

大语言模型提取文献中锂离子电池固态电解质合成信息

韦士豪, 李舒远, 王亚鑫, 孙少瑞

(北京工业大学材料科学与工程学院, 北京 100124)

摘 要: 固态电解质是提升锂离子电池性能和安全性关键材料, 获取其结构化的合成信息对于相关研究至关重要。人工从文献中提取费时费力, 基于规则的方法泛化性能差, 机器学习的方法则需大量标记数据, 都存在一定的局限性。本研究旨在利用大语言模型自动提取文献中的固态电解质合成信息。首先, 训练段落分类模型, 包括贝叶斯模型、BERT 模型和 3 个不同的大语言模型, 筛选出描述固态电解质合成过程的段落。然后, 训练合成信息提取模型, 为 3 个不同的大语言模型, 从这些段落中提取合成信息。在段落分类任务中, Gemma 模型的整体表现最好, F1 分数超过 0.9, LLaMA 模型和 BERT 模型的表现次之, F1 分数都超过 0.85, 贝叶斯模型的表现最差, F1 分数低于 0.85。在合成信息提取任务中, 3 个大语言模型的 F1 分数都超过了 0.9。结果表明, 大语言模型可以高效且准确地提取文献中锂离子电池固态电解质的合成信息, 且小参数模型的性能与大参数模型相当。

关键词: 大语言模型; 锂离子电池; 固态电解质; 提取

中图分类号: TM912 文献标志码: A 文章编号: 0454-5648(2025)07-1844-12

网络出版时间: 2025-06-03



锂离子电池因其高能量密度、长寿命和相对较好的环境兼容性, 已被用于电动汽车领域^[1]。电动汽车具有智能、舒适和环保等优点, 因而成为越来越多消费者购车的首选。然而, 电动汽车的安全和续航问题也在引发人们的担忧^[2], 究其根源在于车载电池。传统锂离子电池使用液态电解质, 在过充放电或碰撞等情况下, 可能会导致电池隔膜破损, 引发电池短路, 出现热失控事件^[3], 使得电池温度急剧升高。当温度达到一定程度时, 电解液分解并产生可燃气体^[4], 导致电动汽车起火燃烧。此外, 使用液态电解质的电池难以满足人们对长续航的需求, 特别是在冬季低温环境下。而且, 低温充放电会导致活性锂损失^[5], 进而加剧电池使用寿命的衰减。

固态电解质作为提高锂离子电池安全性和续航能力的关键材料, 受到了人们的广泛关注。与液态电解质相比, 固态电解质具有不易燃、无泄露和化学稳定性好的优点^[6], 可以从根本上解决电动汽车热失控引发的安全问题。固态电解质可以适配更高比容量的电极材料, 降低电解质的用量^[7], 从而提

高电池的能量密度, 为电动汽车提供更持久的续航。

然而, 固态电解质的研究和开发面临着诸多挑战, 包括离子导电性、界面稳定性和机械性能等方面的问题^[8-10]。为了克服这些困难, 研究人员使用了各种策略探索新型固态电解质材料。人工智能技术, 如机器学习, 可以高效分析材料成分和性能之间的关系, 从而加速新型固态电解质材料的发现和优化过程^[11]。

固态电解质的材料成分与合成过程等信息广泛存在于海量文献文本中, 需要从这些非结构化的文本中提取出结构化的材料数据, 才能用于后续的机器学习分析。人工提取材料数据不仅费时费力, 而且难以保证数据的准确性。基于规则的提取方法简单实用, 已经被用于从专利和文献中提取各种信息^[12-16], 但是只能提取规则范围内的特定信息, 难以覆盖复杂多样的合成过程的文本描述。机器学习技术能够高效地从文献中自动提取材料信息, 并且具有灵活性, 可以根据不同的任务和领域进行调整, 但是这通常需要使用大量人工标注的数据训练模型^[17-22]。

收稿日期: 2025-01-02。 修订日期: 2025-01-23。

基金项目: 国家科技重大专项资助项目(2021YFA1201000)。

第一作者: 韦士豪(1999—), 男, 硕士研究生。

通信作者: 孙少瑞(1978—), 男, 博士, 教授。

Received date: 2025-01-02. Revised date: 2025-01-23.

First author: WEI Shihao (1999—), male, Master candidate.

E-mail: wsh228581@163.com

Correspondent author: SUN Shaorui (1978—), male, Ph.D., Professor.

E-mail: sunsr@bjut.edu.cn

大语言模型(LLMs)凭借其卓越的语言理解和文本处理能力,在自然语言处理(NLP)领域取得了突破性进展,不仅变革了材料科学研究的范式^[23-26],而且为文献信息提取工作开辟了新的道路^[27-29]。本工作提出将 LLMs 的理解与处理能力引入锂离子固态电解质领域,用于自动从文献中筛选固态电解质合成段落,提取并输出结构化合成信息。首先,采用 ChatGPT 辅助标注结合人工检查的方式构建段落分类数据集,一部分用于训练贝叶斯模型、微调 BERT 模型和 LLMs,另一部分用于评估这些模型。然后,使用前期人工构建的少量合成信息提取数据微调 LLMs,使其学习从非结构文本提取结构化信息的过程。最后,将分类模型筛选出的段落输入提取模型,并评估模型性能。本工作为固态电解质的结构化合成信息提取提供了参考,有望加速固态电

解质材料的研发过程。

1 方法

1.1 提取框架

固态电解质合成信息提取流程如图 1 所示,包括文献下载、解析与处理、段落分类、信息提取和合成过程可视化 5 部分。首先下载的 XML 文献会被解析,提取出“Experimental”或“Method”标题下的所有段落,然后逐个输入段落分类模型,筛选出的固态电解质合成段落被输入信息提取模型,最后输出结构化的合成信息,包括产物、原料、方法和步骤 4 部分。提取的结构化合成信息既可以保存为 JSON 文件构建数据集,也可以结合 Graphviz 绘制固态电解质合成路线图,方便查看。图 2 是一个提取固态电解质合成信息并绘制合成路线图的例子。

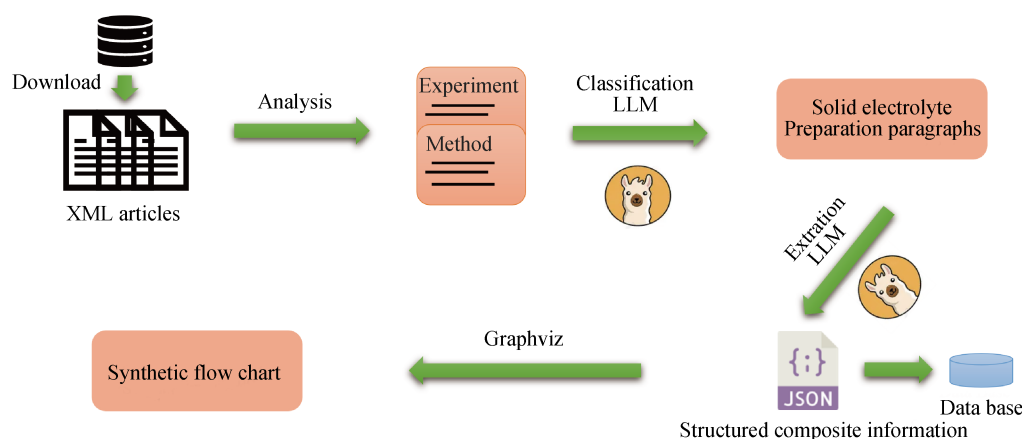
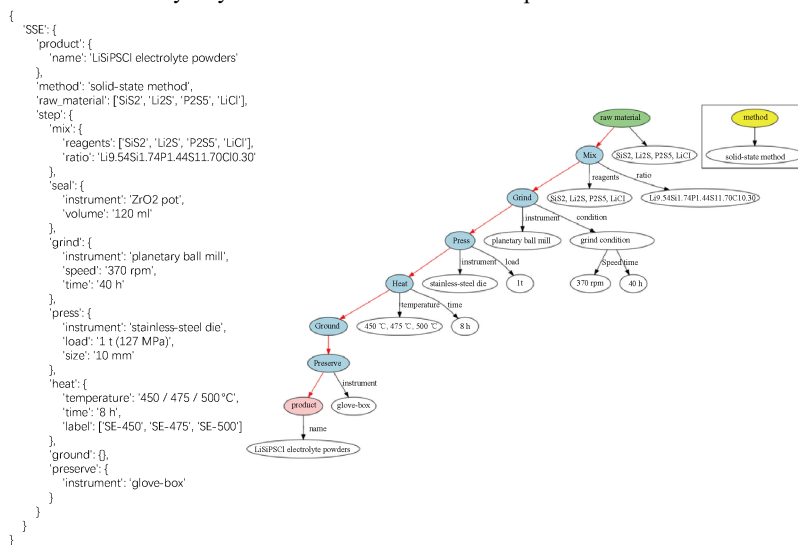


图 1 固态电解质合成信息提取流程

Fig. 1 Solid-state electrolyte synthesis information extraction process

LiSiPSCl electrolyte powders were synthesized by a solid-state method [25,44]. The above-prepared SiS₂ and commercial Li₂S (Aladdin; 99.5 wt%), P₂S₅ (9 Ding Chem. Co., China; 99 wt%) and LiCl (Sigma-Aladdin; 99 wt%) powders were mixed according to the nominal molar ratio of Li_{0.54}Si_{1.74}P_{1.44}S_{11.70}Cl_{0.30}. Then the mixture was sealed in a ZrO₂ pot (120 ml in volume) containing a certain amount of ZrO₂ balls (φ10 mm, balls: powders = 30 : 1, wt%) and ground in a planetary ball mill at 370 rpm for 40 h. The ball-milled powder was pressed into pellets with a stainless-steel die (φ10 mm) under a load of 1 t (127 MPa). Pellets were sealed in tubes under vacuum and then heated at 450 / 475 / 500 °C for 8 h (labeled as SE-450, SE-475, SE-500, respectively). Finally, they were taken out, ground to powders and preserved in the glove-box.



Paragraph

Extraction

Flow chart

图 2 固态电解质合成信息提取示例

Fig. 2 Solid-state electrolyte synthesis information extraction example

1.2 文献获取

为了方便获取论文,使用爬虫的方式从购买的出版社数据库中批量下载。在各出版社中,Elsevier ScienceDirect 数据库文献的获取较为方便,因此,爬取该数据库中的相关文献。为了确保获取的文献相关性更高,检索范围限定在标题,检索关键词为“solid lithium battery OR solid Li battery”,检索年份为 2010~2023。下载的文献格式为 XML,内容组织有序,方便解析和提取。

1.3 模型选择与微调

朴素贝叶斯(NB)模型基于贝叶斯定理,在许多实际应用中表现出色,特别是在文本分类和垃圾邮件过滤等方面。词频-逆文档频率(TF-IDF)是一种统计方法,用于评估一个词语对于一个文档集或一个语料库中的其中一份文档的重要性。词频是指在文档中某个词语出现的次数,用于度量词语的重要性。一个词语在文档中出现的次数越多,其 TF 值就越高。逆文档频率是指在整个文档集合中包含该词语的文档的比例,用于度量一个词语的普遍重要性。一个词语在越多的文档中出现,其 IDF 值就越低,表明该词语在文档集合中不那么具有区分性。将 TF-IDF 与 NB 模型结合,可以提高分类的准确性和可靠性。

BERT 是一种基于 Transformer 架构的预训练语言模型,专为自然语言处理任务设计。它通过遮蔽语言模型(MLM)和下一句预测(NSP)两项预训练任务学习通用的语言知识和语义表示,能够捕捉到丰富的语言模式和上下文信息。BERT 模型凭借其强大的上下文表示能力和广泛的适用性,可以通过微调适应多种下游任务,如文本分类、问答系统和命名实体识别等。

作为对比,还使用了 LLMs 进行段落分类,为确保数据安全,方便自主开发,本工作使用本地部署约开源 LLMs 进行段落分类和信息提取。开源大模型种类众多,参数覆盖 1B 到 72B,代表性的开源大模型有 Meta 的 LLaMA 系列、Google 的 Gemma 系列、智谱的 ChatGLM 系列和通义千问的 Qwen 系列。由于处理的是英文文献,同时考虑计算资源和模型性能之间的平衡,并且为了对比不同 LLMs 的效果,所以选择 LLaMA 系列中的 LLaMA3.2-3B-Instruct 模型和 LLaMA3.1-8B-Instruct 模型, Gemma 系列中的 Gemma-7B-Instruct 模型(下文分别简称为 3B 模型、8B 模型和 7B 模型),“Instruct”表示该模型在预训练的基础上进行了指令调优,能够

更好地遵循任务指令。参数高效微调(PEFT)是一种 LLMs 微调技术,旨在减少模型训练参数数量的同时保持其性能。本工作使用 PEFT 中的 LoRA 方法微调模型,该方法具有训练速度快、计算资源少和多任务学习等优点。

LLaMA-Factory 是一个简单高效、开箱即用的开源框架,集成大模型的微调、推理和融合于一体,适配各种主流开源模型,支持多种微调方法。本文借助该框架和自有数据集微调模型,以完成段落分类和合成信息提取任务。

1.4 段落分类

“Experimental”和“Materials”标题下的段落内容主要有固态电解质制备、电极制备和电池组装 3 部分。为了准确地识别描述固态电解质制备过程的段落,需要训练段落分类模型。首先,对于有小标题的段落,可以直接划分为固态电解质合成段落和非固态电解质合成段落。对于没有小标题的段落或者同一小标题下的多个段落,使用 ChatGPT 辅助标注,将这些段落作为输入,请求 ChatGPT 判断每个段落描述的是否为固态电解质的制备过程。如果是,则标记为 1;否则,标记为 0。所有标注的段落都经过了人工检查,最终得到 300 个带标签的数据,其中固态电解质制备段落和其他段落各占一半,数据示例如图 3 所示。然后,选取 200 个数据进行训练和微调,得到 NB、3B、7B、8B 和 BERT 5 个不同的分类模型。最后,利用剩余的 100 个数据进行测试和评估。

1.5 信息提取

由于此前在无机催化剂合成信息提取工作中已经构建了相关数据集,因此,在合成信息提取任务中直接使用该数据集。其格式为 OpenAI 格式,即每条数据由 system、user、assistant 构成,分别对应要执行的任务、输入的合成文本、输出的结构化数据,共计 150 个数据,数据示例如图 3 所示。输出的结构化信息包括产物、原料、方法和步骤 4 部分,是人为预定义的模板。首先,使用这 150 个数据进行微调,得到 3B、7B 和 8B 3 个合成信息提取模型。然后,使用段落分类模型筛选出的 50 个固态电解质合成段落进行测试和评估。

1.6 任务评估

对于段落分类任务,从精确率、召回率、F1 分数和混淆矩阵 4 个方面评估 NB、3B、7B、8B 和 BERT 5 个模型的分类性能。其中精确率表示模型预测为正类的样本中实际为正类的比例;召回率表示所有实际为正类的样本中,模型正确预测为正类的

比例; F1 分数表示精确率和召回率的调和平均数, 用于综合考虑精确率和召回率; 混淆矩阵展示每个

类别的预测值和实际值的对比, 用于描述分类模型的性能。

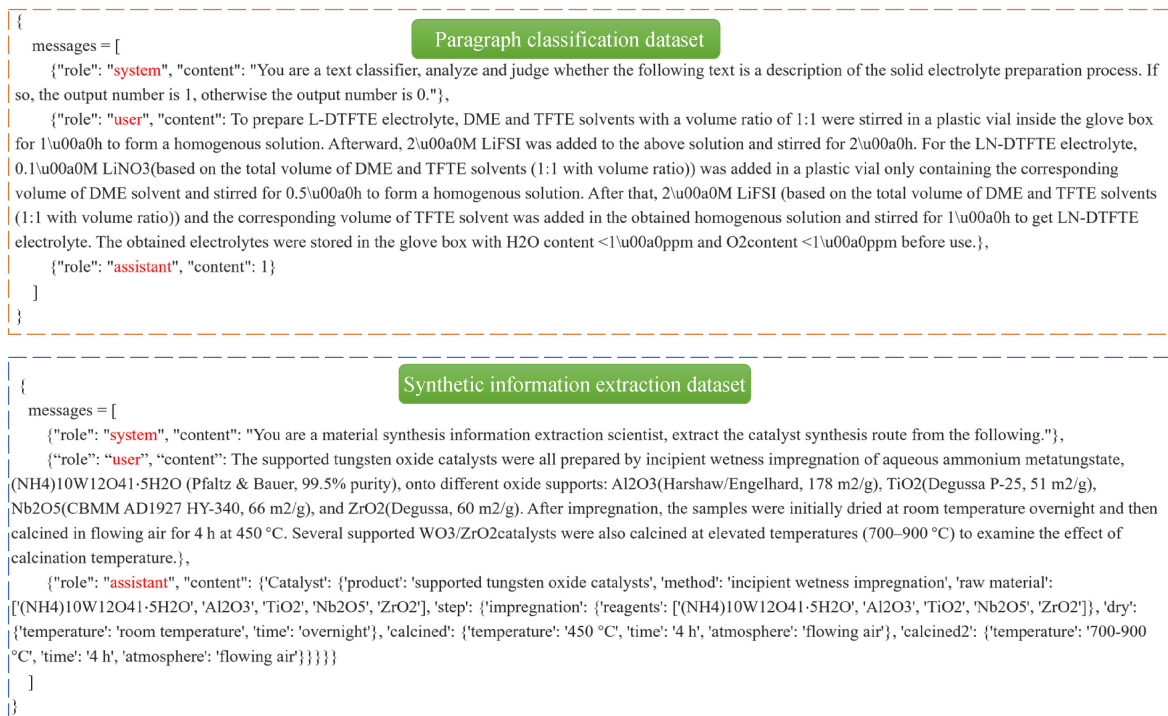


图 3 微调数据集示例

Fig. 3 Examples of fine-tuning datasets

对于信息提取任务, 使用精确率、召回率和 F1 分数指标评估 3B、7B 和 8B 3 个模型的提取性能^[30]。3 个指标的计算方式分别为: 精确率=TP/(TP+FP), 召回率=TP/(TP+FN), F1 分数=(2×精确率×召回率)/(精确率+召回率)。其中, TP 表示模型提到实际也有的实体数量, FP 表示实际没有而模型提到的实体数量, FN 表示模型没有提到而实际有的实体数量。

2 结果与讨论

2.1 段落分类

2.1.1 NB 模型训练 数据集被划分为训练集和验证集, 验证集比例为 0.2, 使用 TF-IDF 提取文本特征并构建特征向量, 利用多项式贝叶斯(MNB)算法进行训练。为了评估模型的性能和泛化能力, 使用 5 折交叉方法测试, 测试结果如图 4 所示。训练得分几乎接近 1, 这表明模型在训练集上表现非常好, 几乎可以完美地拟合训练数据。交叉验证得分随着训练样本数量的增加而提高, 表明模型的泛化能力在提升, 但始终低于训练得分, 这表明模型在未见过的验证数据上的表现不如在训练数据上。交叉验证得分的阴影区域相对较小, 表明模型在不同的数据划分上表现相对稳定。

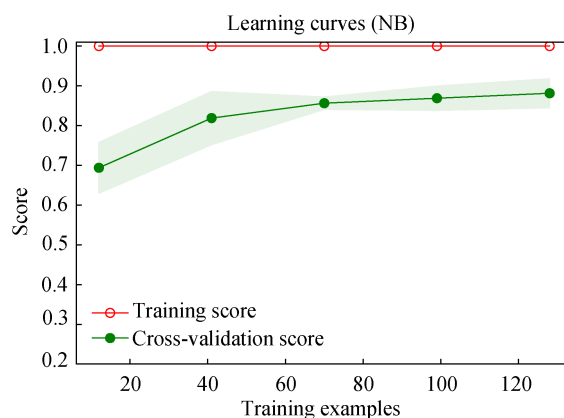


图 4 不同训练样本量对段落分类准确性的影响

Fig. 4 Impact of different training sample sizes on paragraph classification accuracy

2.1.2 3B 模型训练 3B 模型是 Meta 推出的新一代轻量级模型, 结合了监督微调、近端策略优化和直接偏好优化, 具有较强的指令遵循能力。首先, 数据集被划分为训练集和验证集, 验证集比例为 0.1。然后, 设置一系列超参数训练模型, 初始学习率为 1.0×10^{-4} , 并根据 train_loss 和 val_loss 曲线拟合情况调整学习率, 经过多轮调整确定合适的学习率为 5.0×10^{-5} 。如图 5a 所示, 模型在训练集和验证集上的拟合表现一致, 没有过拟合或欠拟合现象。

最后, 固定学习率并在整个数据集上进行训练。如图 5b 所示, train_loss 随着 step 的增加而迅速下降, 表明模型在学习并逐渐拟合训练数据。在 138 个 step 之后, train_loss 几乎不再变化, 表明模型的学习过程趋于稳定。最后将模型微调权重和原始权重合并得到段落分类模型。

2.1.3 7B 模型训练 7B 模型由 Google 推出, 是 Gemma 系列开源模型的开山之作, 基于 Gemini 模型使用的相同技术构建。虽然被称作 7B 模型, 但是该模型的实际参数为 8B 量级。如图 6a 所示, 模型在训练集和验证集上的拟合表现一致, 没有过拟合或欠拟合现象, 该过程的学习率为 5.0×10^{-5} 。最后, 固定学习率并在整个数据集上进行训练。如

图 6b 所示, train_loss 随着 step 的增加迅速下降至 0 附近, 随后在该附近小幅度波动。经过 139 个 step 之后, train_loss 几乎不再变化, 模型保持稳定。

2.1.4 8B 模型训练 8B 模型使用监督式微调和基于人工反馈的强化学习优化过, 同样具有很强的指令遵循能力。如图 7a 所示, 模型在训练集和验证集上的拟合表现一致, 没有过拟合或欠拟合现象, 该过程的学习率为 5.0×10^{-5} 。最后, 固定学习率并在整个数据集上进行训练。如图 7b 所示, train_loss 随着 step 的增加呈现出波动式下降的趋势, 这可能是由该阶段的学习率较高或模型参数更新导致的。在经过 236 个 step 之后, train_loss 几乎不再变化, 模型保持稳定。

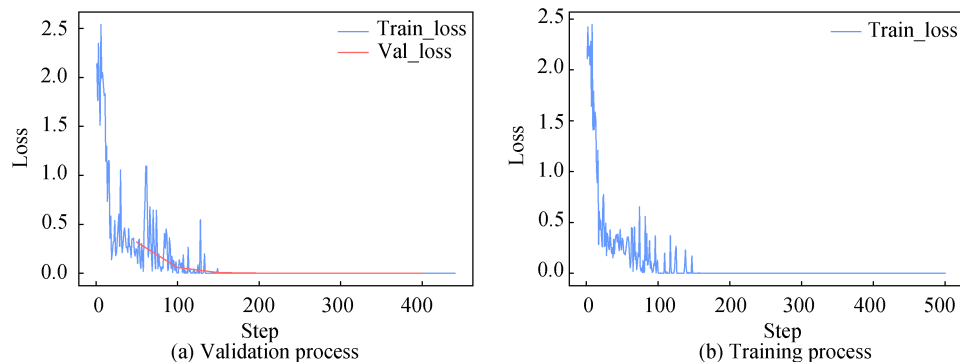


图 5 段落分类任务中 3B 模型的验证过程 loss 曲线和训练过程 loss 曲线

Fig. 5 Validation loss curve and training loss curve of the 3B model in the paragraph classification task

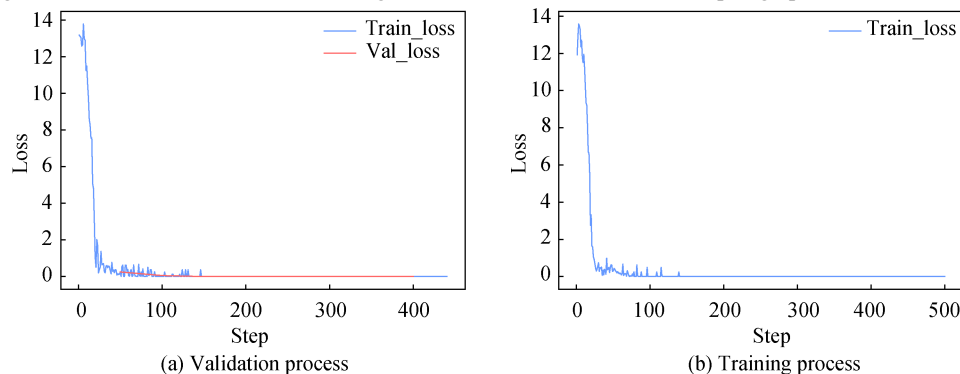


图 6 段落分类任务中 7B 模型的验证过程 loss 曲线和训练过程 loss 曲线

Fig. 6 Validation loss curve and training loss curve of the 7B model in the paragraph classification task

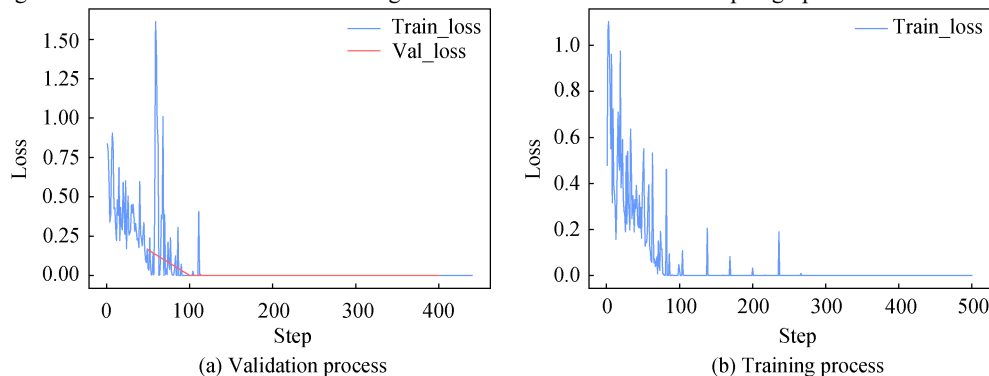


图 7 段落分类任务中 8B 模型的验证过程 loss 曲线和训练过程 loss 曲线

Fig. 7 Validation loss curve and training loss curve of the 8B model in the paragraph classification task

2.1.5 BERT 模型训练 BERT 模型是 Google 推出的 transformers 模型,以自监督的方式在大规模英文语料上进行了预训练,在文本分类、文本生成、机器翻译等 NLP 领域被广泛应用。BERT 模型有 large 和 base 2 种规模,为了更好地挖掘语义特征,本研究使用 large 版本模型。BERT 模型的文本预处理方式有 uncased 和 cased 2 种,分别表示将文本转为小写形式和保留文本大小写,文本分类任务对大小写不敏感,使用 uncased 处理方式可以降低模型的训练复杂度。BERT 模型微调过程如图 8 所示,微调操作和上面章节的微调过程相同,该过程的学习率为 3.0×10^{-6} 。

主要的微调超参数见表 1,其中学习率(learning_rate)决定了在每次迭代中模型权重更新的幅度。学习率如果太高,模型可能会在最优解附近震荡,甚至发散;如果太低,训练过程可能会非常缓慢,甚至陷入局部最优解。batch_size 是指在一次迭代中用于训练模型的样本数量。较小的批量大小可

以提供更频繁的梯度更新,有助于模型更快地收敛,但可能会增加噪声,导致训练不稳定;较大的批量大小可以提供更稳定的梯度估计,减少噪声,但可能会使训练过程变慢,并且需要更多的显存。由于 batch_size 大于 1 会导致显存溢出,因此,只能设置为 1。梯度累积允许模型在不增加内存消耗的情况下,模拟较大批次的训练效果。模型在每个批次上进行前向和反向传播,但不立即更新权重,而是累积一定步数的梯度,然后才进行一次权重更新。gradient_accumulation_step 就是训练时指定的梯度累积步数,这是 LLaMA-Factory 设置的默认值。epochs 表示整个训练数据集被用来训练模型的次数,较小的 epochs 会使得模型没有足够时间学习训练数据的复杂性,导致欠拟合;较大的 epochs 可以帮助模型更好地学习数据特征,但也可能导致过拟合,增加训练时间成本。根据训练数据量并经过多次调整,确定了段落分类任务中 epochs 值。

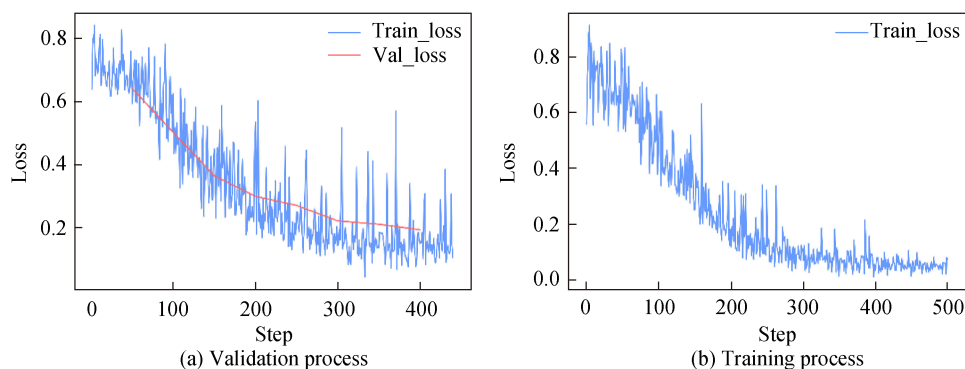


图 8 段落分类任务中 BERT 模型的验证过程 loss 曲线和训练过程 loss 曲线

Fig. 8 Validation loss curve and training loss curve of the BERT model in the paragraph classification task

表 1 段落分类任务中的微调超参数值

Table 1 Fine-tuning hyperparameter values in paragraph classification task

Model	Learning_rate	Batch_size	Gradient_accumulation_steps	Epochs
LLaMA-3.2-3B-Instruct	5.0×10^{-5}	1	8	20
Gemma-7B-Instruct	5.0×10^{-5}	1	8	20
LLaMA3.1-8B-Instruct	5.0×10^{-5}	1	8	20
BERT	3.0×10^{-6}	1	8	20

2.1.6 段落分类评估 图 9 中, NSSE 表示非固态电解质段落, SSE 表示固态电解质段落。NB 模型在精确率、召回率和 F1 分数 3 个指标上都略低于其他模型,这可能是因为 LLMs 和 BERT 在大量文

本数据上的预训练以及在段落分类数据上的微调,使其具有较强的语言理解能力。3B 和 8B 模型的整体表现相当,可能是由于 3B 模型来自 8B 模型的剪枝和知识蒸馏,两者具有相似的性能。7B 模型的表现最好,但是由于本工作的分类任务仅限于固态电解质合成段落的分类,而且训练和测试数据较少,不代表 7B 模型总是优于其他模型。BERT 模型的表现与 LLMs 相当或略差于 LLMs,如果不是追求极致的分类效果,而是更注重训练效率和成本,可以选择微调 BERT 模型。3B 和 8B 模型的对比可以表明,在某些情况下,小参数模型也能比大参数模型更好地完成任务,不必一味追求参数更大的模型,而是应根据具体任务的需求做出合理的选择。

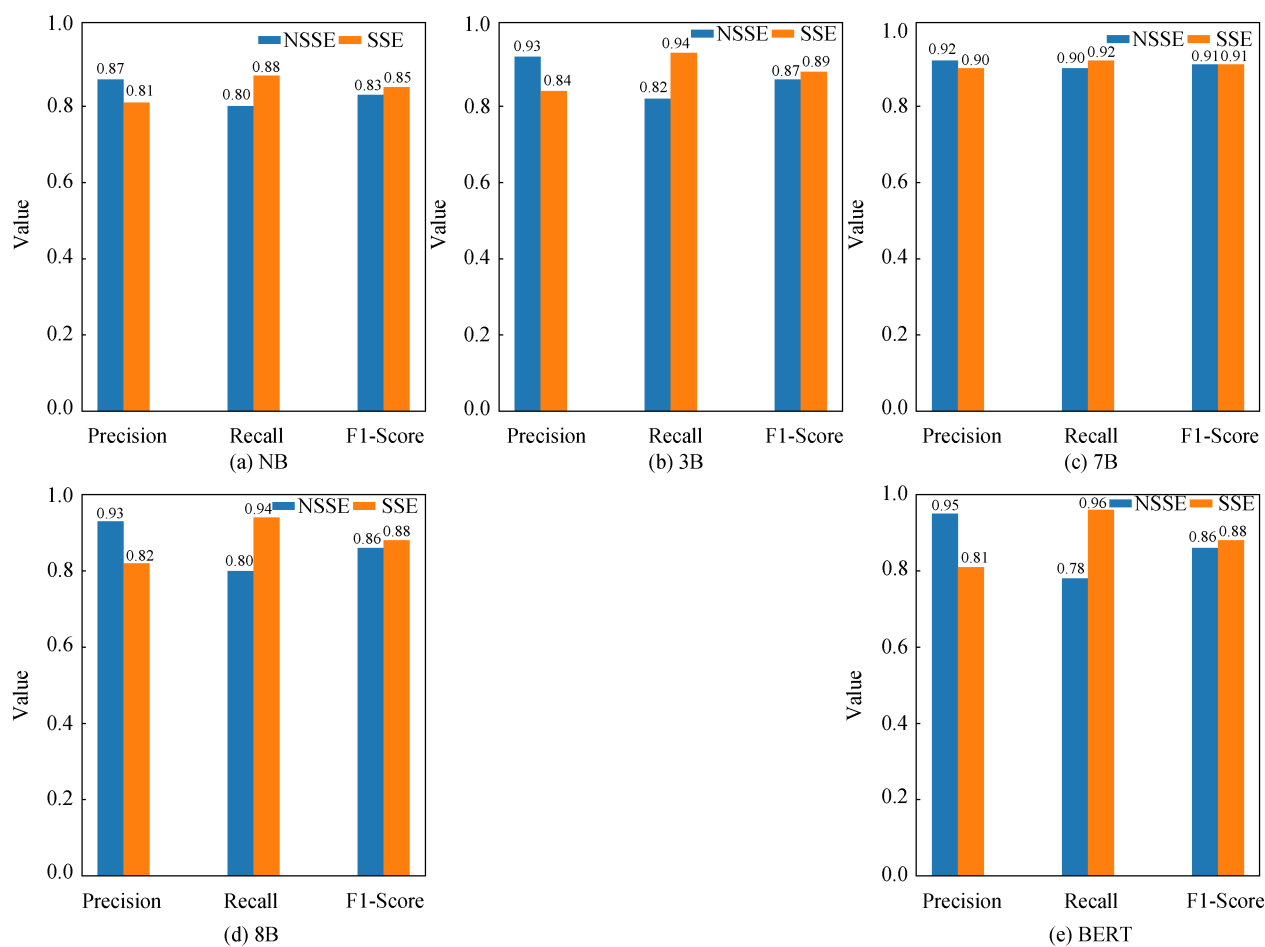


图9 朴素贝叶斯, LLaMA3.2-3B, Gemma-7B, LLaMA3.1-8B, BERT 段落分类效果评估

Fig. 9 Evaluation of paragraph classification performance for Naive Bayes, LLaMA3.2-3B, Gemma-7B, LLaMA3.1-8B, and BERT

在图 10 中, 类别 NSSE 表示非固态电解质段落类别, 类别 SSE 表示固态电解质段落类别。可以看出, BERT 模型在识别固态电解质段落方面表现出色, 识别出 48 个真正的正例(TP)和仅 2 个假的负例(FN), 但是在识别非固态电解质段落方面表现逊色, 结果为 39 个真正的正例(TP)和 11 个假的负例(FN)。7B 模型在段落分类任务中的综合表现最好, 对段落描述具有较强的理解和区分能力。在错误分类中, 将标签 NSSE 预测为 SSE 的情况比较多, 这可能是由于非固态电解质段落中包含了电极制备和电池组装的描述, 其中涉及电解质的使用, 导致模型将这些段落判断为固态电解质段落。为进一步提升 LLMs 的分类效果, 可以在构建数据集时调整提示方式, 以便让 LLMs 能更好地理解不同段落之间的细微差别。

2.2 固态电解质合成信息提取

2.2.1 3B 模型训练 微调过程同 2.1.2 章节, 经过多次调整确定合适的学习率为 8.0×10^{-6} 。如图 11a 所示, 模型在训练集和验证集上的拟合表现一致, 没有明显的过拟合或欠拟合现象。最后, 固定学习

率并在整个数据集上进行训练。如图 11b 所示, train_loss 随着 step 的增加呈现出波动式下降的趋势, 说明模型正在学习并逐渐提高其性能。在大约 400 个 step 之后, train_loss 不再呈现下降趋势, 表明模型在训练集上达到了较好的拟合度。

2.2.2 7B 模型训练 微调过程同 2.1.3 章节, 经过多次调整确定合适的学习率为 5.0×10^{-6} 。如图 12a 所示, 模型在训练集和验证集上的拟合表现一致, 没有明显的过拟合或欠拟合现象。最后, 固定学习率并在整个数据集上进行训练。如图 12b 所示, train_loss 随着 step 的增加呈现出波动式下降的趋势, 但是波动幅度较小, 模型在训练集上达到了较好的拟合度。

2.2.3 8B 模型训练 微调过程同 2.1.4 章节, 经过多次调整确定合适的学习率为 5.0×10^{-6} 。如图 13a 所示, 模型在训练集和验证集上的拟合表现一致, 没有明显的过拟合或欠拟合现象。最后, 固定学习率并在整个数据集上进行训练。如图 13b 所示, 8B 模型的 train_loss 变化趋势与 3B 模型的微调过程相似。

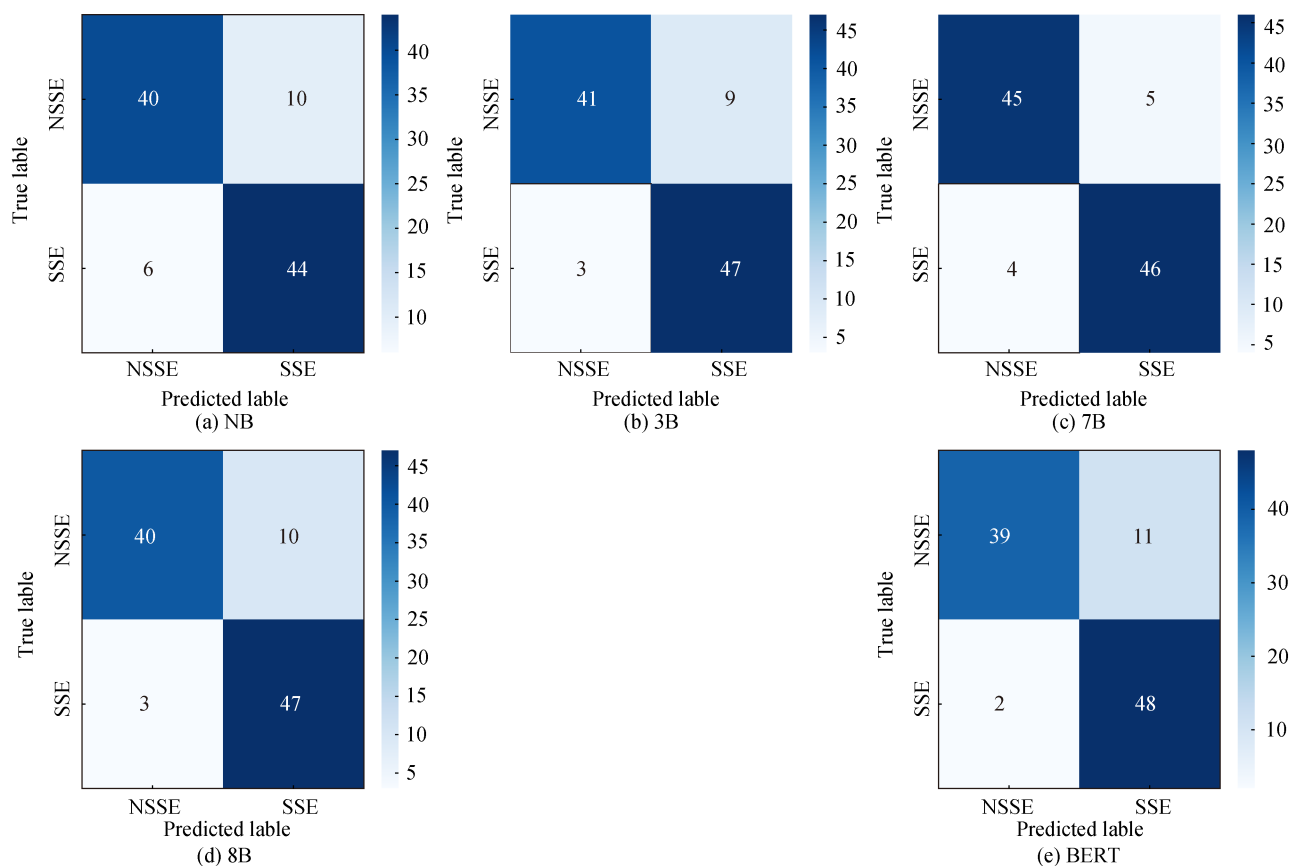


图 10 朴素贝叶斯, LLaMA3.2-3B, Gemma-7B, LLaMA3.1-8B, BERT 段落分类混淆矩阵

Fig. 10 Confusion matrix for paragraph classification using Naive Bayes, LLaMA3.2-3B, Gemma-7B, LLaMA3.1-8B, and BERT

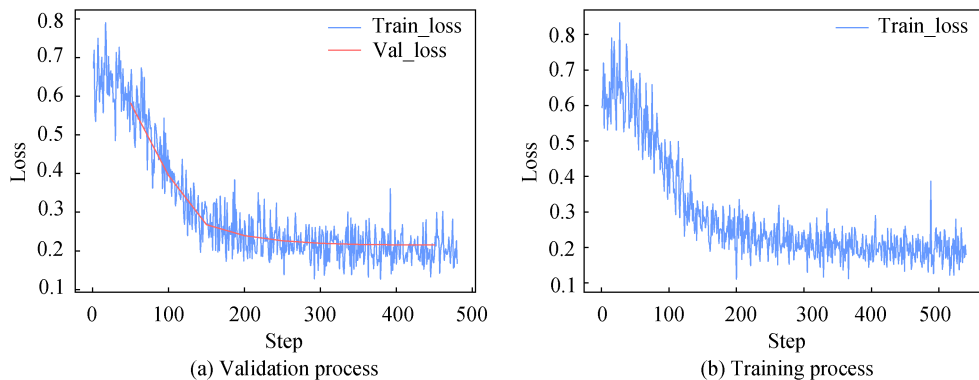


图 11 合成信息提取任务中 3B 模型的验证过程 loss 曲线和训练过程 loss 曲线

Fig. 11 Validation loss curve and training loss curve of the 3B model in synthesis information extraction task

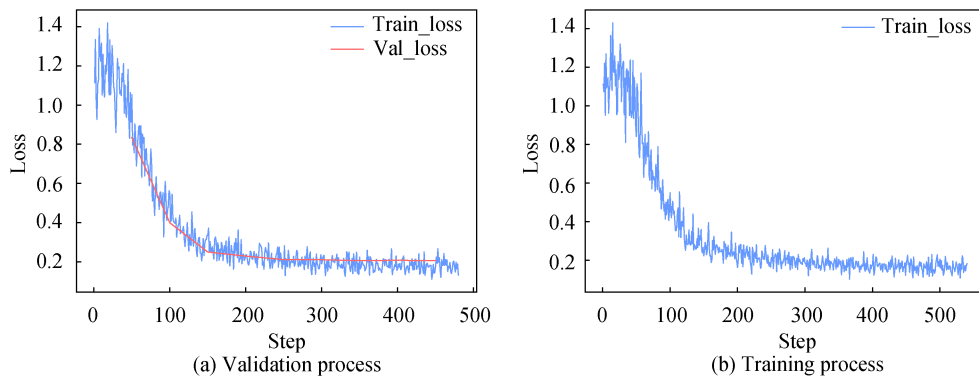


图 12 合成信息提取任务中 7B 模型的验证过程 loss 曲线和训练过程 loss 曲线

Fig. 12 Validation loss curve and training loss curve of the 7B model in synthesis information extraction task

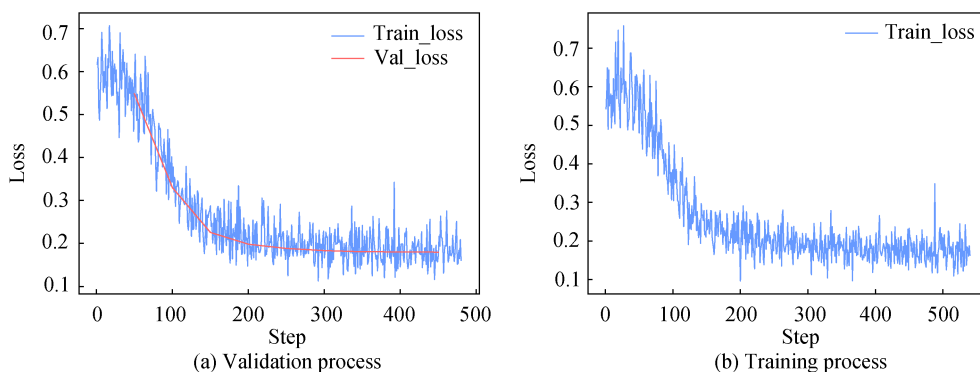


图 13 合成信息提取任务中 8B 模型的验证过程 loss 曲线和训练过程 loss 曲线

Fig. 13 Validation loss curve and training loss curve of the 8B model in synthesis information extraction task

在合成信息任务中，主要的微调超参数见表 2，参数的具体介绍和选择过程同 2.1 章节，鉴于该任务的训练数据集规模较小，适当增加了 epochs 值，以便模型能够更充分地学习合成信息提取的过程。

表 2 信息提取任务中的微调超参数值

Table 2 Fine-tuning hyperparameter values in paragraph extraction task

Model	Learning_rate	Batch_size	Gradient_accumulation_steps	Epochs
LLaMA-3.2-3B-Instruct	8.0×10^{-6}	1	8	30
Gemma-7B-Instruct	5.0×10^{-6}	1	8	30
LLaMA3.1-8B-Instruct	5.0×10^{-6}	1	8	30

2.2.4 合成信息提取评估 选取 50 个分类模型筛选出的固态电解质合成段落进行测试，分别输入微调好的 3B、7B 和 8B 3 个模型进行提取，然后统计每个模型提取的每个段落的实体数量，如图 14 所示，最后使用精准率、召回率和 F1 分数 3 个指标综合评估模型的提取效果，计算方法见 2.1.6 章节。

从图 15 可以看出，3B 模型在精准率、召回率和 F1 分数 3 个指标上的平均值都超过了 0.9，说明

该模型提取固态电解质合成信息的表现出色。尽管存在一定的波动，但模型的性能指标在大多数情况下都保持在较高水平，这表明模型具有一定的稳定性。有一个段落的提取精确率为 0.7，是由于模型提取了规定的 4 类实体范围之外的信息，导致结果偏低，其他精确率低于平均值的段落同样出现了或多或少地提取其他无关的实体的情况。召回率的波动相对较大，对于低于平均值的段落，模型提取时遗漏了较多必要的实体。

从图 16 可以看出，7B 模型在精准率、召回率和 F1 分数 3 个指标上的平均值都超过了 0.9，说明该模型提取固态电解质合成信息的表现同样出色，也具有一定的稳定性。有一个段落的召回率低于 0.6，是由于该模型遗漏了较多的提取动作实体信息。7B 模型的综合提取效果稍逊于 3B 和 8B 模型。

从图 17 可以看出，8B 模型在精准率、召回率和 F1 分数 3 个指标上的平均值都超过了 0.9，说明该模型提取固态电解质合成信息的表现同样出色，也具有一定的稳定性。该模型在提取过程中也出现了与其他模型相同的问题，包括提取了无关实体、遗漏必要实体。

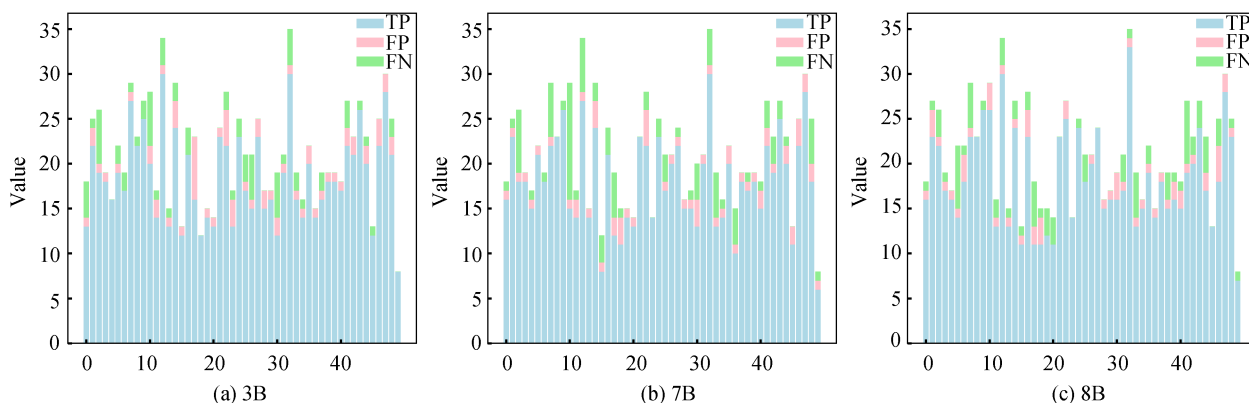


图 14 3B 模型，7B 模型，8B 模型实体提取数量统计

Fig. 14 Entity extraction quantity statistics for the 3B model, 7B model, and 8B model

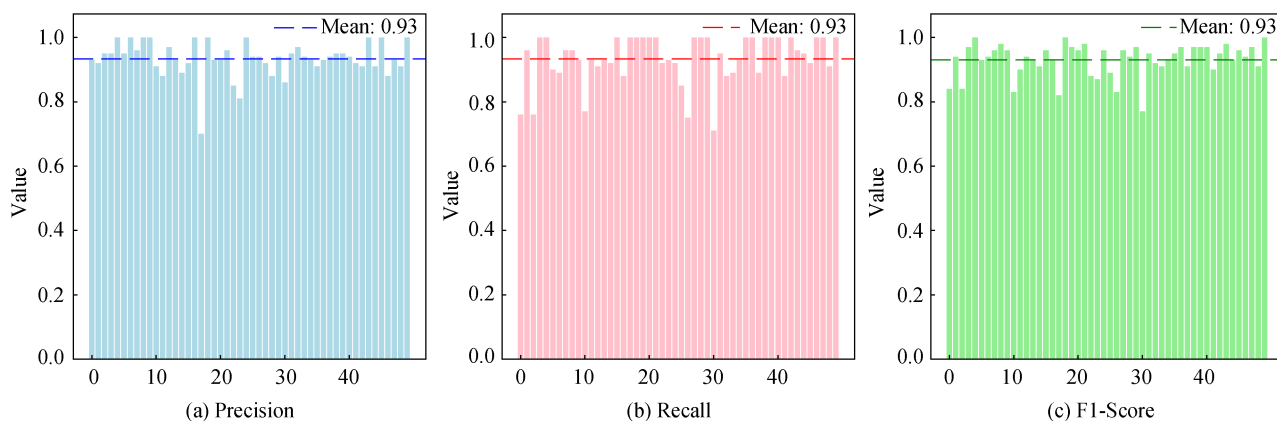


图 15 精确率、召回率和 F1 分数评估 3B 模型

Fig. 15 Evaluation of the 3B model using Precision, Recall, and F1 Score

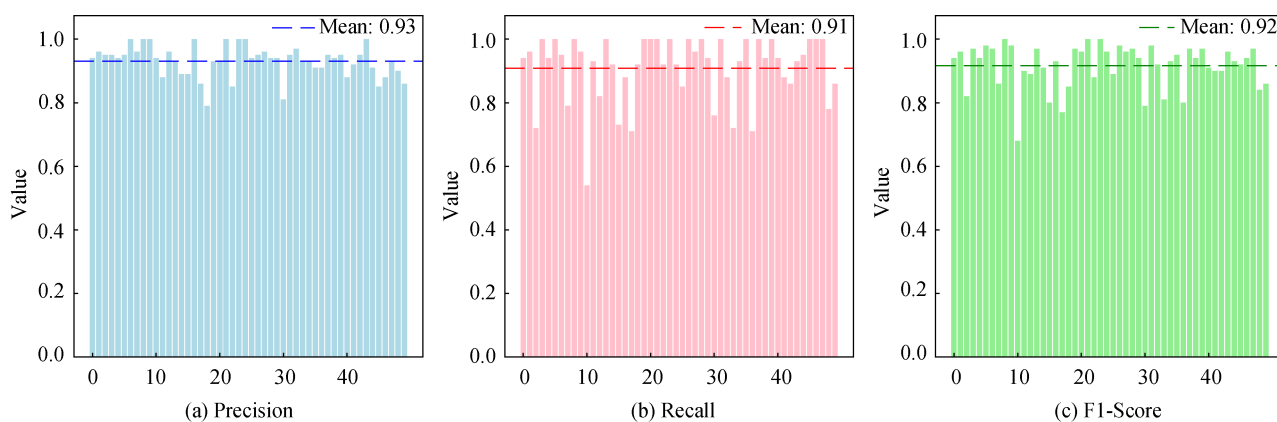


图 16 精确率、召回率和 F1 分数评估 7B 模型

Fig. 16 Evaluation of the 7B model using Precision, Recall, and F1 Score

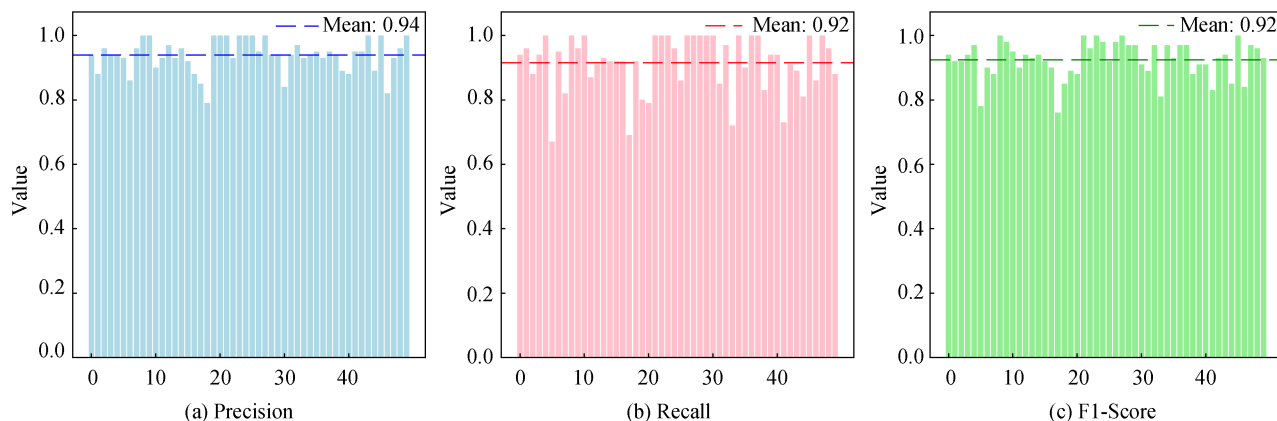


图 17 精确率、召回率和 F1 分数评估 8B 模型

Fig. 17 Evaluation of the 8B model using Precision, Recall, and F1 Score

3 个 LLMs 在固态电解质合成信息提取任务中的表现出色, 说明预训练加微调的方式可以充分发挥大模型的文本理解和指令遵循的能力, 能够胜任固态电解质合成过程中重要信息提取任务。而且使用的微调数据集来自无机催化剂领域, 在固态电解

质领域表现出色, 也说明了大模型的泛化能力强, 学习到了从非结构化文本提取结构化信息的能力, 而非提取固定材料和固定合成过程信息的能力。3B、7B 和 8B 模型的提取效果相当, 表明模型的大小并不是决定性能的唯一因素, 模型的设计和训练

策略因素同样重要。

3 结论

1) 在段落分类任务中,大语言模型和 BERT 模型得益于其在大规模文本数据上的预训练以及针对特定分类任务的微调,展现出了超越贝叶斯模型的分性能。在不同规模大语言模型之间的比较中,3B 模型来自 8B 模型的剪枝和知识蒸馏,两者具有相似的分类效果,7B 模型的 F1 分数超过了 0.9,凸显了其在处理此类任务时的优越性能。

2) 在合成信息提取任务中,使用无机催化剂合成信息提取数据集微调大语言模型,结果 3 个模型的 F1 分数都超过了 0.9,体现了大语言模型的强大的泛化能力,说明大语言模型学习到了从非结构化文本提取结构化信息的能力,而非提取固定材料和固定合成过程信息的能力。3B 模型的提取效果略优于 7B 和 8B 模型,表明参数大小不是决定性能的唯一因素,模型的设计和训练策略因素同样重要。

参考文献:

- [1] XIE J, LU Y C. A retrospective on lithium-ion batteries[J]. *Nat Commun*, 2020, 11(1): 2499.
- [2] YU X Q, CHEN R S, GAN L Y, et al. Battery safety: From lithium-ion to solid-state batteries[J]. *Engineering*, 2023, 21: 9–14.
- [3] ZHOU G, HUANG Q, ZHANG Q, et al. Thermal insulation phase-change hydrogel with enhanced mechanical properties for inhibiting thermal runaway propagation in lithium-ion battery module[J]. *J Energy Storage*, 2024, 102: 114102.
- [4] YANG Z, YAN Y, PAK U. Thermal reliability assessment and sensitivity analysis of 18, 650 cylindrical lithium-ion battery[J]. *J Energy Storage*, 2023, 59: 106504.
- [5] 李义函, 卢世刚, 王晶, 等. 磷酸铁锂离子电池低温不可逆析锂及其对电池性能衰减的影响[J]. *储能科学与技术*, 2024, 13(10): 3656–3665.
LI Yihan, LU Shigang, WANG Jing, et al. *Energy Storage Sci Technol*, 2024, 13(10): 3656–3665.
- [6] YANG Q F, LI C L. Li metal batteries and solid state batteries benefiting from halogen-based strategies[J]. *Energy Storage Mater*, 2018, 14: 100–117.
- [7] ZHANG X Y, CHENG S C, FU C K, et al. Advancements and challenges in organic–inorganic composite solid electrolytes for all-solid-state lithium batteries[J]. *Nano Micro Lett*, 2024, 17(1): 2.
- [8] CHEN X C, LIU X M, SAMUTHIRA PANDIAN A, et al. Determining and minimizing resistance for ion transport at the polymer/ceramic electrolyte interface[J]. *ACS Energy Lett*, 2019, 4(5): 1080–1085.
- [9] LUO W, GONG Y H, ZHU Y Z, et al. Transition from superlithiophobicity to superlithiophilicity of garnet solid-state electrolyte[J]. *J Am Chem Soc*, 2016, 138(37): 12258–12262.
- [10] HARRY K J, LIAO X X, PARKINSON D Y, et al. Electrochemical deposition and stripping behavior of lithium metal across a rigid block copolymer electrolyte membrane[J]. *J Electrochem Soc*, 2015, 162(14): A2699–A2706.
- [11] KIM J, MOK D H, KIM H, et al. Accelerating the search for new solid electrolytes: Exploring vast chemical space with machine learning-enabled computational calculations[J]. *ACS Appl Mater Interfaces*, 2023.
- [12] HAWIZY L, JESSOP D M, ADAMS N, et al. ChemicalTagger: A tool for semantic text-mining in chemistry[J]. *J Cheminform*, 2011, 3(1): 17.
- [13] SWAIN M C, COLE J M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature[J]. *J Chem Inf Model*, 2016, 56(10): 1894–1904.
- [14] MAVRAČIĆ J, COURT C J, ISAZAWA T, et al. ChemDataExtractor 2.0: Autopopulated ontologies for materials science[J]. *J Chem Inf Model*, 2021, 61(9): 4280–4289.
- [15] LIU Y, DING L, YANG Z W, et al. Domain knowledge discovery from abstracts of scientific literature on Nickel-based single crystal superalloys[J]. *Sci China Technol Sci*, 2023, 66(6): 1815–1830.
- [16] WANG W R, JIANG X, TIAN S H, et al. Automated pipeline for superalloy data by text mining[J]. *NPJ Comput Mater*, 2022, 8: 9.
- [17] YAN R E, JIANG X, WANG W R, et al. Materials information extraction via automatically generated corpus[J]. *Sci Data*, 2022, 9(1): 401.
- [18] VAUCHER A C, ZIPOLI F, GELUYKENS J, et al. Automated extraction of chemical synthesis actions from experimental procedures[J]. *Nat Commun*, 2020, 11(1): 3601.
- [19] GUPTA T, ZAKI M, ANOOP KRISHNAN N M, et al. MatSciBERT: A materials domain language model for text mining and information extraction[J]. *NPJ Comput Mater*, 2022, 8: 102.
- [20] GUO J, SANTIAGO IBANEZ-LOPEZ A, GAO H Y, et al. Automated chemical reaction extraction from scientific literature[J]. *J Chem Inf Model*, 2022, 62(9): 2035–2045.
- [21] WANG Z R, CRUSE K, FEI Y X, et al. ULSA: Unified language of synthesis actions for the representation of inorganic synthesis protocols[J]. *Digit Discov*, 2022, 1(3): 313–324.
- [22] CRUSE K, TREWARTHA A, LEE S, et al. Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities[J]. *Sci Data*, 2022, 9(1): 234.
- [23] Zheng Z, Rampal N, Inizan T J, et al. Large language models for reticular chemistry[J]. *Nat Rev Mater*, 2025, <https://doi.org/10.1038/s41578-025-00772-8>.
- [24] HONG Z J. ChatGPT for computational materials science: A perspective[J]. *Energy Mater Adv*, 2023, 4: 26.
- [25] Zheng Z, Zhang O, Borgs C, et al. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis[J]. *J Am Chem Soc*, 2023, 145(32): 18048–18062.
- [26] BOIKO D A, MACKNIGHT R, KLINE B, et al. Autonomous chemical research with large language models[J]. *Nature*, 2023, 624(7992): 570–578.
- [27] DAGDELEN J, DUNN A, LEE S, et al. Structured information extraction from scientific text with large language models[J]. *Nat Commun*, 2024, 15(1): 1418.
- [28] ANSARI M, MOOSAVI S M. Agent-based learning of materials datasets from the scientific literature[J]. *Digit Discov*, 2024, 3(12): 2607–2617.
- [29] 时宗彬, 朱丽雅, 乐小虬. 基于本地大语言模型和提示工程的材料信息抽取方法研究[J]. *数据分析与知识发现*, 2024, 8(7): 23–31.
SHI Zongbin, ZHU Liya, LE Xiaoli, et al. *Data Ana Know discov*, 2024, 8(7): 23–31.
- [30] Li S, Zhang Y, Fang Z, et al. Extracting the Synthetic Route of Pd-Based Catalysts in Methanol Steam Reforming from the Scientific Literature[J]. *J Chem Inf Model*, 2023, 63(20): 6249–6260.

Large Language Models Extract Synthesis Information of Lithium-ion Battery Solid-state Electrolytes from Literature

WEI Shihao, LI Shuyuan, WANG Yaxin, SUN Shaorui

(College of Materials Science and Engineering, Beijing University of Technology, Beijing 100124, China)

Extended Abstract

Introduction Lithium-ion batteries are widely used in electric vehicles due to their high energy density and environmental compatibility, but their safety and endurance are limited by liquid-state electrolytes. Solid-state electrolytes have attracted recent attention for their high safety and energy density, but their research faces some challenges. Artificial intelligence technologies, such as machine learning, can accelerate the discovery and optimization of solid-state electrolyte materials. However, extracting synthesis information from literature is time-consuming and prone to errors. Large language models (LLMs) become popular in the field of natural language processing. This paper was to propose the use of LLMs to automatically extract solid-state electrolyte synthesis information from literature *via* building datasets to train and fine-tune models.

Methods The process of extracting synthesis information for solid-state electrolytes included five parts, *i.e.*, literature downloading, parsing and processing, paragraph classification, information extraction, and synthesis process visualization. Initially, the downloaded XML documents were parsed to extract all paragraphs under the headings of "Experimental" or "Method". A classification dataset was constructed using these paragraphs, with the training set used to train and fine-tune a Bayesian model, a BERT model, and three different LLMs. The test set was used to evaluate the classification performance of different models. Three LLMs are fine-tuned using a dataset previously built for the extraction of synthesis information in inorganic catalysts, enabling them to complete the task of extracting synthesis information for solid-state electrolytes, outputting four parts of contents, *i.e.*, product, raw materials, methods, and steps, and testing and evaluating with the synthesis paragraphs filtered by the classification models.

Results and discussion The NB model slightly underperforms the other models in terms of precision, recall, and F1 score. This may be attributed to the pre-training of LLMs and BERT on extensive text data and their fine-tuning on paragraph classification data, endowing them with strong language comprehension abilities. The comparable overall performance of the 3B and 8B models may be attributed to the 3B model being derived from the 8B model through pruning and knowledge distillation, resulting in similar performance characteristics. The classification task in this paper is limited to the classification of solid-state electrolyte synthesis paragraphs and the training and testing data are relatively scarce, the superior performance of the 7B model does not imply that it is always superior to other models. The comparison between the 3B and 8B models demonstrates that in certain scenarios, smaller-parameter models can outperform their larger-parameter counterparts. Therefore, it is not always necessary to pursue models with more parameters. Instead, reasonable choices should be made based on the specific requirements of the task. In misclassifications, there is a higher incidence of predicting label 0 as 1, which may be due to the inclusion of descriptions of electrode preparation and battery assembly in paragraphs about non-solid electrolytes, involving the use of electrolytes, leading to the model to misjudge these paragraphs as solid-state electrolyte paragraphs. To further enhance the classification performance of LLMs, the way prompts the construction of the dataset that can be adjusted to allow LLMs to better understand the subtle differences between different paragraphs.

The three LLMs perform exceptionally well in the task of extracting synthesis information for solid-state electrolytes, demonstrating that the combination of pre-training and fine-tuning can fully leverage the text understanding and instruction following capabilities of large models, making them competent for the crucial task of extracting important information from the synthesis process of solid-state electrolytes. Moreover, the fine-tuning dataset used originates from the field of inorganic catalysts and shows excellent performance in the field of solid-state electrolytes, indicating a great generalization ability of large models. They learn the ability to extract structured information from unstructured text, rather than just extracting information about fixed materials and synthesis processes. The extraction effects of the 3B, 7B, and 8B models are comparable, indicating that the size of the model is not a sole factor determining performance. The design and training strategies of the models are equally important.

Conclusions In the paragraph classification task, LLMs and BERT, benefiting from their pre-training on large-scale text data and fine-tuning for specific classification tasks, demonstrated classification performance that surpassed that of Bayesian models. Among the LLMs of different sizes, the 3B model, derived from the 8B model through pruning and knowledge distillation, showed similar classification effectiveness to the 8B model. The 7B model achieved an F1 score exceeding 0.9, thus highlighting its superior performance in handling such tasks.

In the synthesis information extraction task, LLMs were fine-tuned using a dataset of inorganic catalyst synthesis information. The results showed that the F1 scores of all three models exceeded 0.9, demonstrating great generalization capabilities of LLMs. This indicated that LLMs could learn the ability to extract structured information from unstructured text, rather than just extracting information about fixed materials and synthesis processes. The extraction performance of the 3B model was slightly better than that of the 7B and 8B models, thus indicating that the number of parameters could not be a sole factor determining performance. The design and training strategies of the models could be equally important.

Keywords large language models; lithium-ion batteries; solid-state electrolytes; extraction