

基于大语言模型的材料科学信息抽取

时宗彬^{1,2}, 乐小虬^{1,2}

(1. 中国科学院文献情报中心; 2. 中国科学院大学经济与管理学院, 北京 100190)

摘要: 科学文献因具备特有的专业术语和复杂的语义关系, 使得直接利用已有的大语言模型(LLM)从中提取专业领域知识充满挑战。针对LLM在材料科学领域知识抽取性能不足的问题, 提出一种通过领域对齐微调提升抽取效果的方法, 并展示了如何利用微调后的LLM提取材料文献中的关键信息。首先, 通过具备较强语言能力的LLM结合提示工程, 辅助人工标注材料领域的数据集。其次, 在具有7B参数的LLM上采用QLoRA技术进行指令微调, 使模型在微调后能够根据指令从材料科学文献中准确提取信息。微调后的LLM在多个任务上表现优异: 材料实体识别的F1值达到0.94, 材料类型识别的准确率达到0.91, 材料属性数值抽取的准确率达到0.89。实验结果表明, 该方法在处理材料科学文献中的复杂术语和语义关系方面表现出色, 并展现出较强的泛化能力, 为材料科学领域的知识抽取提供了一种新的解决思路。

关键词: 大语言模型; 提示工程; QLoRA; 指令微调; 信息抽取

DOI: 10.11907/rjtk.241885

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2026)001-0001-09



扫描二维码阅读全文:

Information Extraction in Materials Science Based on Large Language Model

SHI Zongbin^{1,2}, LE Xiaoqiu^{1,2}

(1. National Science Library, Chinese Academy of Sciences;

2. School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The distinctive terminology and intricate semantic relationships inherent to scientific literature continue to present a significant obstacle to directly utilizing existing LLM for the extraction of specialized domain knowledge. This study aims to address the limitations in the knowledge extraction capabilities of LLM within the field of materials science. We put forth a methodology to augment their efficacy through domain-specific fine-tuning and illustrate the deployment of the fine-tuned LLM for the extraction of pivotal information from materials science literature. The preliminary stage of the methodology utilises LLM with robust linguistic capabilities and prompt engineering to facilitate the manual annotation of the materials domain dataset. Subsequently, fine-tuning of QLoRA is performed on LLM with parameter 7B, thereby enabling the fine-tuned LLM to accurately extract information from materials science literature in accordance with the instructions. The fine-tuned LLM demonstrated exceptional performance in material entity recognition, achieving an F1 score of 0.94. It also showed high accuracy in material type recognition and material property value extraction, with scores of 0.91 and 0.89, respectively. The experimental results demonstrate that the method is effective in addressing complex terminology and semantic relationships in materials science literature, exhibiting robust generalization capabilities and offers a novel approach to knowledge extraction in the field of materials science.

Key Words: large language models; prompt engineering; QLoRA; instruction fine-tuning; information extraction

0 引言

科学文献通常由专业的学者、研究人员撰写, 针对特定的科学领域进行深入研究。文章内容通常包含复杂的

理论、方法和技术细节, 即使是具备相应专业知识的科研人员也难以快速理解所有内容, 并在设计实验时充分、有效地利用文献中已有的知识。材料知识信息抽取是指利用自然语言处理技术(Natural Language Processing, NLP), 基于BERT等模型从科学文献中抽取结构化的知识, 如材

收稿日期: 2024-12-26

基金项目: 国家社会科学基金项目(23BTQ102)

作者简介: 时宗彬(1993-), 男, 中国科学院文献情报中心博士研究生, 研究方向为语义理解与知识挖掘; 乐小虬(1967-), 男, 博士, 中国科学院文献情报中心研究员、博士生导师, 研究方向为语义理解与知识挖掘。本文通讯作者: 乐小虬。

料类型、属性和数值等,从而帮助研究人员从海量学术文献中提取有用的信息,为材料研发和技术创新提供有力的数据支持,对推动材料科学进步、加速研发进程及其领域的技术创新具有重要意义^[1]。

在科学文献中,材料通常以特殊符号表示,例如 SMILES、IUPAC 等,这与自然语言有很大的不同。另外,材料科学文本中材料之间的关系复杂,涉及大量的专业术语以及复杂的关系表示。例如,在句子“*Our ST-OPVs achieved champion PCEs as high as 13.9% at an AVT of 22.8%, while the PCE of the device with a bulk heterojunction (BHJ) D18/Y6 at 100 nm is 12.3% at an AVT of 17.0%*”中准确找到各个材料设备和对应的光电转换效率(Power Conversion Efficiency, PCE)值,需要 NLP 模型能够准确推理出材料设备与 PCE 之间的关系。因此,从材料科学文献中抽取信息的一个关键是使 NLP 模型能够理解材料领域的专业知识。

在早期的研究中,对于文献中材料信息抽取的研究主要关注于机器学习结合规则的方法。具有代表性的研究是 Mavracic 等^[2]利用 Snowball 算法开发出半自动抽取工具 ChemDataExtractor2.0。随着深度学习技术的崛起,特别是 BERT 预训练模型的出现^[3],Zhao 等^[4]采用 BERT 预训练模型加微调的方法,成功实现了对材料实体的抽取。为进一步提高抽取的准确性,Yoshitake 等^[5]专门针对材料领域训练了 MaterialBERT 模型。同样的,Huang 等^[6]针对金属氧化物电池领域的实体识别任务提出 BatteryBERT 模型,用于自动抽取金属氧化物相关实体,其在该任务上的 F1 值达到了 0.85,显示出优异的性能。然而,这些方法主要聚焦于材料实体信息抽取任务。由于不同研究领域或应用场景可能涉及不同类型的材料实体,基于预训练微调的方法在实际应用中需要花费大量精力设计规则或构建专用数据集^[7]。因此,这些技术虽然取得了一定成效,但仍存在局限性,难以广泛适用于各个材料领域。

随着以 ChatGPT 为代表的大语言模型(Large Language Model, LLM)的崛起,其在理解和利用自然语言序列中的深层语义信息方面展现出了卓越性能^[8]。特别值得一提的是,这些 LLM 所具备的上下文学习能力(In-Context Learning, ICL)使其能够通过输入中巧妙地引入提示和示例,从而出色地完成语言翻译、摘要生成、一般知识问答等多样化的任务。如今,这种能力也成功应用于材料信息抽取任务中。例如,Polak 等^[9]通过与 LLM 对话的方式实现了材料信息抽取。Bran 等^[10]则进一步结合链式思维(Chain-of-Thought, CoT)^[11]和工具调用代理策略^[12],利用 LangChain^[13]框架增强 LLM 在材料领域的信息抽取能力。然而,这些基于 LLM 的 ICL 能力、通过对话方式提取材料信息的方法,通常需要复杂的提示工程设计和具备较大参数数量的 LLM 模型才能满足需求,并且在面对材料领域特有的专业术语时,一般的 LLM 可能会出现理解上的困难,导

致在更复杂、专业的细节性问题上表现不佳^[14]。为了解决该问题,Zhao 等^[15]尝试从头开始训练专注于化学材料领域的 LLM,并在化学合成、性质预测等任务中取得了显著成果。然而,该模型目前主要聚焦于材料合成和性质预测,在材料信息抽取方面的表现仍有待提升。

虽然在已有的利用 LLM 从材料文献提取信息的方案中,基于提示工程的方法被证明是一种有效方法,能够帮助用户在海量文献中快速、准确地提取所需信息,但也需要在设计提示和保存提取结果时进行一定的人工干预与后处理,以确保提取的信息满足需求^[16]。另外,基于提示工程的方法更依赖 LLM 本身的语言能力,且由于 LLM 对提示输入较为敏感,回答的内容难以控制,导致模型的输出有时可能缺乏一致性,甚至是事实错误的^[17]。随着任务复杂性的增加,缺乏领域知识的 LLM 往往难以在专业领域知识问答中保持逻辑的连贯性、事实的准确性和稳健的推理能力^[18]。

考虑到 LLM 在预训练和指令微调(Instruct Fine Tuning, IFT)中通过海量数据训练已经具备一定的领域知识,其在材料专业领域表现较差是由于在 IFT 过程中,已有的 LLM 更关注通用性,在指令微调中缺乏相关专业领域的对齐数据^[19-20]。因此,为了解决基于提示工程从材料科学文献中抽取知识时,因 LLM 缺乏专业领域对齐知识,容易产生错误回答的问题,本文使用材料领域的指令集对 LLM 进行领域知识对齐微调。同时,为了降低 LLM 微调时对显卡资源的消耗,本文使用 QLoRA 方式对 LLM 进行微调^[21]。具体工作如下:

(1)利用语言能力较强的 LLM 结合提示工程,辅助人工标注出有机光电材料领域的数据集,并将该数据集基于事先定义好的模板指令转换成适合 LLM 训练的对话指令集形式。

(2)在占用较低本地资源的情况下,通过 QLoRA 方式对 LLM 进行微调,证明了即使只有 7B 参数的 LLM 在微调后也可很好地完成材料科学文献信息抽取工作。

(3)对比了具备更强语言能力的通用 LLM 与专注于材料科学领域的 LLM 在微调后的性能表现,并探讨了不同超参数设置对 QLoRA 微调后 LLM 性能的影响。

1 研究方法

1.1 高效微调 LLM 方法

随着 LLM 的出现,由于模型参数数量的增大,受训练成本和硬件条件限制,如何以较少成本高效微调 LLM 成为一个难点。Hu 等^[22]提出 Lora 微调方法,通过引入低秩矩阵的方法对 LLM 已有权重进行重参数化,大大降低了训练成本,且能取得与全参微调相近的结果。具体原理如下:

对于待微调参数 $W_0 \in R^{d \times d}$,用 ΔW 表示微调后的变化。在进行 LoRA 微调时,可以固定 W_0 ,将 ΔW 表示成如下

形式:

$$\Delta W = BA \quad (1)$$

其中, $B \in R^{d \times r}$, $A \in R^{r \times d}$, B 采用零值初始化, A 则采用高斯函数初始化。因此, 在一开始训练时, $BA = 0$, LoRA 模块对原模型不造成影响。当输入为 x 时, 此时模型在 W_0 层的输出可表示为:

$$h = W_0 x + \frac{\alpha}{r} \Delta W x = h = W_0 x + \frac{\alpha}{r} BA x \quad (2)$$

在计算梯度时, 只需计算 BA 参数的梯度, 即只需计算 $2r \times d$ 个参数的梯度, 而不使用 LoRA 时, 需要计算梯度的参数为 $r \times r$ 个。当 $d \ll r$ 时, 即可减少很多训练参数, 从而大幅节省微调所需的显存。其中, α 表示缩放因子, r 表示 BA 的秩。

由于 W_0 没有直接参与梯度计算, Tim 等^[21] 进一步提出 QLoRA 方法, 通过对 W_0 进行量化, 当需要计算梯度时再进行逆量化, 并结合分页优化策略减少微调 LLM 时所需的显存, 使得在仅具备 20 GB 显存的硬件条件下, 也能够完成对具有约 7 B 参数规模模型的微调。

基于公式(2)选取不同的 W_0 , 意味着对 LLM 中不同类型的参数进行微调, 这会影响模型在微调后的表现。本文通过选择微调 LLM 自注意力层中 Key、Query 和 Value 等不同参数的组合来寻找最佳的 W_0 设置。实验使用经过指令微调的参数为 7 B 的 LLM 开源模型作为基础模型用于微调实验, 包括 Llama2-7b-chat^[23]、Mistral-7b-instruct^[24]、Gemma-7b-it^[25]、Stack-Llama2-7b 以及 ChemLLM-7b^[26] 模型, 以对比不同模型经领域数据集微调后的效果。

1.2 数据集制作

有机光电材料领域研究热点突出, 涉及多学科, 且信息繁杂, 手动分析难度大。同时, 该领域的术语众多且复杂, 常常存在语义模糊和术语不一致的问题。因此, 该领域非常适合进行信息抽取研究。因此, 本文选择有机光电材料领域科学文献作为实验的抽取对象。由于缺乏相关训练数据, 需要通过人工标注来构建相应的数据集。为减轻人工标注的负担, 本文结合 LLM 与提示工程, 利用 LLM 的 ICL 能力辅助人工标注。通过 LLM 自动标注并结合人工审核的方法, 获取高质量的数据集。

本文使用预先设计好的提示模板, 将有机光电材料领域的文本以句子为单位与模板提示一起输入到本地 WizardLM-70B 模型中进行推理^[27]。如图 1 所示, 依次获得相应的材料实体、类型和属性数值信息。为了确保生成的信息质量, 降低人工审核工作量, 本文限制了 LLM 生成范围, 并对结果进行筛选, 排除与原文不匹配或重复的答案。具体标注过程为: 首先, 利用 PubMed 的 OA 接口获取到 1 000 篇有机光电领域的文献。其次, 使用规则匹配技术从这些文献中找出包含材料光电转换效率(PCE)、开路电压(Voc)等属性数值的信息, 以及材料给体(donor)和受体(acceptor)类型信息的文本。再次, 结合提示工程初步得到材料实体、类型和属性的相关信息。在此基础上, 采用 Doccano 标注工具对结果进行人工审核^[28]。最后, 构建了 3 个数据集, 分别用于材料实体识别、材料类型识别和材料属性抽取, 数据条目数量分别为 1 600 条、3 466 条和 1 192 条。

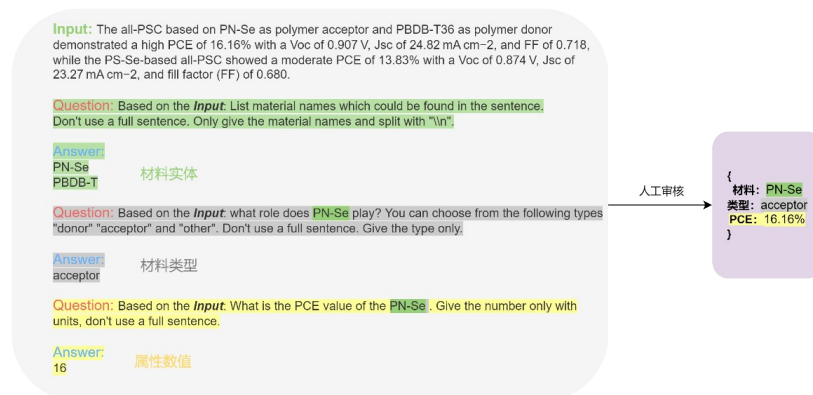


Fig. 1 The example of manual annotation combined with prompts

图 1 结合提示工程的人工标注示例

Wei 等^[29] 的研究表明, 在未经过指令微调的情况下, LLM 可能难以有效响应相关指令。因此, 需要将人工标注的数据转换成一个通过自然语言指令描述的数据集。为了提升微调后 LLM 的语义理解能力, 在指令数据制作过程中对问题进行了详细描述。每条训练数据包含任务说明、问题和答案 3 部分内容, 如图 2 所示。

为了确保指令集的质量, 本文还对提问方式进行了同义替换, 以保证微调后的 LLM 依然有着较强的语义理解能力, 而非仅依赖训练过程中出现过的固定指令进行回答。

本文使用的指令集如表 1 所示, 包含材料实体识别、材料类型判断以及材料属性数值推理 3 个任务。

此外, 考虑到实际输入的文本中可能并不包含问题的答案, 在制作指令数据集时, 给出了允许模型进行否定回答的提示。另外, 在标注的训练集中加入负样本, 将这些负样本对应的答案设置为“None”或“Other”, 从而通过微调训练让 LLM 根据输入的实际学会拒答, 以提升回答事实的准确性。

在构建指令集时, 本文还采用了一些其他材料领域的

```
### Instruction: You are a chemical and material assistant. Your task is to find a specific type of material.

### Question: Based on the following sentence: The absorption spectra for the donor D149 and the acceptor C70 and PC70BM are quite similar in the absorption region, but with differences in the density of each absorption peak. What is the acceptor material that appears in the text. Don't use a full sentence. Only give the material names and split with "\n". If the acceptor material is not present give "None".

### Answer:
C70
PC70BM
```

Fig. 2 The example of instruction dataset

图2 指令数据集示例

Table 1 The questions and instructions included in the instruction dataset

表1 指令集中包含的问题和指令

任务类型	指令	问题
材料名称	You are a chemical and material assistant. Your task is to extract material name.	List material names which could be found in the sentence. Don't use a full sentence. Only give the material names and split with "\n". If the material name is not present type "None". Don't explain why and generate other words which is not in the sentence. Provide a list of materials present in the given sentence, separated by "\n". If no materials are mentioned, simply state "None"
材料类型	You are a chemical and material assistant. Your task is to find a specific type of material.	According to the above sentence, answer what role does {MAT} play? You can choose from the following types "donor" "acceptor" and "other". Don't use a full sentence. Give the type only. the type of {MAT} is What is the {TYPE} material that appears in the text. Don't use a full sentence. Only give the material names and split with "\n". If the {TYPE} material is not present give "None". Don't explain why and generate other words which is not in the sentence.
属性数值	You are a chemical and material assistant. Your task is to find an attribute value of material.	What is the {PROPERTY} value of the {MAT}. Give the number only with units, don't use a full sentence. if the value is not present type "None". Don't explain why.

公开数据集对现有数据集进行扩充,以提升 LLM 的通用性,包括 ChEMU^[30]、SOFC-exp^[31]和 BatteryQA^[6]数据集。其中,ChEMU 是来自材料专利领域的数据集,可用于材料实体识别;SOFC-exp 是来自金属氧化物电池领域的数据集;BatteryQA 是有关电池材料类型问答的数据集。本文将这些公开数据集与标注的有机光电材料数据集一起加入到指令数据集中,最终形成的指令数据集详情如表 2 所示。数据集中训练样本 Token 长度分布如图 3 所示。

Table 2 The source and quantity of instruction dataset

表2 指令集来源和数量

任务类型	有机光电材料	ChEMU	BatteryQA	SOFC-exp
材料实体	1 600	1 200	2 000	418
材料类型	3 466	685	663	-
材料属性数值	1 192	-	-	-

注:“-”表示没有相关数据

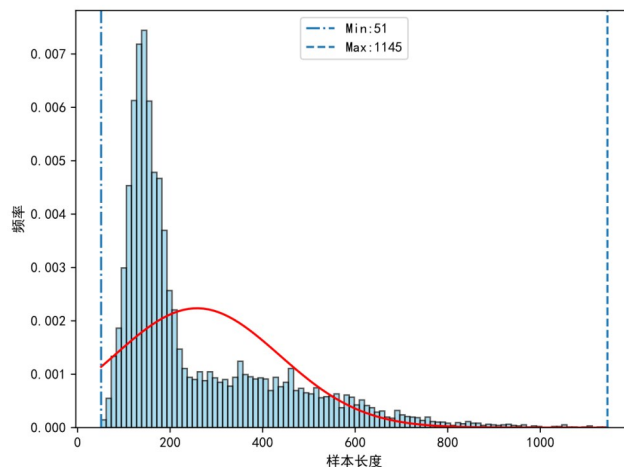


Fig. 3 The distribution of Token lengths in training samples

图3 数据集中训练样本 Token 长度分布

2 实验结果与分析

2.1 实验设置

在进行 QLoRA 微调实验时,统一选择 Adam 作为优化器,学习率设置为 $1e-4$, LLM 模型参数加载选择 INT8 方式,Epoch 设置为 2。每块 GPU 上训练 Batch Size 为 1,权重衰减系数为 0, LoRA 的 Dropout 为 0.05,最大梯度裁剪为 0.3。公式(2)中的超参数 α 固定为 r 的两倍,在有着两块 24 GB 显存的服务器上进行微调实验。在微调过程中,仅计算 LLM 根据指令生成的输出与目标答案部分的损失进

行梯度更新。

2.2 在不同 LLM 上的微调结果

基于不同开源 LLM 所采用的训练数据和模型结构的差异,结合常见的模型结构、预训练数据领域偏好以及模型的语言能力,本文选取 Llama2-7b-chat、Mistral-7b-instruct、Gemma-7b-it、Stack-Llama2-7b 和 ChemLLM-7b 模型进行实验,以确定最佳模型。其中, Mistral-7b-instruct 和 Gemma-7b-it 是最近公布的在 7 B 参数模型中拥有较强语言能力的模型; ChemLLM 是在 InternLM-2 模型^[32]基础上,在化学领域微调后专门用于 AI4SCI 领域的模型;

Stack-Llama2 是基于 LLaMA2 架构, 通过 DPO fine-tune^[33] 使其模型输出更加符合人类偏好。在进行 QLoRA 微调时, 公式(2)中的 r 值默认为 4, 微调对象为模型注意力层中的 Query 和 Value 参数。

图 4 展示了不同 LLM 利用有机光电材料领域的训练数据, 在 1 000 条材料实体和 3 000 条材料类型指令数据集上进行微调时的损失变化图。从图中可以看出, 各个模型的损失值在最初几百次迭代内快速下降, 然后逐渐趋于稳定, 但每个模型的损失波动幅度不同。其中, ChemLLM 的损失曲线波动更为明显, 在稳定阶段时的损失值也比其他模型稍高, 可能是因为其模型使用的 InternLM-2 架构相较 Llama2 架构适应性较差, 使得 ChemLLM 微调的表示不如 Llama2 系列模型稳定, 从而出现更大的损失波动。各个模型最终损失基本在 0.2 上下波动, 但在部分迭代批次中波动较大。可能是因为这些批次的训练数据与模型在预训练阶段接触的数据分布存在较大差异, 或者包含更复杂或不常见的特征, 导致模型在这些批次的学习过程中出现损失较大的波动。

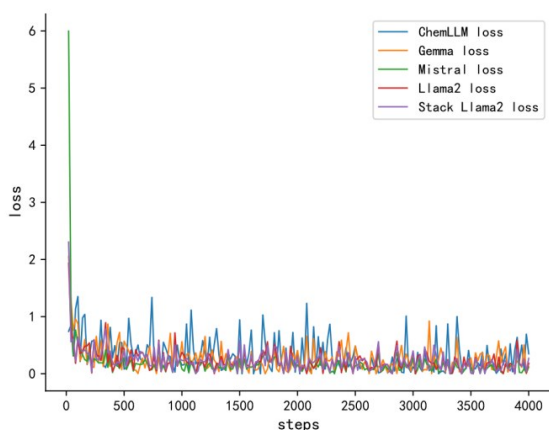


Fig. 4 The loss variation of different LLMs during fine-tuning

图 4 不同 LLM 在微调时的损失变化

在材料实体识别任务中, 将本文方法与传统的预训练加微调方法进行对比实验, 选用了材料领域的预训练模型 MatBERT 加 CRF 的方式进行微调^[34]。微调后各个模型在 600 条材料实体识别和 466 条材料类型判断任务上的实验结果见表 3。由于材料类型判别任务只需验证模型的回答与真实类型是否一致, 因此只给出了准确率。

根据表 3 可以看出, Mistral-7b-instruct 和 Gemma-7b-it 模型在材料实体识别任务上取得了相同的最优值, F1 值达到了 0.85, 而 Llama2-7b-chat 表现最差, 只有 0.82。专门用于化学材料领域的 ChemLLM-7b 虽然只比最优值低了 0.01, 但在测试过程中出现了生成幻觉的问题, 其他模型则没有此现象。而在预训练微调方法中, 尽管基于 MatBERT 与 CRF 的方法在准确率上接近微调后的 LLM 模型, 达到了 0.81, 但其召回率仅为 0.75。这表明当训练数据量不足时, 传统的预训练加微调方法更倾向于依赖模式匹配

Table 3 The results of entity recognition and type classification in the field of organic optoelectronic materials for fine-tuning LLMs and pre-trained modules

表 3 微调后的 LLM 和预训练模型在有机光电材料领域实体识别与类型识别结果

模型	材料实体			材料类型
	准确率	召回率	F1 值	准确率
Mistral-7b-instruct	0.90	0.81	0.85	0.87
Llama2-7b-chat	0.86	0.80	0.82	0.82
Gemma-7b-it	0.90	0.83	0.85	0.86
Stack-Llama2-7b	0.88	0.81	0.85	0.85
ChemLLM-7b	0.86	0.83	0.84	0.88
MatBERT+CRF	0.81	0.75	0.78	-

注: “-”表示没有相关数据

的方式来识别材料实体, 而非基于深层语义信息进行识别, 因而导致召回率偏低。这一现象反映出在数据量较小的情况下, 微调后预训练模型的泛化能力相比 LLM 受到更多限制, 难以有效识别出训练数据中未曾出现过的目标信息。

在材料类型识别上, ChemLLM-7b 获得了最高的准确率 0.88, 但这一结果与其他模型差别不大。该结果表明, 一方面模型微调过程中损失的稳定性在一定程度上与 LLM 的语言能力相关, 具有较强语言能力的模型训练过程中损失比较稳定, 另一方面从模型训练在各任务上表现相差不大以及训练后的生成质量考虑, 相比模型是否具有领域知识, 选择具有更强语言能力的模型才是微调的关键。

在观察这些 LLM 微调后的输出时发现, 影响材料实体识别任务准确率的一个重要因素是实体边界问题。例如在人工标注的材料实体为 “indole dye” 时, Mistral-7b-instruct 只输出了 “indole”, 以及当材料实体以复合结构形式出现时, “C70/MoO₃/Ag” 受训练标注方式的影响, 模型的输出倾向于将其分开为 “C70 MoO₃ Ag”。另外, 在观察这些微调模型在材料类型判断中的错误回答时发现, 相比从隐含的关系中推断出材料类型, 模型的回答更倾向给出不确定的答案, 如图 5 所示。

以上实验结果表明, 相比模型巨大的参数量, 只需采用小量样本进行 QLoRA 指令微调, LLM 即能够有效理解有机电池材料领域的相关知识。相比 LLM 固有的领域知识, 其语言能力更为重要。语言能力越强的模型越容易通过指令微调的形式实现领域知识的对齐。

2.3 域外数据集对结果的影响

由于只使用有机光电材料领域的的数据, 导致训练集中出现的材料实体多为有机材料, 不利于大语言模型对 “材料实体” 这一概念的全面建模与泛化学学习, 导致其在材料实体识别任务上的 F1 值仅为 0.85 左右。而基于 BERT 方法的 F1 值在 0.9 以上, 与其相比还有一定差距^[4]。因此, 本文使用已有的公开数据集对有机光电材料指令集进行了扩充, 详情见表 2。其中, ChEMU 材料类型识别的是溶剂, BatteryQA 材料类型识别的是阴极、阳极和电解质, 而

输入: We observed that the CT emissions gradually decreased and finally disappeared from J71:ZITI-S, J71:ZITI-C to J71:ZITI-N blends, which means that, when the active layers have both optimized morphology and excellent absorption, by narrowing HOMO offset between donor and acceptor materials, the superiority of NFAs can be fully realized.

问题: According to the above sentence, answer what role does ZITI-S play? You can choose from the following types ""donor"" ""acceptor"" and ""other"".

LLM答案: other

人工标注: acceptor

Fig. 5 Examples of incorrect answers in the material type discrimination task

图5 在材料类型判别任务中的回答错误示例

SOFC-exp 侧重于金属氧化物电池领域的材料实体识别,以期通过增加与有机光电材料类型识别相似任务的方式,增强 LLM 的指令跟随能力。对于材料实体识别任务来说,可通过扩充其他领域的材料实体识别任务,以增加模型对材料实体的理解和识别能力。同时,训练过程中增加了 1 000 条材料属性数值识别任务数据,剩余 192 条数据作为测试集。在 Mistral-7b-instruct 模型上进行微调实验的结果如图 6 所示。

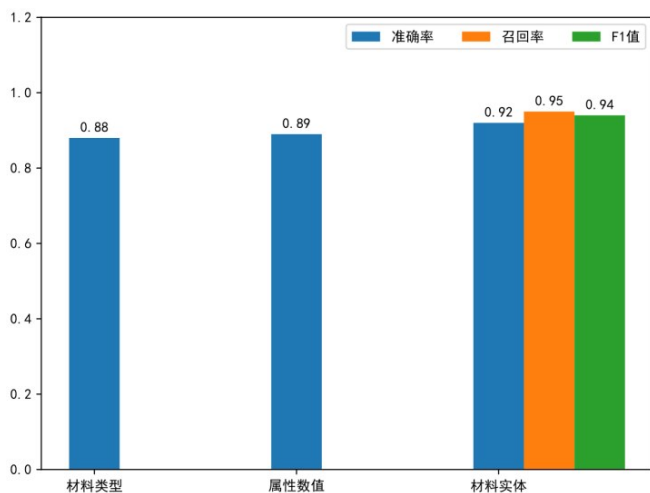


Fig. 6 The results of the Mistral model after fine-tuning with training data from other domains

图6 引入其他领域训练数据进行微调后的 Mistral 模型结果

图6的结果相比表3中的材料实体识别和材料类型判别结果都有所提升。因此,通过增加其他领域的材料实体识别任务,模型可以接触到更多种类的材料实体,从而提高其在材料科学领域的泛化能力和准确性。这种跨领域数据集的扩充还有助于模型学习不同领域材料实体的上下文信息和语境,进而更好地抽取复杂的材料实体。此外,扩充其他领域的抽取任务可以帮助模型识别和理解材料实体之间的关系,从而提升模型在有机光电材料领域的信息抽取能力。因此,在模型微调过程中可以通过适当扩充其他领域任务的方式促进微调后 LLM 性能的提升。

2.4 超参数对比实验

为了比较公式(2)中不同 r 值对实验结果的影响,本文分别选择 $r = 4, 8, 16$,即通过在微调时改变公式(2)中 A, B 矩阵秩的方式,改变微调中可调的参数 ΔW 规模。本文在

Mistral-7b-instruct 上进行了实验,以验证不同 r 值对经过指令数据集微调后的 LLM 在材料实体识别、材料类型判断和材料属性数值三元组抽取 3 个任务中的影响。实验结果见表 4。

Table 4 The impact of different values of r on the fine-tuning results

表4 不同 r 值对微调结果的影响

Mistral-7b-instruct($r=4, 8, 16$)	材料实体			材料类型	材料属性数值
	准确率	召回率	F1 值	准确率	准确率
$r=4$	0.92	0.95	0.94	0.89	0.88
$r=8$	0.94	0.92	0.93	0.91	0.89
$r=16$	0.84	0.94	0.89	0.72	0.81

由表 4 可见,适当增大 r 值会略微提高微调后模型的性能。然而,当 $r = 16$ 时,微调后模型的性能却明显下降。这可能是由于较大的 r 值引入了更多可调参数,但由于训练数据集规模有限,导致模型出现了过拟合的情况,使微调后模型的性能受到了影响。观察微调后模型的输出发现,当 $r = 16$ 时,模型更倾向于拒绝回答,在材料类型判别任务中将材料类型多误归为“other”类型。此外,对于材料属性的数值预测结果也不够完整,常常出现漏掉单位的情况。

在上述微调实验中,仅针对多头自注意力机制中的 Query 和 Value 矩阵进行了微调。为了研究微调不同注意力头对模型的影响,本文还设计了针对不同注意力头进行微调的对比实验,以比较微调不同注意力头的效果。实验结果如表 5 所示,训练过程中的步均损失如图 7 所示。

由表 5 可知,单独对 Key 和 Query 注意力头进行微调时效果最差。相反,同时对 Key-Query 和 Key-Value 注意力头进行微调时,在材料实体识别任务中取得了最佳的 F1 值 0.95,材料属性数值预测的 F1 值达到了 0.89。对于材料类型识别任务,微调 Key 和 Value 注意力头的准确率达到

Table 5 The results of fine-tuning with different combinations of key, query and value

表5 选择不同的 Key、Query 和 Value 组合进行微调的结果

Mistral-7b-instruct($r=4$)	材料实体	材料类型	材料属性数值	Mistral-7b-instruct($r=4$)	材料实体
	准确率	召回率	F1 值	准确率	准确率
KQV	0.96	0.94	0.95	0.85	0.89
KV	0.94	0.94	0.94	0.91	0.85
KQ	0.75	0.88	0.81	0.84	0.87
QV	0.92	0.95	0.94	0.89	0.88

0.91。综合 3 种任务的表现, 基于 Query 和 Value 的微调方式为最优方案, 在各任务中表现更稳定。由图 7 可以看出, 迭代 8 000 步后, 除 Key-Query 微调方式外, 其他 3 种方式的损失趋于一致, 三者在实际任务中的性能也较为接近。

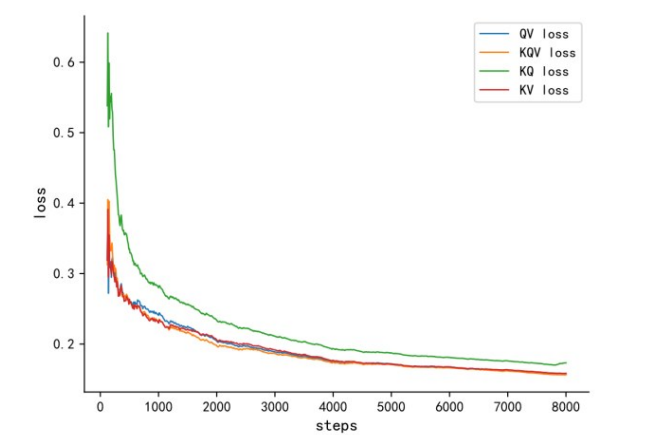


Fig. 7 The training loss of the Mistral model with different combinations of attention heads

图 7 微调 Mistral 模型不同注意力头组合时的训练损失

综上可知, 在进行 QLoRA 微调时选择合适的 r 至关重要。当用于训练的数据集较小时, r 不宜选择太大。而对自注意力块中 Key、Query 和 Value 注意力头进行 QLoRA 微调的组合中, 仅有基于 Key 和 Query 的组合表现较差, 其他组合方式的结果相差并不大, 说明 LLM 在预训练过程中, Key、Query 与 Value 形成了一定的关联性, 需要同时进行微调。因此, 在进行 QLoRA 微调时应避免只对 Key 和 Query 进行微调, 在其他情况下可以根据具体任务选择最优组合。

2.5 泛化能力实验

在有机光电材料指令集中, 材料类型识别任务中已有的训练语料更多集中在给体、受体、溶剂类型上。为了验证微调后的 LLM 确实有跟随相关指令的能力, 本文使用只

在有机光电材料数据集上进行微调的模型 (见表 3), 在 BatteryQA 中类型识别任务数据集上进行了实验。将对给体和受体的识别改为预测电池材料中的 “cathode”、“anode” 和 “electrolyte”, 用于验证模型在相似任务中对其他领域的泛化能力, 并与 GPT4o 上的结果进行了比较, 以评估微调后模型在跨领域任务上的适应性及其与 SOTA 模型的竞争力。为了比较的公平性, 实验采用 Zero-Shot 的形式, 即只使用指令而不提供实例, 实验结果见表 6。

Table 6 The generalization ability of different LLMs in material type recognition task

表 6 不同 LLM 在材料类型识别任务中的泛化能力						
模型	Mistral-7b-instruct	Llama2-7b-chat	Gemma-7b-it	Stack-Llama2-7b	ChemLLM-7b	GPT4o
准确率	0.85	0.71	0.75	0.82	0.78	0.89

可以看出, 不同 LLM 中 Mistral-7b-instruct 和 Stack-Llama2-7b 的泛化能力最强, 准确率分别只下降了 0.02 和 0.03, 说明具有较强语言能力的模型通过微调可以很容易迁移到其他领域的相似任务中。当缺乏相关领域的数据集时, 使用其他领域相似任务的数据集进行等效训练也能取得不错的结果。然而, 其准确率仍比最新的 GPT-4o 模型低 0.04, 表明模型性能除受训练数据影响外, 更大参数量的模型在语言理解和生成能力方面优势显著, 从而具备更强的泛化性能。尽管如此, 通过精细微调, 小参数量模型也能达到与之接近的性能水平。

另外, 本文还对微调后的 Mistral-7b-instruct 模型进行了跨语言能力测试, 将输入内容替换为中文, 观察只在英文数据集上进行微调后模型在中文有机光电材料领域执行相关任务的能力。结果发现, 仅通过英语语料的训练, 模型同样拥有了在中文领域执行相同任务的能力。在中文领域测试的示例如表 7 所示, 表明微调后的模型在中文领域也保持了良好的适应性。

Table 7 Examples extracted by the Mistral model in Chinese domain
表 7 Mistral 模型在中文领域抽取示例

任务类型	输入	问题	Mistral-7b-instruct 输出
材料实体识别	本文以 4,4'-二氯二苯甲酮为起始原料, 经硼氢化钠还原、氯代、哌嗪取代 3 步反应得到 1-(双(4-氯苯基)甲基)哌嗪。合成的 1-(双(4-氯苯基)甲基)哌嗪通过 H NMR 和 LC-MS 进行结构确认, 从谱图中也可看出所得产物的纯度较高, 结构准确。该路线简化了后处理过程, 提高了收率, 有利于工业化生产	List material names which could be found in the sentence. Don't use a full sentence. Only give the material names and split with "\n". If the material name is not present type "None".	4,4'-二氯二苯甲酮 硼氢化钠 氯代 哌嗪 1-(双(4-氯苯基)甲基)哌嗪
		According to the above sentence, answer what role does 二苯胺 play? You can choose from the following types "给体" "受体" "溶剂" and "other".	给体
材料属性数值	通过计算得到 D-PTZ、D-PTZa、D-PTZb 和 D-PTZc 的前线轨道分布如图 3 所示 (等值面数值=0.02), 其中化合物 D-PTZ、D-PTZb 和 D-PTZc 的 HOMO 主要分布在两个吩噻嗪基团或两个苯并吩噻嗪基团上, 其 HOMO 能级分别为 -5.31、-5.39 和 -5.21 e V	What is the HOMO value of the D-PTZb. Give the number only with units, don't use a full sentence. if the value is not present type "None".	-5.39 e V

最后,本文从美国化学会(ACS)数据库中以“organic photoelectric material”为关键词检索到 1 000 篇相关文献,并利用经过微调的 Mistral-7b-instruct 模型从这些文献中提取了关于给体、受体材料及其相应 PCE 指标的数据。基于这些提取的数据,按照文献发布年份对近年来有机光电材料领域 PCE 的研究进展进行了分析,并绘制了相应的趋势图。此外,还给出了该领域常见的给体和受体材料分

布,如图 8 所示。从图中可以清晰地看出,在有机光电材料领域,给体材料 P3HT 和受体材料 PCBM 是研究的焦点。在 1 000 篇文献中,分别有约 337 篇和 216 篇文献提及了这两种材料。此外,该领域的研究在 2017–2020 年期间显著加速,期间 PCE 指标从大约 10% 提高到超过 20%。截至 2023 年,在有机光电材料领域,PB2:FTCC 材料实现了最高的 PCE,达到了 30.2%。

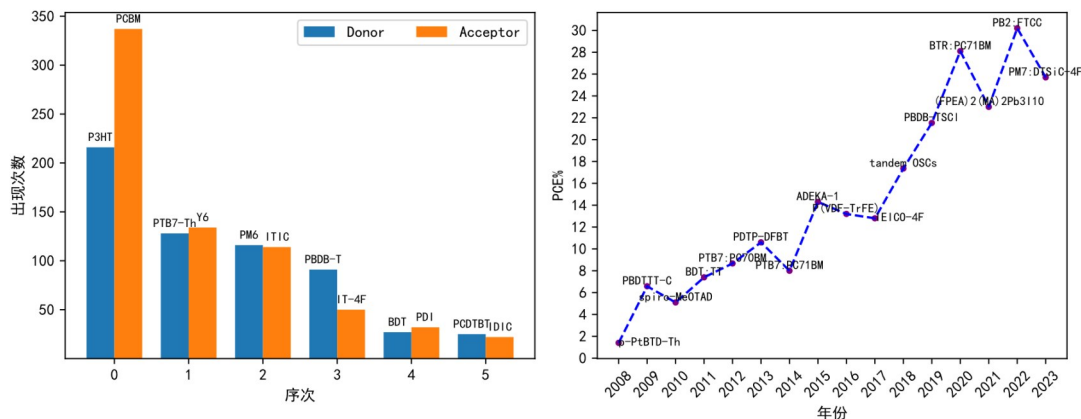


Fig. 8 Common donor and acceptor materials and trend of PCE changes

图 8 常见的给体和受体材料及 PCE 演变趋势

3 结语

本文研究表明,只需少量材料领域的指令数据集微调 LLM 即能够实现从非结构化的材料科学文献中提取信息,从而快速、高效地实现材料领域的信息抽取,为从科技文献中提取专业领域信息提供了一种高效、可靠的方法。本文通过提示工程辅助人工标注的方法,构建了涵盖材料实体识别、材料类型判别和材料属性数值问答 3 个任务的有机光电材料领域数据集。之后,利用 QLoRA 调参技术对拥有 7 B 参数的 LLM 进行微调实验。实验结果显示,经过少量数据微调的 LLM 在相关任务中取得了显著成效。在进一步的分析中,发现 QLoRA 微调方式对于 r 值的选择比对注意力头 Key、Query 和 Value 的选择更为敏感。此外,相较于 LLM 本身固有的领域知识,其语言能力在微调过程中显得更为重要。具有更强语言能力的模型在微调后不仅在相关任务中表现更佳,而且展现出更强的泛化能力。

受限于本地设备,本文只针对具有 7 B 大小参数的模型进行了微调实验,而没有选择具有更大参数的 LLM,且训练样本长度为句子级别,导致其在长文本问答中性能有所下降。未来计划在具有更大参数的模型上进行相关实验,在提升相关任务准确率的同时,通过增加材料制备、实验条件等信息抽取相关任务,以及增加训练样本长度等方式,进一步增强 LLM 在材料科学领域的信息抽取能力。

参考文献:

[1] SMITH A, BHAT V, AI Q X, et al. Challenges in information-mining the

materials literature: a case study and perspective[J]. Chemistry of Materials, 2022, 34(11): 4821–4827.

[2] MAVRACIC J, COURT C J, ISAZAWA T, et al. Chemdataextractor 2.0: autopopulated ontologies for materials science[J]. Journal of Chemical Information and Modeling, 2021, 61(9): 4280–4289.

[3] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171–4186.

[4] ZHAO X T, GREENBERG J, AN Y, et al. Fine-tuning BERT model for materials named entity recognition[C]. 2021 IEEE International Conference on Big Data, 2021: 3717–3720.

[5] YOSHITAKE M, SATO F, KAWANO H, et al. MaterialBERT for natural language processing of materials science texts[J]. Science and Technology of Advanced Materials: Methods, 2022, 2(1): 372–380.

[6] HUANG S, COLE J M. BatteryBERT: a pretrained language model for battery database enhancement[J]. Journal of Chemical Information and Modeling, 2022, 62(24): 6365–6377.

[7] WU T Y, HE S Z, LIU J P, et al. A brief overview of ChatGPT: the history, status quo and potential future development[J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(5): 1122–1136.

[8] DONG Q X, LI L, DAI D M, et al. A survey on in-context learning[C]. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024: 1107–1128.

[9] POLAK M P, MORGAN D. Extracting accurate materials data from research papers with conversational language models and prompt engineering—example of ChatGPT[J]. Nature Communications, 2024, 15(1): 1569.

- [10] BRAN A M, COX S, SCHILTER O, et al. Augmenting large language models with chemistry tools[J]. *Nature Machine Intelligence*, 2024, 6(5): 525–535.
- [11] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 24824–24837.
- [12] SCHICK T, DWIVEDI-YU J, DESSÌ R, et al. Toolformer: language models can teach themselves to use tools[DB/OL]. https://proceedings.neurips.cc/paper_files/paper/2023/file/d842425e4bf79ba039352da0f658a906-Supplemental-Conference.pdf.
- [13] TOPSAKAL O, AKINCI T C. Creating large language model applications utilizing LangChain: a primer on developing LLM APPs fast[C]//*International Conference on Applied Engineering and Natural Sciences*, 2023: 1050–1056.
- [14] PIMENTEL A, WAGENER A, DA S E F, et al. Challenging ChatGPT with chemistry-related subjects [DB/OL]. <https://arxiv.org/pdf/2305.11890>.
- [15] ZHAO Z H, MA D, CHEN L, et al. ChemDFM: dialogue foundation model for chemistry[DB/OL]. <https://arxiv.org/pdf/2401.14818v1>.
- [16] JABLONKA K M, AI Q X, AL-FEGHALI A, et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon[J]. *Digital Discovery*, 2023, 2(5): 1233–1250.
- [17] KAMOI R, DAS S S S, LOU R, et al. Evaluating LLMs at detecting errors in LLM responses[DB/OL]. <https://arxiv.org/pdf/2404.03602>.
- [18] DAGDELEN J, DUNN A, LEE S, et al. Structured information extraction from scientific text with large language models[J]. *Nature Communications*, 2024, 15(1): 1418.
- [19] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 27730–27744.
- [20] ZHOU C T, LIU P F, XU P X, et al. LIMA: less is more for alignment [C]// *Proceedings of the 37th International Conference on Neural Information Processing System*, 2023: 55006–55021.
- [21] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. Qlora: efficient finetuning of quantized LLMs[C]// *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023: 10088–10115.
- [22] HU E J, SHEN Y L, WALLIS P, et al. Lora: low-rank adaptation of large language models[DB/OL]. <https://arxiv.org/pdf/2106.09685>.
- [23] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundation and fine-tuned chat models[DB/OL]. <https://arxiv.org/pdf/2307.09288>.
- [24] JIANG A Q, SABLAYROLLES A, MENSCH A, et al. Mistral 7B[DB/OL]. <https://arxiv.org/pdf/2310.06825>.
- [25] TEAM G, MESNARD T, HARDIN C, et al. Gemma: open models based on gemini research and technology[DB/OL]. <https://arxiv.org/pdf/2403.08295>.
- [26] ZHANG D, LIU W, TAN Q, et al. ChemLLM: a chemical large language model[DB/OL]. <https://arxiv.org/pdf/2402.06852>.
- [27] XU C, SUN Q, ZHENG K, et al. WizardLM: empowering large language models to follow complex instructions[DB/OL]. <https://arxiv.org/pdf/2304.12244v1>.
- [28] NAKAYAMA H, KUBO T, KAMURA J, et al. Doccano: text annotation tool for human[EB/OL]. <https://github.com/doccano/doccano>.
- [29] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot learners[DB/OL]. <https://arxiv.org/pdf/2109.01652>.
- [30] HE J, NGUYEN D Q, AKHONDI S A, et al. An extended overview of the CLEF 2020 ChEMU lab: information extraction of chemical reactions from patents [C]//*Proceedings of the CLEF 2020 Conference*, 2020: 22–25.
- [31] FRIEDRICH A, ADEL H, TOMAZIC F, et al. The SOFC-Exp corpus and neural approaches to information extraction in the materials science domain[C]//*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020: 1255–1268.
- [32] CAI Z, CAO M S, CHEN H J, et al. InternLM2 technical report[DB/OL]. <https://arxiv.org/pdf/2403.17297>.
- [33] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: your language model is secretly a reward model[C]// *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023: 53728–53741.
- [34] WALKER N, TREWARTHA A, HUO H Y, et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science[J]. *Patterns*, 2022, 3(4): 100488.

(责任编辑:黄健)