矿物岩石地球化学通报
Bulletin of Mineralogy, Petrology and Geochemistry
Vol.44 No.3, May, 2025

大语言模型及其在矿物问答系统中的应用

季晓慧[1],刘成健[1],杨眉[2],何明跃[2],张招崇[3,4],曾姗[1],王玉柱[1]

1. 中国地质大学(北京) 信息工程学院, 北京 100083; 2. 中国地质大学(北京) 国家岩矿化石标本资源库, 北京 100083;
3. 中国地质大学(北京) 深时数字地球前沿科学中心, 北京 100083; 4. 中国地质大学(北京) 地质过程与矿产资源国家重点实验室, 北京 100083

摘　要: 大语言模型(LLMs,Large Language Models)具有极强的自然语言理解和复杂问题求解能力,本文基于大语言模型构建了矿物问答系统,以高效地获取矿物知识。该系统首先从互联网资源获取矿物数据,清洗后将矿物数据结构化为矿物文档和问答对;将矿物文档经过格式转换和建立索引后转化为矿物知识库,用于检索增强大语言模型生成,问答对用于微调大语言模型。使用矿物知识库检索增强大语言模型生成时,采用先召回再精排的两级检索模式,以获得更好的大语言模型生成结果。矿物大语言模型微调采用了主流的低秩适配(Low-Rank Adaption,LoRA)方法,以较少的训练参数获得了与全参微调性能相当的效果,节省了计算资源。实验结果表明,基于检索增强生成的大语言模型的矿物问答系统能以较高的准确率快捷地获取矿物知识。

关　键　词: 大语言模型;矿物;检索增强生成;低秩适配;问答系统

中图分类号: TP391.7;P579　**文章编号:** 1007-2802(2025)03-0453-09　**doi:** 10.3724/j.issn.1007-2802.20240155

Large language model and its application in the questioning-answering system of minerals

JI Xiao-hui[1], LIU Cheng-jian[1], YANG Mei[2], HE Ming-yue[2], ZHANG Zhao-chong[3,4], ZENG Shan[1], WANG Yu-zhu[1]

1. *School of Information Engineering, China University of Geosciences(Beijing), Beijing* 100083*, China;*
2. *National Mineral Rock and Fossil Specimens Resource Center from MOST, China University of Geosciences(Beijing), Beijing* 100083*, China;*
3. *Frontiers Science Center for Deep-time Digital Earth, China University of Geosciences (Beijing), Beijing* 100083*, China;*
4. *State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences(Beijing), Beijing* 100083*, China*

Abstract: Large Language Models (LLMs) have strong capabilities for understanding natural languages and solving complex problems. This article constructs a questioning-answering system of minerals based on the LLM in order to efficiently acquire mineral knowledge. In the system, data of minerals were firstly obtained from Internet resources, and then they were structured to mineral documents and Q&A pairs after the data cleaning. The mineral knowledge base formed through the format conversion and indexing of mineral documents is used for the retrieval-augmented generation of an large language model. The questioning-answering pairs are used for the fine-tuning of the large language model. When the mineral knowledge base retrieval is used to enhance the generation of large language model, a two-level retrieval mode of first recalling and then refining is adopted in order to obtain better results of the large language model generation. The mainstream Low-Rank Adaptation (LoRA) method is used to fine-tune the large language model of minerals, in order to achieve comparable performance of the full parameter fine-tuning by using relatively small numbers of training parameters and to save computational resources. The experimental results show that the questioning-answering system of minerals based on the retrieval-augmented generation of large language models can quickly and

收稿编号:2024-0206,2024-12-06 收到,2025-01-05 改回
基金项目:国家科技资源共享服务平台——国家岩矿化石标本资源库子项目(NCSTI-RMF20240109)
第一作者简介:季晓慧(1977—),女,博士,副教授,博士生导师,研究方向:人工智能及其应用. email:xhji@cugb.edu.cn

引用此文:
季晓慧,刘成健,杨眉,何明跃,张招崇,曾姗,王玉柱. 2025. 大语言模型及其在矿物问答系统中的应用. 矿物岩石地球化学通报,44(3):453–461
Ji X H, Liu C J, Yang M, He M Y, Zhang Z C, Zeng S, Wang Y Z. 2025. Large language model and its application in the questioning-answering system of minerals. Bulletin of Mineralogy, Petrology and Geochemistry, 44(3):453–461

accurately obtain mineral knowledge.

Key words: large language model; mineral; retrieval-augmented generation; Low-Rank Adaptation; the questioning-answering system

0 引言

大语言模型(Large Language Models,LLMs)是指在海量文本语料上训练、包含至少数十亿参数、具有极强自然语言理解和复杂问题求解的小语言模型不具备的“涌现”能力的神经网络模型(Zhao et al.,2023)。大语言模型的“涌现”能力可类比物理学中的“相变”,具有卓越性能(Zhao et al.,2023;Shanahan,2024)。如GPT系列(Radford et al.,2018;Radford et al.,2019;Brown et al.,2020;Dai et al.,2023;Ouyang et al.,2022;OpenAI,2022,2023)、LLaMa家族(Touvron et al.,2023a,2023b;Tworkowski et al.,2023)和GLM系列(Du et al.,2022;Zeng et al.,2022)等。众多垂直领域已使用LLMs推动其相关研究(Zhao et al.,2023;Shanahan et al.,2024),在知识问答上包括医疗领域的Med-PaLM2(Singhal et al.,2023)、中国法律中的LawGPT(Zhou et al.,2024)、生物学领域的LucaOne(He et al.,2024)、化学领域的ChemDFM(Zhao et al.,2024),以及地理空间领域的GeoGPT(Zhang et al.,2023)等。

矿物知识在地球科学中的重要性已被广泛认可(周永章等,2021a;周永章等,2021b;Liu et al.,2023;季晓慧等,2024;周永章等,2024)。传统上,由美国地质调查局、英国地质调查局、国际矿物学协会和国家岩矿化石标本资源库等机构提供的专业矿物数据库是矿物知识的主要来源。然而,这些数据库在灵活响应自然语言查询方面存在局限,并且它们往往只专注于特定地区、国家或矿物类型,需要使用多个数据库来获取全面信息,因此耗时长。谷歌、必应和百度等搜索引擎虽然可以进行自然语言的矿物知识问答,但用户仍需在搜索结果中筛选以找到答案,也十分耗时。

知识图谱具有强大的语义处理能力,已应用于问答系统并具有良好性能(周永章等,2021a,2021b;Liu et al.,2023;季晓慧等,2024;周永章等,2024)。为进行快捷矿物知识问答,也已出现基于知识图谱问答(Knowledge Graph Question Answering,KGQA)的矿物知识系统:通过大量三元组<实体,关系,实体>来表示矿物知识,并基于预定义的模板进行推理(Liu et al.,2023;季晓慧等,2024),如Liu等(2023)通过单个三元组推理进行矿物知识问答,季晓慧等(2024)在此基础上采用多跳推理进行矿物知识问答,但基于KGQA的矿物知识问答方法在问题的语言结构复杂或具有上下文关联时会不准确。此外,由于KGQA依赖于显式的实体、关系三元组进行推理,所以不能对超出其预定义模板的矿物知识进行问答(Liu et al.,2023;季晓慧等,2024)。

鉴于上述现状,本文提出使用检索增强生成(Retrieval-augmented Generation,RAG)的大语言模型来设计和实现矿物问答系统。由于大语言模型能回答任意自然语言形式的问题,因此本文提出的方法可以避免已有矿物数据库或搜索引擎以及基于预定义模板知识图谱的矿物问答系统只能回答特定问题的限制,同时也避免了使用搜索引擎搜索答案时对给出的各参考网页进行答案筛选的过程,可见本文方法能更高效且全面地回答矿物相关问题。

1 大语言模型

从语言模型的发展历程来看,大语言模型处于第四阶段,前三个阶段分别为统计语言模型(Statistical Language Model,SLM)、神经语言模型(Neural Language Model,NLM)和预训练语言模型(Pre-trained Language Model,PLM)(Zhao et al.,2023)。

SLM兴起于20世纪90年代,其使用马尔可夫假设根据一个固定长度n的前缀来预测目标单词,因此通常被称为n元(n-gram)语言模型;NLM使用神经网络,如循环神经网络(Recurrent Neural Networks,RNN)来建模文本序列的生成,其使用低维稠密向量,即“词嵌入”(Word Embedding)来表示词汇语义,能够刻画丰富的隐含语义特征,且对于复杂语言模型的搭建非常友好,能够有效克服SLM中的数据稀疏问题,为自然语言处理提供了一种较为通用的解决途径;PLM以“预训练-微调”为范式,预训练通过大规模无标注文本建立模型的基础能力,微调使用有标注数据对于模型进行特定任务的适配,以更好地解决下游的自然语言处理任务。

大语言模型通过训练更大的PLM来扩展语言模型性能极限,在解决复杂任务时表现出与小型PLM不同的行为,通常被称为“涌现能力”(Emergent Abilities),使其不仅能进行自然语言建模和生成,也能求解复杂任务(Zhao et al.,2023)。

1.1　大语言模型的参数高效微调

大语言模型的知识由海量文本训练而来,存储于模型的大量参数中,因此大模型适用于各种自然语言处理任务。但是,将通用训练的大语言模型直接用于垂直领域,效果并不好,需要进一步使用垂直领域数据对大模型进行训练,需将垂直领域知识融合到大语言模型中,提升其垂直领域应用能力(Zhao et al.,2023;Shanahan et al.,2024)。然而,大语言模型的参数量巨大,进行全参数微调需要相当巨大的算力资源,因此采用参数高效微调(Parameter-Efficient Fine-tuning,PEFT)方式可以减少需要训练的模型参数,同时保证微调后的模型性能与全量微调的性能相当(Zhao et al.,2023)。

低秩适配(Low-Rank Adaptation,LoRA)是性能较好、使用广泛的PEFT方法(Zhao et al.,2023;Hu et al.,2021),如图1所示。微调时LoRA冻结经过通用数据预训练的大语言模型原始参数矩阵$W_0 \in \mathbb{R}^{H\times H}$,添加可训练的低秩分解矩阵$A \in \mathbb{R}^{H\times r}$和$B \in \mathbb{R}^{H\times r}$来近似参数更新矩阵$\Delta W = A \cdot B^{\mathrm{T}}$来适配下游任务,其中$r \ll H$,因而减少了所需要训练的参数。在前向传播过程中,中间状态$h = W_0 \cdot x$变为$h = W_0 \cdot x + A \cdot B^{\mathrm{T}} \cdot x$,训练完成后,原始参数矩阵$W_0$和训练得到的权重$A$和$B$合并得到微调之后的权重$W = W_0 + A \cdot B^{\mathrm{T}}$,为$H \times H$维。

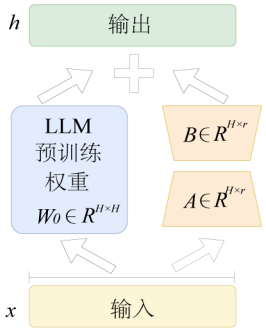

图1 基于LoRA的大语言模型参数高效微调

Fig.1 The efficient fine-tuning of parameters of LLMs based on the LoRA

1.2　大语言模型的检索增强生成

大语言模型虽然具有出色的自然语言生成能力,但在回答没有训练过的问题时往往给出其编造的错误回答,因此需要补充外部知识以缓解此“幻觉”问题(Gao et al.,2023;Zhao et al.,2023;Shanahan et al.,2024;Fan et al.,2024;Zhao et al.,2024)。检索增强生成(Retrieval-Augmented Generation,RAG)通过检索获取相关外部知识,将其融入提示(Prompt),使大语言模型通过参考检索出的上下文给出合理回答,即“检索+生成”,“检索”利用向量数据库的高效存储和检索能力,通过融合相似性检索和全文检索等多种检索方式召回问题相关的外部目标知识;“生成”则利用大语言模型的生成能力和提示工程,将检索召回的知识作为上下文提供给大语言模型,生成最终答案。RAG提高了问答的可解释性与可信度,并且外部知识以即插即用的方式显式引入,具备良好的可扩展性,同时可作为生成答案的参考依据。

1.3　大语言模型的提示工程

大语言模型推理时通过上文推测下一个词,即通过NTP(Next Token Prediction)完成相关任务(Zhao et al.,2023;Shanahan et al.,2024),为了取得较好的性能需要有足够和有效的上文,以做出准确预测。为此可通过设计特定的提示来“引导”模型的注意力和生成过程,以激发模型生成下游任务期望的输出,即Prompt工程。设计提示时需考虑任务描述、输入数据、上下文信息和提示策略等关键要素(Zhao et al.,2023)。其中上下文信息可以把检索增强时提供的参考文档以上下文的形式引入提示作为大语言模型的输入,任务示例数据也可以作为上下文以提升大语言模型处理复杂任务的能力,大语言模型通过示例数据学习任务目标、输出格式以及输入和输出之间的映射关系。提示策略通过添加特定的前缀或后缀来引导大语言模型解决复杂任务,例如,使用前缀“你是这项任务(或这个领域)的专家”可以提高大语言模型在某些特定任务(或领域)中的表现。还可以使用关键词搭配构建多种Prompt组合。

2　基于大语言模型的矿物知识问答

本文采用检索增强生成(Retrieval-Augmented Generation,RAG)(Gao et al.,2023;Fan et al.,2024;Zhao et al.,2024)大语言模型构建矿物问答系统,如图2所示。首先从百度百科、百度问答、维基百科和国家岩矿化石标本资源库等互联网资源爬取中文矿物数据,并进行去重、校正、格式标准化和过滤无关信息等清洗,清洗后将数据结构化为矿物文档和问答对。矿物文档经格式转换和建立索引后转化为矿物知识库,问答对用于微调开源大语言模型,即矿物问答系统由矿物知识库和矿物大语言模型组成。当用户提出一个矿物相关问题时,系统从知识库中检索出相关文档,并进行排序以选取出与问题最相关的知识文档,将排名前k的文档连同用户问题和专门设计的提示词一起输入到微调后的大语言模型中,生成答案。

2.1　参数高效微调获得矿物大语言模型

本文通过将矿物专业术语库添加到基础大语言模型ChatGLM(Du et al.,2022),并对其使用清洗后的

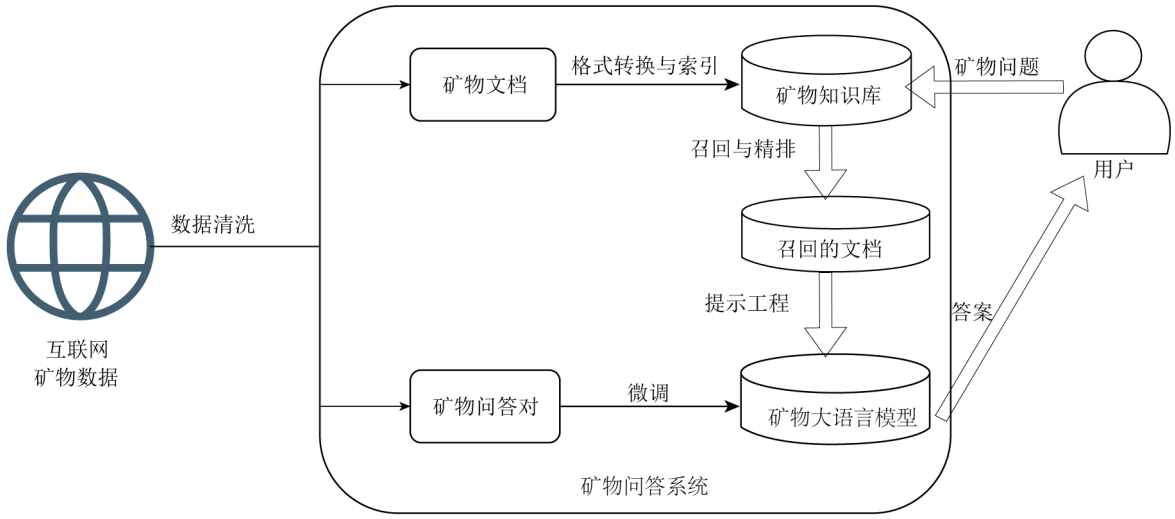

图2 基于检索增强生成大语言模型的矿物问答系统

Fig.2 The questioning-answering system of minerals based on the retrieval-augmented LLM

矿物问答对微调的方式构建矿物大语言模型，使其具有更丰富和专业的矿物问答知识。微调采用如图1所示的LoRA(Zhao et al.,2023;Hu et al.,2021)方法。微调期间，ChatGLM的预训练权重W_0保持不变，而适配矩阵A和B则在清洗后的矿物问答对上进行训练。训练开始时，A设置为一个随机高斯矩阵、B设置为零矩阵，使用交叉熵损失函数，并以DeepSpeed(Rasley et al.,2020)做为训练框架。微调后，矿物LLM能更好地理解复杂矿物语言。

2.2 矿物大语言模型的检索增强生成

为减轻矿物大语言模型生成答案时的“幻觉”问题，将互联网获取的矿物文档经清洗去除无关或多余信息后存储到Elastic Search(ES)搜索引擎(ES,2024)中作为矿物大语言模型的外部知识，存储时按表1所示的格式建立索引，以高效地进行矿物知识查询和检索。

矿物问答系统使用Jieba(2024)对用户问题进行分词、去停用词、提取关键词以及识别核心问题词后，ES使用BM25算法(Robertson et al.,2009)计算各文档与问题之间的相关性分数，得分高的文档被召回以输入大语言模型作为问题的上下文。但是由于大语言模型的输入长度有限，并非所有召回的文档都能输入到大语言模型中，因此本文对召回的矿物文档进行进一步的精细排序，以找出前k个(k<5)最相关的文档，这k个文档被输入到大语言模型中作为上下文。

表 1 矿物文档索引结构

Table 1 The indexing structure of mineral documents

索引字段	ES类型	说明
name	text	矿物中文名
English_name	text	矿物英文名
the_provider	text	提供者
web_url	text	网页链接
description	text	矿物的性质、成分、用途等各类详细信息
save_date	date	记录保存时间

为了找出前k个文档，使用问答对训练XGBoost(Chen et al.,2016)模型。除正样本外，通过对每个问题从其他问题答案中随机选择得到错误答案，形成负样本。问题与正确答案/文档之间的相关性分数为1，与错误答案/文档之间的相关性分数为0。训练时根据下述三类特征计算召回答案/文档与问题之间的相关性分数：(1)专业术语数量和数据源特征，用于衡量文档的专业性；(2)文本长度和单词数量特征，用于评估文档内知识的丰富性；(3)TF-IDF(term frequency-inverse document frequency，词频-逆文档频率)分数(TF-IDF,2021)、余弦相似度、杰卡德相似度、最小编辑距离和ES召回分数特征。

2.3 矿物大语言模型的提示工程

为了指导矿物大语言模型生成专用于矿物领域的内容，设计了以下5个专用提示：

(1)专家角色提示：为矿物大语言模型分配矿物专家角色；

(2)矿物问答任务描述：为矿物大语言模型提供操作指南，包括从矿物知识文档中提取答案、匹配用户查询、简洁地总结答案以及避免重复；

(3)加入矿物知识文档：加入从知识库中检索到的精细排序后的矿物文档答案，如[文档1，文档2，…，文档k]；

(4)领域特定种子词:设定“矿物”、“问答”和“检索增强生成”为种子词,将矿物大语言模型的输出聚焦于矿物领域;

(5)拒答:指导矿物LLM不要生成跑题或错误的答案、不创建与矿物无关的内容、在矿物知识文档不包含与用户查询相关的信息时,不要编造或猜测答案。

上述提示以自然语言形式写出,如矿物专家提示可写为“请以矿物专家身份回答下面问题”。可根据输出结果对提示的描述进行调整以得到更满意的结果。

系统支持多轮问答,在回答本轮问题时,将设计的提示词、所有问题、前一回合的答案输入到大语言模型中,以生成最相关的答案。

3 矿物大语言模型实验结果

基于大语言模型的矿物问答系统chatMineral部署在配备Tesla P100 GPU的服务器上,使用PyTorch 1.10.0版本作为深度学习框架,使用的关键库及其版本和用途如表2所示,使用的数据集及其数量和用途如表3所示。

3.1 矿物大语言模型的参数高效微调

系统所使用的基座模型为ChatGLM3-6B,在配备了Tesla V100 GPU的服务器上使用表3所示的矿物问答对,采用LoRA进行微调,微调训练参数如表4所示。

生成200个矿物问答对进行测试,测试自动进行,因此闭源模型如GPT4.0等未参与测试,与开源大语言模型T5(Arora et al.,2022)和GPT2(Radford et al.,2019)的对比如表5所示,其中准确率(Accuracy)为正确回答数量与总回答数量之比,正确回答指连贯且包含与对应问题相关的所有相关和准确信息的答案,如果一个回答缺乏连贯性、包含错误细节或冗余信息,则为不正确。

实验表明,经过矿物数据微调并针对中文语言处理进行优化的矿物大语言模型,在生成矿物领域内准确且连贯的答案方面表现更出色。

3.2 矿物大语言模型的检索增强生成

通过在百度问答数据上训练XGBoost获得相关性分数,来选择矿物知识库中检索出的前k个文档,按照8:1:1的比例将百度问答数据划分为训练集、验证集和测试集,训练XGBoost过程如图3所示。

为了评估训练后的XGBoost模型性能,将百度问答测试数据集中的问题输入到矿物知识库中,每个问题检索出97个相关文档(实验表明,97个文档可以覆盖95%的答案文档),XGBoost对这97个召回文档进行进一步精确排序。使用精确度(Precision)指标来评估XGBoost的性能,精确度(Precision)通过将前3个文档中正确答案/文档的数量除以召回文档的总数来计算。

表6显示,利用本研究提出的基于多个文档特征计算文档与问题的相关性分数的XGBoost模型达到了

表 2 矿物问答系统使用的相关软件库

Table 2 Software libraries used by the questioning-answering system of minerals

库	版本	用途
Elastic Search	7.14.0	存储矿物知识库
XGBoost	1.7.5	精排召回的矿物知识文档
Jieba	0.42.1	文本分析
PEFT	0.5.0	基于LoRA微调大模型

表 4 ChatGLM的微调参数

Table 4 The fine-tuned parameters of the ChatGLM

训练参数	值
Batch Size	32
Learning Rate	$1e^{-5}$
Epoch	3
Dropout	0.1
Optimizer	Adam
LoRA_rank	8
LoRA_alpha	16
beam size	4

表 3 数据集

Table 3 Datasets

	数据源	清洗前	清洗后	去重后	用途
1	国家岩矿化石标本资源库/条	69 657	29 065	27 402	构建矿物知识库
2	维基百科/条	21 986	18 503	2 980	构建矿物知识库
3	百度百科/条	13 491	9 301	3 102	构建矿物知识库
4	百度问答/条	38 834	16 049	6 408	精排矿物召回文档;微调大模型
5	自建问答对/对	200	200	200	测试
6	矿物术语/条	21 490	10 130	6 529	数据清洗;问题分析

表 5　各模型准确率

Table 5　Accuracies of the models

模型	准确度（no RAG）	准确度（RAG）
chatMineral	82%	93%
T5	41%	48%
GPT2	40%	49%

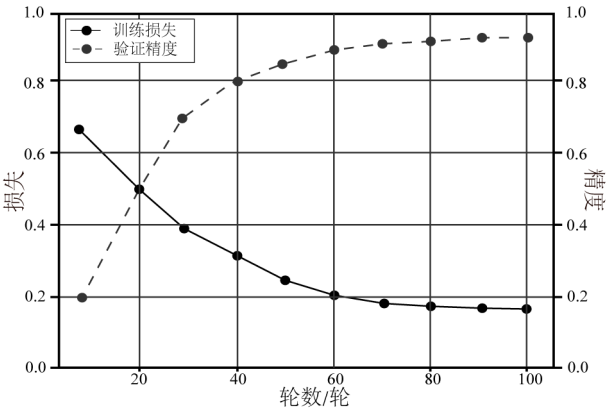

图3　精排模型XGBoost的训练曲线

Fig.3　Training curves of the XGBoost fine-ranking model

表 6　top k 为3时的精确度对比

Table 6　A comparison of precisions for models when top *k* is 3

模型	精度
XGBoost	95%
Word2Vec	76%
BERT	82%
DeBERTa	86%

95%的精确度，超过了其他模型如Word2Vec、BERT和DeBERTa。使用XGBoost对所构建的大语言模型检索增强后对3.1中所列200个问题的准确率为93%，远高于未增强前的82%。

3.3　基于大语言模型的矿物问答系统

采用开源Python框架Streamlit(2024)实现矿物问答系统交互，如图4和图5所示，分别对应PC端和移动端与用户的交互界面。用户输入问题后，答案以流式形式输出，即以打字机的方式逐字显示答案，以增强答案生成时的用户体验。为了帮助用户验证和参考答案，答案还包含了答案源链接，如图4和5中所示的问题是“硅钨锰矿属于哪个晶系？”。除答案外，还包含与答案相关的链接。图6为使用未经矿物知识增强的ChatGLM回答示例问题生成的答案，是错误的，可见本文基于矿物知识增强大语言模型的有效性。如表5所示，系统在回答200个随机选择的矿物问题时的准确

图4　PC端界面

Fig.4　Interface on PCs

图5　移动端界面

Fig.5　Interface on mobile devices

率为93%，未来将进一步获取更多矿物数据以扩展知识库，并进一步优化基础大语言模型的微调，以进一步提高系统准确率。

4　矿物大语言模型的未来工作

在当前已有矿物问答系统的基础上，可使用知识图谱增强大语言模型及融合多个模态，进一步提升矿物知识问答能力。

图6 未经矿物知识增强的ChatGLM对示例问题的问答

Fig.6 The questioning-answering example of the ChatGLM which is not retrieval-augmented by the mineral knowlodge base

4.1 矿物大模型的知识图谱增强

当前的矿物大模型仅使用了相关矿物文档对大语言模型进行领域增强,而知识图谱通过显式和结构化方式存储和表示领域特定知识,具有良好的准确性、解释性和知识扩展性,可进一步增强大语言模型的知识感知。

知识图谱可以在预训练阶段整合到大语言模型中,以帮助大语言模型从知识图谱中学习知识(Zeng et al.,2022;Singhal et al.,2023);也可以在推理阶段将知识图谱整合到大语言模型中,通过从知识图谱中检索知识,提高大语言模型在获取领域特定知识方面的表现;还可以利用知识图谱来解释事实及大语言模型的推理过程,以提高大语言模型的可解释性。

4.2 多模态矿物大模型

当前的矿物问答系统只能处理文本数据,无法处理图像等其他模态数据,需要将不同模态数据(如文本、图像、音频和视频等)进行融合,通过学习不同模态之间的关联,实现更智能化的问答。

多模态大模型在不同模态的数据经过预处理后输入到深度神经网络中,经过多层的特征提取和融合,最终输出相应结果。因此多模态大模型的结构主要包括模态编码器、输入模态投影组件、大语言模型主干网络、输出模态投影组件和模态生成器5个部分,训练时模态编码器、大语言模型主干网络和模态生成器3个部分通常保持冻结,主要训练输入和输出投影组件来进行模态之间的对齐与融合,且模态投影组件的可训练参数小于大语言模型和模态编码器,使得训练多模态大模型经济且高效。

5 结论

本文介绍了大语言模型,并重点介绍了大语言模型的参数高效微调、检索增强生成及提示工程。基于大语言模型构建了矿物问答系统,该系统从在线资源中收集与矿物相关的数据,清洗后形成矿物知识库及矿物问答对用于增强和微调大语言模型。提出了包括召回和精细排序的两阶段检索知识库方法以增强大语言模型生成性能。该问答系统能对任意自然语言形式的矿物问题直接给出答案,避免了矿物数据库或搜索引擎以及基于预定义模板知识图谱的矿物问答系统只能回答特定问题的限制,同时也避免了使用搜索引擎搜索矿物答案时对给出的各参考网页进行答案筛选的过程,能更高效且全面地回答矿物相关问题,在测试数据集上的准确率为93%。未来将继续获取更多矿物数据并定期更新来扩展现有知识库,并将采用知识图谱以增强大语言模型的生成,及加入视觉等更多模态构建多模态矿物大模型。

作者贡献声明：季晓慧，论文构思和撰写、图件绘制、分析测试；刘成健，数据采集、模型训练；杨眉、何明跃、张招崇、曾姗、王玉柱，论文修改、数据分析。

利益冲突声明：作者保证本文无利益冲突。

致谢：感谢审稿人和编辑提出的宝贵意见和建议。

参考文献 (References):

Arora S, Narayan A, Chen M F, Orr L J, Guha N, Bhatia K S, Chami I, Sala F, R′e C. 2022. Ask Me Anything: A simple strategy for prompting language models. ArXiv Preprint ArXiv:2210.02441

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D, Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. 2020. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems,: 1877–1901

Chen T Q, Guestrin C, Chen T Q, Guestrin C. 2016. XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,: 785–794

Dai D M, Sun Y T, Dong L, Hao Y R, Ma S M, Sui Z F, Wei F R. 2023. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In: Anna Rogers, Jordan Boyd-Graber, Naoaki Okazaki (eds.). Findings of the Association for Computational Linguistics. Toronto, Canada: Association for Computational Linguistics, 4005–4019

Du Z X, Qian Y J, Liu X, Ding M, Qiu J Z, Yang Z L, Tang J. 2022. GLM: General language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 320–335.

ES. 2024. https://www.elastic.co/elasticsearch/

Fan W, Ding Y, Ning L, Wang S, Li H, Yin D, Chua T, Li Q. 2024. A survey on rag meets llms: Towards retrieval-augmented large language models. arXiv preprint arXiv:2405.06211

Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Guo Q, Wang M, Wang H. 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997

He Y, Fang P, Shan Y, Pan Y, Wei Y, Chen Y, Chen Y, Liu Y, Zeng Z, Zhou Z, Zhu F, Holmes E C, Ye J, Li J, Shu Y, Shi M, Li Z. 2024. LucaOne: Generalized biological foundation model with unified nucleic acid and protein language. bioRxiv

Hu J E, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Chen W. 2021. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685

季晓慧, 董雨航, 杨中基, 杨眉, 何明跃, 王玉柱. 2024. 基于知识图谱多跳推理的中文矿物知识问答方法与系统. 地学前缘, 31(4): 37–46［Ji X H, Dong Y H, Yang Z J, Yang M, He M Y, Wang Y Z. 2024. Mineral question-answering system in Chinese based on multi-hop reasoning in knowledge graphs. Earth Science Frontiers, 31(4): 37–46 (in Chinese with English abstract)］

Jieba. 2024. https://github.com/fxsjy/jieba

Liu C J, Ji X H, Dong Y H, He M Y, Yang M, Wang Y Z. 2023. Chinese mineral question and answering system based on knowledge graph. Expert Systems with Applications, 231: 120841

Ouyang L, Wu J, Xu J, Almeida D, Wainwright C L, Mishkin P, Chong Z, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R, Ouyang L, Wu J, Xu J, Almeida D, Wainwright C L, Mishkin P, Chong Z, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R. 2022. Training language models to follow instructions with human feedback. In: Koyejo S, Mohamed S，Agarwal, A (eds.). Proceedings of the 36th International Conference on Neural Information Processing Systems. NYUnited States: Red Hook, 27730–27744

OpenAI. 2022. Introducing ChatGPT. OpenAI Blog, November 2022

OpenAI. 2023. GPT-4 Technical Report. ArXiv Preprint ArXiv:2303.08774

Radford A, Narasimhan K, Salimans T, Sutskever I. 2018. Improving language understanding by generative pre-training. OpenAI

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. 2019. Language Models are Unsupervised Multitask Learners. Language models are unsupervised multitask learners. OpenAI Blog, 1(8): 9

Rasley J, Rajbhandari S, Ruwase O, He Y X. 2020. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 3505–3506

Robertson S, Zaragoza H. 2009 The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4): 333–389

Shanahan M. 2024 Talking about large language models. Communications of the ACM, 67(2): 68–79

Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, Clark K, Pfohl S, Cole-Lewis H, Neal D, Schaekermann M, Wang A, Amin M, Lachgar S, Mansfield P, Prakash S, Green B, Dominowska E, Arcas B A Y, Tomasev N, Liu Y, Wong R, Semturs C, Mahdavi S S, Barral J, Webster D, Corrado G S, Matias Y, Azizi S, Karthikesalingam A, Natarajan V. 2023. Towards expert-level medical question answering with large language models: 2305.09617. https://arxiv.org/abs/2305.09617v1

Streamlit. 2024. https://streamlit.io/TF-IDF. 2011. In: Sammut C, Webb G I, eds. Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978–0–387–30164–8_832

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. 2023a. LLaMA: Open and efficient foundation language models. 2302.13971. https://arxiv.org/abs/2302.13971v1

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D M, Blecher L, Ferrer C C, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A S, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I M, Korenev A V, Koura P S,

Lachaux M, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith E M, Subramanian R, Tan X, Tang B, Taylor R, Williams A, Kuan J X, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T. 2023b. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288

Tworkowski S, Staniszewski K, Pacek M. Wu Y, Michalewski H, Milo′s P. 2023. Focused Transformer: Contrastive Training for Context Scaling. In: Proceeding of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), 37: 1–28

Zeng A, Liu X, Du Z, Wang Z, Lai H, Ding M, Yang Z, Xu Y, Zheng W, Xia X, Tam W L, Ma Z, Xue Y, Zhai J, Chen W, Zhang P, Dong Y, Tang J. 2022. GLM-130B: An Open Bilingual Pre-trained Model. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)

Zhang Y F, Wei C, Wu S Y, He Z T, Yu W H. 2023. GeoGPT: Understanding and processing geospatial tasks through an autonomous GPT.: 2307.07930. https://arxiv.org/abs/2307.07930v1

Zhao P, Zhang H, Yu Q, Wang Z, Geng Y, Fu F, Yang L, Zhang W, Cui B. 2024. Retrieval-augmented generation for ai-generated content: A survey. arXiv preprint arXiv:2402.19473

Zhao W X, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, Du Y, Yang C, Chen Y, Chen Z, Jiang J, Ren R, Li Y, Tang X, Liu Z, Liu P, Nie J, Wen J. 2023. A Survey of Large Language Models. ArXiv Preprint ArXiv:2303.18223

Zhao Z, Ma D, Chen L, Sun L, Li Z, Xu H, Zhu Z, Zhu S, Fan S, Shen G, Chen X, Yu K. 2024. ChemDFM: Dialogue foundation model for chemistry. arXiv preprint arXiv:2401.14818. doi.org/10.48550/arXiv.2401.14818

周永章, 张前龙, 黄永健, 杨威, 肖凡, 吉俊杰, 韩枫, 唐磊, 欧阳冲, 沈文杰. 2021a. 钦杭成矿带斑岩铜矿知识图谱构建及应用展望. 地学前缘, 28(3): 67–75［Zhou Y Z, Zhang Q L, Huang Y J, Yang W, Xiao F, Ji J J, Han F, Tang L, Ouyang C, Shen W J. 2021a. Constructing knowledge graph for the porphyry copper deposit in the Qingzhou-Hangzhou Bay area: Insight into knowledge graph based mineral resource prediction and evaluation. Earth Science Frontiers, 28(3): 67–75 (in Chinese with English abstract)］

周永章, 左仁广, 刘刚, 袁峰, 毛先成, 郭艳军, 肖凡, 廖杰, 刘艳鹏. 2021b. 数学地球科学跨越发展的十年: 大数据、人工智能算法正在改变地质学. 矿物岩石地球化学通报, 40: 556–573［Zhou Y Z, Zuo R G, Liu G, Yang F, Mao X C, Guo Y J, Xiao F, Liao J, Liu Y P. 2021b. Overview: A glimpse of the latest advances in artificial intelligence and big data geoscience research. Earth Science Frontiers, 31: 1–6 (in Chinese with English abstract)］

周永章, 肖凡. 2024. 管窥人工智能与大数据地球科学研究新进展. 地学前缘, 31(4): 1–6［Zhou Y Z, Xiao F. 2024. Overview: A glimpse of the latest advances in artificial intelligence and big data geoscience research. Earth Science Frontiers, 31(4): 1–6 (in Chinese with English abstract)］

Zhou Z, Shi J, Song P, Yang X, Jin Y, Guo L, Li Y. 2024. LawGPT: A Chinese Legal Knowledge-Enhanced Large Language Model. arXiv preprint arXiv:2406.04614

（本文责任编辑：李溪遥；英文审校：张兴春）

·亮点速读·

内太阳系第一代星子的中等挥发性元素富集机制

目前,人们对类地行星中等挥发性元素亏损问题的认识仍然存在不足。已有的解释包括中等挥发性元素亏损源于星云凝结不完全或行星分异过程中的损失。针对这一问题,美国加州理工大学的研究团队通过分析岩浆铁陨石(内太阳系的非碳质球粒陨石、外太阳系的碳质球粒陨石)的化学成分,首次系统性重建了太阳系早期星子的中等挥发性元素储存,并揭示了其与地球、火星的同源关联。这些岩浆铁陨石的金属核在星子分异过程中经历快速冷却(<1百万年),有效地封存了原始中等挥发性元素丰度(如Ga、Cu、Cr),避免了后期碰撞或热扰动的改造。

研究者通过Fe/Ni约束了核-慢分异的氧化还原条件(f_{O_2}=IW-3至IW-1,Fe-FeO以下3至1个对数单位)。然后,通过金属-硅酸盐分配系数(亲铁元素Fe、Ni、Co、Mo与亲铜元素S、Sb)与质量平衡模型的联合约束,反推了岩浆铁陨石母体的全岩化学组成。结果表明,非碳质和碳质铁陨石母体表现出与球粒陨石相似的中等挥发性元素丰度,这表明第一代内太阳系星子具有显著丰富的中等挥发性元素特征。同时,铁陨石母体的中等挥发性元素丰度的变化,特别是亏损中等挥发性元素的非碳质和碳质岩浆铁陨石母体,反映了其母体碰撞后的次生挥发性元素流失。除了元素特征,同位素的证据(如$\varepsilon^{54}Cr$、$\Delta^{17}O$)进一步表明,地球与火星的物质组成以贫挥发分非碳质球粒陨石为主,其中地球约有4%~5%来自外太阳系的碳质球粒陨石的贡献,火星的碳质球粒陨石的贡献小于1%。

结合元素和同位素证据,研究者们认为类地行星中等挥发性元素的亏损,更多地与星子碰撞过程中的中等挥发性元素的丢失相关,而非星子分分异过程中不充分的凝结或中等挥发性元素流失。前人的星云凝结不完全观点存在不足:“第一代”内太阳系星子具有显著丰富的中等挥发性元素特征。前人的行星分异观点也存在不足:核-慢分异并未显著损耗中等挥发性元素。

［以上成果发表在国际著名科学期刊***Science Advances***上。Damanveer S. Grewal, Surjyendu Bhattacharjee, Bidong Zhang, Nicole X. Nie, Yoshinori Miyazaki. 2025. Enrichment of moderately volatile elements in first-generation planetesimals of the inner Solar System. Science Advances, 11, eadq7848］

(夏群科 编译)