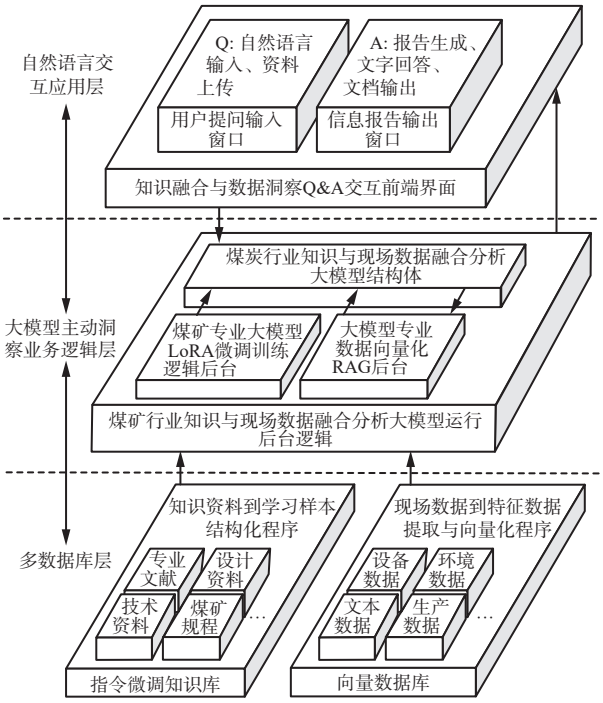


图 1　基于 LoRA 微调与 RAG 融合的煤矿专业大模型技术架构

Fig. 1　Technical architecture of coal mine domain large model based on LoRA fine-tuning and RAG fusion

向量数据输入微调后大模型，并结合查询生成最终的回答，使得 RAG 能够从海量最新数据中精准检索出与查询相关的上下文信息，提升生成模型的效果，解决模型泛化能力不足和知识利用率低的问题，为自然语言交互应用层提供强大的模型支持。

自然语言交互应用层通过用户提问输入窗口与信息报告输出窗口完成交互，利用微调后大模型理解用户的自然语言输入，结合构建的煤矿行业向量数据库进行精准的知识检索与生成，支持上下文理解及复杂查询方式。

2　LoRA 微调与 RAG 增强融合技术

LoRA 微调与 RAG 增强融合技术路线如图 2 所示。

2.1　煤矿历史数据 LoRA 微调技术

煤矿历史数据 LoRA 微调技术流程如图 3 所示，分为数据集构建、LoRA 微调与模型量化 3 个部分。

1）数据集构建。首先修复煤炭行业知识资料中的文本编码错误，利用 BeautifulSoup 解析 HTML/XML，提取出纯文本，并借助 spaCy 完成分词、分句和实体识别，使文本结构保持一致。其次根据固定窗口大小对清洗后的文本进行截断，确保上下文完整性。然后构建 Alpaca 格式模板，将任务指令“instruction”、截断文本“input”和预期输出“output”填充为结构化数据。最后将生成的 Alpaca 格式数据存储为 JSON 文件格式并存储于 MySQL 数据库中，

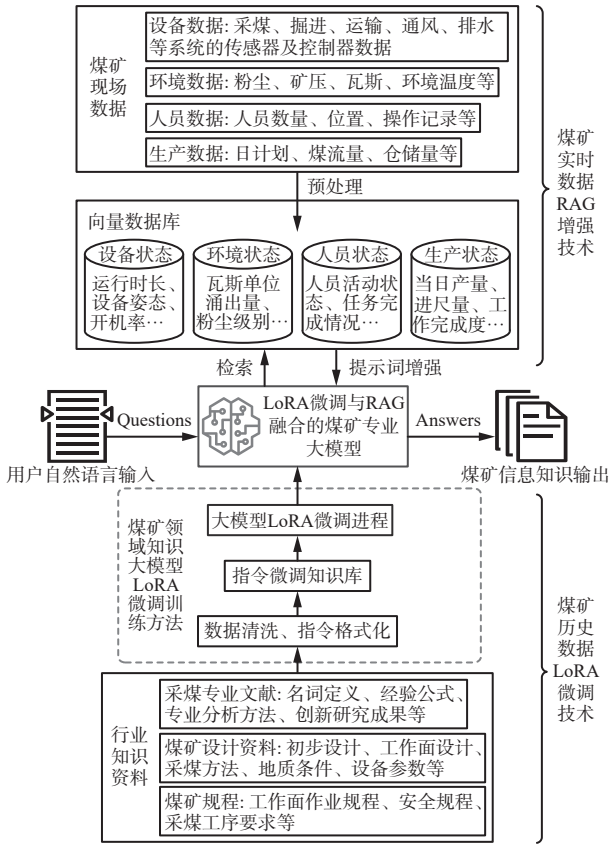


图 2　LoRA 微调与 RAG 融合技术路线

Fig. 2　Technical route of LoRA fine-tuning and RAG fusion

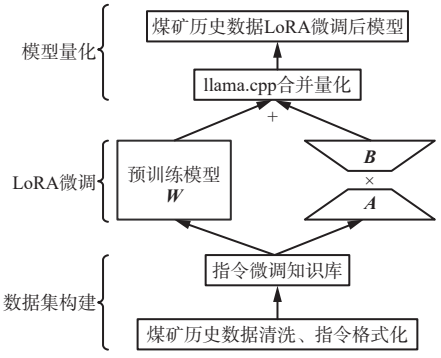


图 3　煤矿历史数据 LoRA 微调技术流程

Fig. 3　Technical flow of LoRA fine-tuning for historical coal mine data

构建指令微调知识库。

2）LoRA 微调。预训练模型的参数虽然庞大，但在解决具体任务时，真正起作用的参数可能极少，LoRA 技术在大模型上能以更低的秩实现与小模型较高秩相当的性能，同时大幅减少可训练参数：通过冻结预训练模型的权重，并在 Transformer 架构的每一层注入可训练的秩分解矩阵，只训练降维矩阵 $\boldsymbol{A}$ 与升维矩阵 $\boldsymbol{B}$，而模型的输入输出维度不变，输出时将 $\boldsymbol{BA}$ 与预训练模型的参数叠加合并，在保证微调模型性能的基础上，大大降低训练参数与 GPU 内存使用。

$$\boldsymbol{h}=\boldsymbol{W}_0x+\Delta\boldsymbol{W}x=\boldsymbol{W}_0x+\boldsymbol{BA}x,\boldsymbol{B}\in\mathbf{R}^{d\times r},\boldsymbol{A}\in\mathbf{R}^{r\times k}\quad(1)$$

式中：$\boldsymbol{h}$ 为输入 x 通过微调后模型的输出向量；$\boldsymbol{W}_0$ 为预训练权重矩阵；$\Delta\boldsymbol{W}$ 为需要更新的参数，即 LoRA 需要微调的部分；d 为 $\boldsymbol{W}_0$ 的输出特征维度；r 为低秩矩阵的秩，$r\ll\min(d,k)$；k 为 $\boldsymbol{W}_0$ 的输入特征维度。

3）模型量化。模型量化通过 llama.cpp 统一完成。量化位数用 q 表示，即权重存储精度；llama.cpp 提供 q_2, q_3, q_4 等多种档位。q 越大，精度损失越小，回答质量越接近原始模型，但磁盘占用与算力需求也随之增加。综合评估后，应在 q_3 或 q_4 中选择最贴近煤矿业务场景的量化参数，从而在成本可控的前提下，最大化微调模型的整体效益。

通过 Python 下载基础大模型 deepseekR1−7b，将指令微调知识库中的数据集内容按大模型的对话格式进行转换处理，并输入到设置好的 batch_size, learning_rate, lora_alpha, lora_dropout 等 LoRA 训练参数中进行微调训练，使用基于计算统一设备架构（Compute Unified Device Architecture，CUDA）的 GPU 进行训练，其性能利用率达到了 100%，微调完成后输出信息显示预训练模型参数共计 75.2 亿个，需要训练的参数约为 2 100 万个，运用 LoRA 技术微调了其中 0.28% 的参数。之后运用 llama.cpp 中 convert_hf_to_gguf.py 脚本进行模型的合并操作，完成合并后运用 q_{4_0} 量化对模型进行量化，结果表明量化后模型较微调合并后模型缩小了 70%，节省了 10 GiB 的存储空间。

2.2　煤矿实时数据 RAG 增强技术

煤矿实时数据 RAG 增强技术的核心是检索、生成相关信息来辅助微调后大模型生成更准确、更丰富的文本内容。检索主要是利用煤矿向量数据库的高效存储和检索能力，召回目标知识；生成是利用微调后大模型和 Prompt 工程，将召回的知识合理利用，生成目标答案。煤矿实时数据 RAG 增强技术流程如图 4 所示。

1）数据处理。针对煤矿设备现场数据包含设备 ID、时间戳（结构化数据）及设备传感器运行状态的数据向量表示（非结构化数据）的多源异构特性，提出一种双层索引优化方法，旨在解决结构化元数据与高维向量数据的协同检索难题。该方法首先将设备 ID 和时间戳作为核心元数据进行结构化存储，其中设备 ID 采用唯一标志符编码体系确保全局唯一性，时间戳统一转换为标准化时间格式以支持时间维度的语义解析。对于多传感器采集的设备运行状态数据，通过特征提取与降维处理构建 128～512

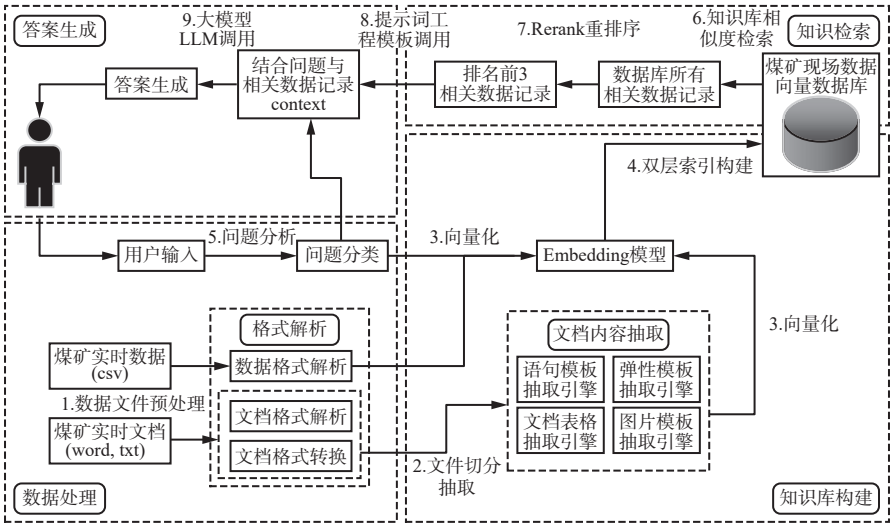


图 4　煤矿实时数据 RAG 增强技术流程

Fig. 4　Technical flow of RAG for real-time coal mine data

维的实值向量表示，并采用列存储方式构建高维向量数据库，每个向量样本通过主键与对应元数据建立关联，最终将不同来源的数据统一进行索引和检索。

2）知识库构建。应用 Milvus 向量数据库构建煤矿实时数据知识库，主要包括文档加载、切分、向量化及索引入库 4 个模块，以实现煤矿实时数据的高效存储。首先将实时采集的煤矿数据（如液压支架顶板压力数据）与最新文件通过拉取的 Embedding 模型（BAAl/bge−large−zh−v1.5）进行切分分块处理，生成语义向量表示，然后将向量化后的数据结合结构化元数据（如支架 ID、时间戳等）构建混合索引，最后写入煤矿向量数据库（Milvus），支持高效的多模态查询。

3）知识检索。构建基于分层可导航小世界图（Hierarchical Navigable Small World, HNSW）的近似最近邻索引层和基于倒排索引的结构化过滤层双层索引体系。

当接收到用户包含结构化条件与向量相似度的复合查询时，首先通过结构化过滤层进行元数据过滤，利用设备 ID 倒排表快速定位目标设备，结合时间戳 B+树索引实现时间范围的精确扫描，将检索范围缩小至特定设备的历史数据子集；然后将过滤后的向量子集输入近似最近邻索引层进行相似性搜索，通过多层图结构的导航优化，实现高维空间中的近似最近邻检索。这种先结构化过滤、后向量检索的处理流水线，通过结构化元数据的语义约束有效降低向量检索空间的复杂度，实现异构数据的高效协同检索，如图 5 所示。

在知识检索后，仅依赖向量检索或关键词匹配的初始排序，将导致候选集中混入大量低相关、重复

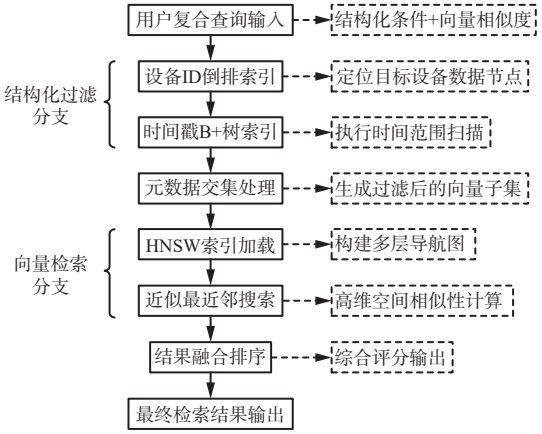


图 5　检索优化策略

Fig. 5　Retrieval optimization strategy

或逻辑冲突的信息，关键知识片段可能因排序靠后而被排除在模型输入范围外，直接影响大模型生成内容的准确性与完整性。为进一步提升检索精度，引入重排序 Rerank 模型（BAAl/bge−reranker−v2−m3）对初始召回结果进行重排精细化处理，以提升检索环节的召回率与排序效果，如图 6 所示。首先，融合检索结果与输入问题的语义相似度值，利用 Rerank 模型对初始召回的文档进行二次语义匹配打分，精准捕捉深层语义关联。然后，结合知识库类型进行分层排序，在召回阶段将信息按业务知识库、技术文档库、实时数据库等类型分组，每组内再依据相似度值从高到低排列，形成类型聚类−相似度优先的有序候选集。这种结构化重排策略确保大模型在生成 SQL 时，能够优先获取同类型知识库中相关性最高的信息，例如优先调用业务知识库中与查询字段匹配的表结构定义、技术文档库中对应的 SQL 语句规则、实时数据库中相似场景的查询模板，从而为模型生成逻辑清晰、语法准确、贴合业务需求的 SQL 语

句提供更有序、更聚焦的上下文输入，从根本上提升复杂查询场景下的生成质量与效率。

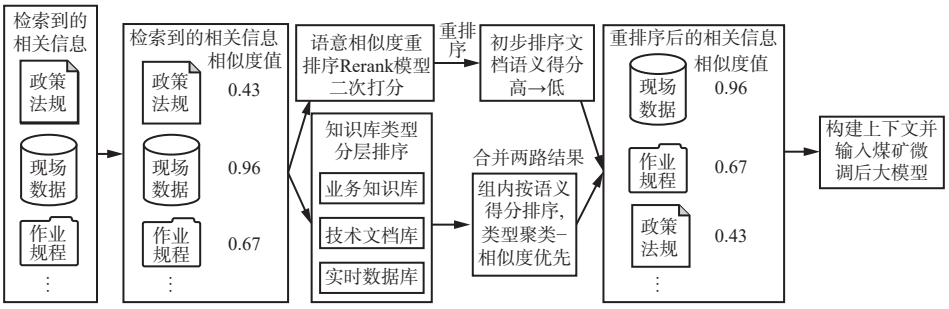


图 6　双重维度优化重排序机制

Fig. 6　Dual-dimension optimization of re-ranking mechanism

4）答案生成。提示词（Prompt）设计作为激活大模型答案生成潜能的关键媒介，通过融合知识库召回信息与用户问题构建语义增强 context 输入。因此，在设计煤矿设备现场数据任务的提示词模板时，引导大模型实现时空维度精准定位能力、设备关联关系理解能力、多系统数据融合分析能力与复杂逻辑分层推理能力 4 点核心回答能力，优化模型对问题的理解能力，并提高其生成回答的准确率。

时空维度精准定位能力通过提示词明确生产时间（如班次、日期等）和设备空间位置（如综采工作面区域、巷道编号等），使大模型能理解并生成包含时空过滤条件的查询。例如，当用户询问“早班 A 区的设备运行数据”时，大模型需自动关联“早班”的时间范围（如 6:00—14:00）和“A 区”的地理区域，精准筛选对应时段、地点的设备数据，避免跨时空的信息混淆。

设备关联关系理解能力是指煤矿设备常以系统形式协同工作，提示词需引导大模型理解设备拓扑结构（如“主设备→子设备”的层级关系）。当用户提问“采煤机故障影响哪些区域”时，大模型需通过设备关联关系，推断出受影响的子设备或下游区域，生成关联分析的查询逻辑，而非孤立查询单一设备。

多系统数据融合分析能力是指煤矿数据涉及安全监控、人员定位、设备管理等多个系统，提示词需引导模型跨系统关联数据。例如，用户查询“瓦斯超限期间的人员分布”时，模型需自动关联安全监控系统的瓦斯数据与人员定位系统的位置信息，生成多表联合查询，实现“数据孤岛”的融合分析，而非局限于单一系统的数据调用。

复杂逻辑分层推理能力是指对于多层级查询需求（如“先统计设备运行时长，再筛选超限设备，最后分析故障规律”），提示词需引导模型拆解问题为多个子步骤，按逻辑顺序生成嵌套查询。例如，用户提问“运行超过 500 h 设备的故障频率”时，模型需先过滤出符合时长条件的设备，再统计其故障次数，最终输出分层分析结果，避免直接返回未处理的原始数据。

本文设计的一个提示词模板示例如下。

question = "查询 2023−04−24T09:40:55 时 89 号支架与 59 号支架的行程均值，并分析是否存在安全隐患情况"

检索召回的表结构信息

tables_info = """

1. equipment_monitoring（设备监测表）

- equipment_id（设备编号）

- monitor_time（监测时间）

- stroke（行程值）

- status_code（状态码）

2. safety_threshold（安全阈值表）

- parameter_name（参数名称）

- min_value（最小值）

- max_value（最大值）

- safety_status（安全状态描述）

3. equipment_info（设备信息表）

- equipment_id（设备编号）

- equipment_type（设备类型）

"""

检索召回的专业名词解释

context_info = """

- 支架行程: 液压支架的伸缩距离，反映顶板支护状态

- 安全隐患判定规则:当支架行程均值超出安全阈值表中 stroke 参数的范围（1.2 m≤store≤1.8 m）时，存在安全隐患

"""

响应指南

response_guidelines = """

1. 如果提供的上下文足够，请生成一个有效的 SQL 查询，不需要对问题进行任何解释。

2. 如果提供的上下文几乎足够，但需要知道特定列中的特定字符串，请生成一个中间 SQL 查询来找到该列中不同的字符串。在查询

前加上注释 intermediate_sql。

3. 如果提供的上下文不充分，请解释为什么不能生成。

4. 请使用最相关的表。

5. 如果问题已经被问过并回答过，请完全重复之前给出的答案。

"""

生成最终 Prompt

prompt = generate_coal_mine_prompt(question, tables_info, context_info, response_guidelines)

print(prompt)

3 技术测试

为测试基于 LoRA 微调与 RAG 融合的煤矿专业大模型的可行性，在一台独立搭载 Windows 操作系统的工作站上进行大模型的微调部署与数据库构建。该工作站配备 24 核心 i9−13900k 处理器、64 GiB 运存、6 TiB 内存和 RTX4090 显卡。

3.1 大模型 LoRA 微调前后应用测试

LoRA 微调前后大模型应用效果如图 7 所示。微调前，大模型回答仅能提供排放瓦斯安全措施的通用原则，像加强通风、监测浓度等宽泛内容，无法针对某一煤矿的特定文件作答；而 LoRA 微调后，回答可以精准契合该煤矿“一通三防”管理制度汇编，不仅详细阐述了控制瓦斯排放的增阻限风、分风限风、逐段排放等具体方法，还对排放时间计算、传感器设置、图纸绘制及断电撤人等操作细则进行说明，实现了从泛泛而谈到精准定位具体煤矿特定文件内容的跨越。

3.2 煤矿设备数据写入与查询测试

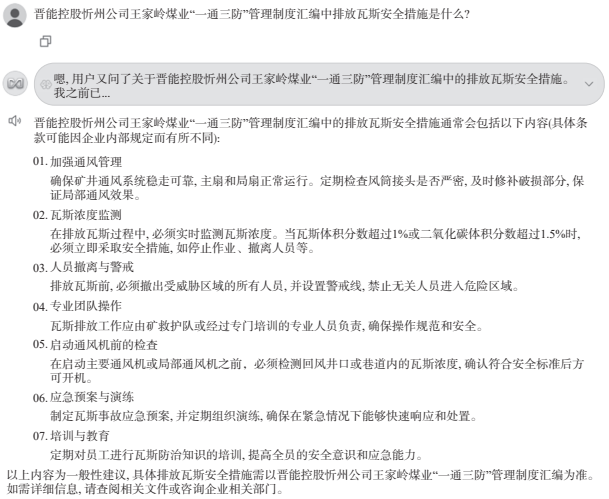
选取现场 143 万条液压支架时序数据（含压力、位移量、设备 ID、时间戳等），分别存入 Milvus 向量数据库与 MySQL 关系型数据库，从写入效率与查询性能 2 个维度进行对比，结果见表 1和表 2。

由表 1 可看出，Milvus 向量数据库的写入速度达到 MySQL 的 2.4 倍，向量索引构建效率显著；而 MySQL 受限于行式存储和事务日志带来的额外输入输出操作，写入性能相对较低。

由表 2 可看出，在向量检索场景方面，Milvus 的向量相似度检索延迟稳定在 20 ms，而 MySQL 无法原生支持向量搜索；在混合查询场景效率方面，MySQL 需全表扫描后排序，143 万条数据延迟超 100 ms，而 Milvus 将设备 ID 过滤后的子集输入 HNSW 层，仅读取查询涉及的向量字段，避免全表扫描，查询速率明显优于传统数据库。

3.3 煤矿专业大模型应用测试

将本地基于 LoRA 微调与 RAG 融合的煤矿专业大模型与本地离线 deepseekR1−7b 模型系统进行部署，并以 Web 与 AnythingLLM 作为前端展示，对多项指标进行测试，以验证 LoRA 微调与 RAG 融合的煤矿专业大模型的专业性与技术有效性。其中使用 10 人 20 轮问答分级打分评价指标，将模型面对不同类型问题的回答分为 4 个等级，4 分代表完全符合标准，3 分代表基本满足但有小瑕疵，2 分代表有明显



（a）LoRA微调前

（b）LoRA微调后

图 7　LoRA 微调前后大模型后应用效果

Fig. 7　Application effects of large model before and after LoRA fine-tuning

表 1　数据写入性能对比

Table 1　Data write performance comparison

指标	Milvus向量数据库	MySQL数据库
写入速度/(条·s^{-1})	12 000	5 000
写入耗时/s	119	286
资源占用	CPU 65%(内存18 GiB)	CPU 85%(内存9 GiB)

表 2　数据查询性能对比

Table 2　Data query performance comparison

查询类型	Milvus向量数据库	MySQL数据库
2023−04−24 T09:40:52 第53号支架行程	架号: 53, 最大行程617.00 mm, 最小行程42.00 mm	—
向量相似度搜索	20 ms	不支持
混合查询(设备ID+时间范围+向量)	≈119 s	≈286 s

不足但大意清晰，1 分代表完全不符合要求，满分 40 分，20 轮得分取均值。打分与应用测试结果见表 3，智能问答对比结果见表 4。

从表 3、表 4 可看出，使用历史数据 LoRa 微调与实时数据 RAG 增强后的大模型在煤矿专业领域知识学习性、知识动态化更新时效性、模型泛化与回答精确度方面具有显著优势，为工业级 AI 落地提供了可行路径。

表 3　打分与应用测试结果

Table 3　Scoring and application test results

指标	通用大模型	煤矿垂域大模型	提升幅度
现场数据问答	12.5分	31.5分	+19分
专业知识问答	28分	34分	+6分
Token	20 token/s	24 token/s	+4 token/s
知识更新	无更新	实时更新	显著提升

表 4　智能问答对比结果

Table 4　Comparison results of intelligent question-answer

问题（用户）	通用大模型（deepseekR1−7b）	煤矿专业大模型
问题1：3101工作面2023−04−24T 09:40:52—09:50:20第66号支架的平均压力是多少？	由于无法实时访问数据，请通过上述步骤查询系统或联系监测部门获取具体数值。若数据正常且无异常波动，平均值通常反映支架的稳定状态，需结合安全标准进一步分析	根据3101工作面支架压力监测系统数据，2023−04−24T 09:40:52—09:50:20第66号支架的平均压力为30 MPa（数据采样间隔10 s，共56个有效数据点计算得出）。该压力值处于该支架额定工作压力（35　MPa）的85.7%范围内，符合工作面正常矿压显现规律
问题2：行驶5台及以上胶轮车时，应装备车辆位置监测系统，并符合哪些规定？	具体以当地安全生产标准或行业规范为准： 1）　系统功能要求：系统应能实时监测和显示胶轮车的位置、运行状态及速度等信息。具备车辆调度管理功能，支持对多辆胶轮车的统一监控和协调 2）　数据精度：位置监测的精度应满足实际需求，通常应在米级范围内，确保准确掌握车辆分布情况	1）无轨胶轮车运输井筒、辅运大巷、采区辅运巷等主要运输线路应具备车辆位置监测功能，能够反映运行区域车辆的位置状态。在采煤工作面两巷道口和掘进工作面巷口应设置定位分站，能够确认车辆进出采掘工作面的状态 2）根据车辆车牌号对车辆进行编码，对应车辆车牌号实行“一车一卡”或“一车一接收器”，保证每台车辆信息在监测系统中的唯一性

4　结论

1）煤矿专业大模型垂域应用技术架构包括多模态数据库层、大模型主动洞察业务逻辑层和自然语言交互应用层。其中多模态数据库层通过指令微调知识库与向量数据库解决了煤矿行业多源异构数据的统一管理与高效检索问题，实现了大模型对于数据的高效整合与利用；大模型主动洞察业务逻辑层通过预训练大模型的 LoRA 微调与 RAG 技术，解决了煤矿行业复杂业务场景下的模型泛化能力不足与知识利用率低的问题，提升了大模型对于煤矿专业领域智能化分析与决策能力；自然语言交互应用层解决了煤矿行业人机交互效率低、知识获取困难的问题，通过自然语言理解与生成技术，提供了直观、高效的智能化服务支持。

2）通过实验测试，LoRA 微调与 RAG 融合的专业模型在专业响应精度、实时数据分析及知识更新能力上实现质的突破，验证了该技术可实现煤矿行业大模型多源数据的智能整合、高效分析与精准检索推理，显著提升大模型在煤矿行业的专业应用水平和知识利用率。

参考文献（References）：

[1] 戴春雷. 我国矿山智能化研究进展及大模型应用前景[J]. 工矿自动化，2024，50（增刊 2）：1-11，22.
DAI Chunlei. Progress of intelligent mine research and the prospects for mine large model applications in China[J]. Journal of Mine Automation，2024，50（S2）：1-11，22.

[2] 付翔，秦一凡，李浩杰，等. 新一代智能煤矿人工智能赋能技术研究综述[J]. 工矿自动化，2023，49（9）：122-131，139.
FU Xiang，QIN Yifan，LI Haojie，et al. Summary of research on artificial intelligence empowerment technology for new generation intelligent coal mine[J]. Journal of Mine Automation，2023，49（9）：122-131，139.

[3] 刘具，秦坤. 我国煤炭绿色开采技术进展[J]. 矿业安全与环保，2023，50（6）：7-15.
LIU Ju，QIN Kun. Progress of green coal mining technology in China[J]. Mining Safety & Environmental Protection，2023，50（6）：7-15.

[4] 王国法. 煤矿智能化最新技术进展与问题探讨[J]. 煤炭科学技术，2022，50（1）：1-27.
WANG Guofa. New technological progress of coal mine intelligence and its problems[J]. Coal Science and Technology，2022，50（1）：1-27.

[5] 王国法，庞义辉，任怀伟. 智慧矿山技术体系研究与发展路径[J]. 金属矿山，2022（5）：1-9.
WANG Guofa，PANG Yihui，REN Huaiwei. Research and development path of smart mine technology system[J]. Metal Mine，2022（5）：1-9.

[6] 冯杨洋，汪庆，谢旻晖，等. 从 BERT 到 ChatGPT：大模型训练中的存储系统挑战与技术发展[J]. 计算机研究与发展，2024，61（4）：809-823.
FENG Yangyang，WANG Qing，XIE Minhui，et al.

From BERT to ChatGPT: challenges and technical development of storage systems for large model training[J]. Journal of Computer Research and Development, 2024, 61(4): 809-823.

[7] 张浩然，李君，邢立宁，等. 大模型与智能优化算法集成研究综述[J/OL]. 控制与决策：1-20[2025-08-03]. https://doi.org/10.13195/j.kzyjc.2025.0121.
ZHANG Haoran, LI Jun, XING Lining, et al. A research review on the integration of large models and intelligentoptimization algorithms[J/OL]. Control and Decision: 1-20[2025-08-03]. https://doi.org/10.13195/j.kzyjc.2025.0121.

[8] 王睿，张留洋，高志涌，等. 面向边缘智能的大模型研究进展[J/OL]. 计算机研究与发展：1-18[2025-02-13]. https://link.cnki.net/urlid/11.1777.TP.20250127.0921.006.
WANG Rui, ZHANG Liuyang, GAO Zhiyong, et al. Research progress on large models for edge intelligence[J/OL]. Journal of Computer Research and Development: 1-18[2025-02-13]. https://link.cnki.net/urlid/11.1777.TP.20250127.0921.006.

[9] 赵浜，曹树金. 生成式 AI 大模型结合知识库与 AI Agent 开展知识挖掘的探析[J/OL]. 图书情报知识：1-14[2025-02-25]. http://kns.cnki.net/kcms/detail/42.1085.G2.20241103.1003.002.html.
ZHAO Bang, CAO Shujin. Exploring knowledge mining using large language model combinedwith knowledge base and AI agent[J/OL]. Documentation, Information & Knowledge: 1-14[2025-02-25]. http://kns.cnki.net/kcms/detail/42.1085.G2.20241103.1003.002.html.

[10] 任海玉，刘建平，王健，等. 基于大语言模型的智能问答系统研究综述[J]. 计算机工程与应用，2025，61(7)：1-24.
REN Haiyu, LIU Jianping, WANG Jian, et al. Research on intelligent question answering system based on large language model[J]. Computer Engineering and Applications, 2025, 61(7): 1-24.

[11] 江耀森，杨超宇，刘晓蕾. 基于 ChatGLM3−6B 的煤矿事故知识图谱构建及应用研究[J]. 中国安全生产科学技术，2024，20(10)：12-21.
JIANG Yaosen, YANG Chaoyu, LIU Xiaolei. Research on construction and application of knowledge graph for coal mine accidents based on ChatGLM3-6B[J]. Journal of Safety Science and Technology, 2024, 20(10): 12-21.

[12] 许洁，龚青，项逸达，等. 一种基于 LLM 的煤矿安全助手的设计与应用[J]. 内蒙古煤炭经济，2024(1)：133-135.
XU Jie, GONG Qing, XIANG Yida, et al. Design and application of a coal mine safety assistant based on LLM[J]. Inner Mongolia Coal Economy, 2024(1): 133-135.

[13] 王海玲，刘兴丽，邵宗曦，等. 基于 ChatGLM 的煤矿安全预警知识抽取及应用[J]. 煤炭技术，2024，43(8)：219-221.
WANG Hailing, LIU Xingli, SHAO Zongxi, et al. Knowledge extraction and application of coal mine safety early warning based on ChatGLM[J]. Coal Technology, 2024, 43(8): 219-221.

[14] 王永志，温世博，李博文，等. 大模型驱动的矿产资源智能预测超级智能体构建方法探索[J]. 地学前缘，2025，32(4)：38-45.
WANG Yongzhi, WEN Shibo, LI Bowen, et al. Construction technology of super-agents for intelligent mineral resources prediction driven by large model[J]. Earth Science Frontiers, 2025, 32(4): 38-45.

[15] 戴宗超，李杰，张扬，等. 基于生成式特征图的深度学习模型局部性能可视分析方法[J/OL]. 计算机辅助设计与图形学报：1-17[2025-02-13]. https://link.cnki.net/urlid/11.2925.tp.20241029.1354.029.
DAI Zongchao, LI Jie, ZHANG Yang, et al. Visual analysis method for local performance of deep learning models based on generative feature maps[J/OL]. Journal of Computer-Aided Design & Computer Graphics: 1-17[2025-02-13]. https://link.cnki.net/urlid/11.2925.tp.20241029.1354.029.

[16] SUN Yi, JI Faxiu. An embodied intelligence system for coal mine safety assessment based on multi-level large language models[J]. Sensors, 2025, 25(2). DOI: 10.3390/s25020488.

[17] 籍欣萌，咎红英，崔婷婷，等. 大模型在垂直领域应用的现状与挑战[J]. 计算机工程与应用，2025，61(12)：1-11.
JI Xinmeng, ZAN Hongying, CUI Tingting, et al. Status and challenges of large language models applications in vertical domains[J]. Computer Engineering and Applications, 2025, 61(12): 1-11.

[18] 余娜. 煤矿行业首个 AI 大模型商用落地[N]. 中国工业报，2023-07-25(第 3 版).
YU Na. The first AI large model commercial landing in the coal mining industry[N]. China Industry News, 2023-07-25(3th ed.).

[19] 朱俊仪，朱尚明. 利用检索增强生成技术开发本地知识库应用[J]. 通信学报，2024，45(增刊 2)：242-247.
ZHU Junyi, ZHU Shangming. Development of local knowledge base application using retrieval augmented generation technology[J]. Journal on Communications, 2024, 45(S2): 242-247.

[20] 张浩然，郝文宁，靳大尉，等. DF−RAG：基于查询重写和知识选择的检索增强生成方法[J/OL]. 计算机科学：1-12[2025-02-25]. http://kns.cnki.net/kcms/detail/50.1075.TP.20250226.1646.002.html.
ZHANG Haoran, HAO Wenning, JIN Dawei, et al. DF-RAG: a retrieval-augmented generation method based on query rewriting and knowledge selection[J/OL]. Computer Science: 1-12[2025-02-25]. http://kns.cnki.net/kcms/detail/50.1075.TP.20250226.1646.002.html.

（下转第 50 页）