

AI 辅助先进电池设计与应用专刊



基于大语言模型 RAG 架构的电池加速研究: 现状与展望

钟 逸¹, 冷 彦¹, 陈思慧¹, 李培义¹, 邹 智¹, 刘 洋², 万佳雨¹

(¹上海交通大学溥渊未来技术学院, 未来电池研究中心, 上海 200240; ²昆山杜克大学数据科学研究中心, 江苏 昆山 215316)

摘要: 随着近年电池领域研究投入的激增, 研究人员面临着前所未有的信息过载和知识盲区的挑战。针对这一问题, 本文探讨了大语言模型(large language model, LLM)的检索增强生成(retrieval augmented generation, RAG)架构在电池领域的应用潜力, 在此基础上对近期的研究文献进行综述, 并提出展望。本文介绍了大语言模型 RAG 架构的工作原理, 强调了该架构在垂直领域的可靠性, 并基于此综述探讨了该架构在电池材料设计、电池单元设计和制造、电动交通与电网的电池管理系统三个领域的潜在应用。在电池材料设计部分, 本文着重分析了大语言模型 RAG 架构的无幻觉生成能力在数据提取、研究方案设计和多模态数据问答中的优势。在电池单元设计和制造部分, 本文从科研端指出该架构对电池单元设计方案分析的辅助作用, 从制造端指出该架构桥接产业和科研的鸿沟、辅助产业管控的作用。在电动交通和电网的电池管理系统部分, 本文指出该架构在跨领域知识联结、辅助系统级运维的作用。最后, 本文讨论了多模态 RAG 技术在电池研究领域的应用潜力及其对电池研究效率的提升, 并展望了 RAG 在电池领域的更多应用前景。

关键词: 大语言模型; 检索增强生成; 电池材料; 电芯; 电池管理系统

doi: 10.19799/j.cnki.2095-4239.2024.0604

中图分类号: O 6-39

文献标志码: A

文章编号: 2095-4239 (2024) 09-3214-12

Accelerating battery research with retrieval-augmented large language models: Present and future

ZHONG Yi¹, LENG Yan¹, CHEN Sihui¹, LI Peiyi¹, ZOU Zhi¹, LIU Yang², WAN Jiayu¹

(¹Future Battery Research Center, Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai 200240, China; ²Data Science Research Center, Duke Kunshan University, Kunshan 215316, Jiangsu, China)

Abstract: In recent years, the surge in research investment within the battery field has presented researchers with challenges of information overload and knowledge gaps. This study examines the Retrieval-Augmented Generation (RAG) architecture of large language models in the battery domain, offering a review of contemporary research and future prospects. We describe the working principles of the RAG architecture, affirm its reliability in specialized domains, and discuss its applications across three key areas as follows: battery material design, battery cell design and manufacturing, and battery management systems for e-mobility and electric grids. In the section on battery material design, the study highlights the hallucination-free generation capabilities of RAG in data extraction, research protocol design,

收稿日期: 2024-07-03; 修改稿日期: 2024-08-11。

第一作者: 钟逸 (2005—), 男, 本科, 从事人工智能的可再生能源应用, E-mail: zhongyeah@sjtu.edu.cn; 冷彦 (2001—), 男, 硕士研究生, 从事锂离子电池先进材料的开发、人工智能的可再生能源应用, E-mail: ly234244@sjtu.edu.cn; 通信作者: 万佳雨, 副教授, 研究方向为储能材料与器件、先进制造, E-mail: wanjiy@sjtu.edu.cn; 刘洋, 副教授, 从事新能源、人工智能, E-mail: yang.liu2@dukekunshan.edu.cn。

引用本文: 钟逸, 冷彦, 陈思慧, 等. 基于大语言模型 RAG 架构的电池加速研究: 现状与展望[J]. 储能科学与技术, 2024, 13(9): 3214-3225.

Citation: ZHONG Yi, LENG Yan, CHEN Sihui, et al. Accelerating battery research with retrieval-augmented large language models: Present and future[J]. Energy Storage Science and Technology, 2024, 13(9): 3214-3225.

and multimodal data querying. The section on battery cell design and manufacturing elucidates RAG's role in enhancing research-driven battery cell design and bridging the gap between industry and academia, thereby aiding industrial control processes. The discussion on battery management systems for e-mobility and electric grids underscores RAG's contribution to cross-domain knowledge integration and system-level operation and maintenance support. The study concludes by considering the application of multimodal RAG technology in battery research and anticipates further expansion of RAG applications in this field.

Keywords: large language model; retrieval augmented generation; battery material; battery cell; battery management system

1 绪 论

1.1 电池研究现状

近年来, 电池研究和产业领域取得了显著的成就。从微观层面对电池材料的深入研究, 到器件层面电芯的设计和制造, 再到宏观层面电动交通工具和储能电站的电池管理系统(**battery management system, BMS**), 众多创新技术不断涌现。这些技术进步推动了电池领域学术论文、研究报告、专利、学术会议数量的急剧增长, 每年发表的相关研究文献数以万计。因此, 领域信息增长速度迅猛, 数据量庞大且呈现多模态、高度分散的特点。这导致研究和产业人员经常面临信息过载和知识盲区的挑战, 严重影响了研究的效率和创新的速度。在这种背景下, 传统的研究范式显得效率不足, 亟需改进以适应快速发展的行业需求。

1.2 大语言模型的优势

大语言模型的出现为提高研究效率提供了可行方案。首先, 相较**BERT**等语言模型, 大语言模型在具有同样强大的文本理解能力的同时, 还具有强大的语言生成能力。这种强大的生成能力源于其基于**Transformer**架构的大规模预训练和自回归生成机制^[1]。具体而言, 大语言模型通过在海量文本数据上进行预训练, 学会捕捉语言的复杂模式、语法规则和上下文关系, 并通过自回归机制在生成时逐步预测每一个词语, 从而能够在广泛的上下文中生成连贯且富有逻辑性的内容。这种架构使得模型不仅能理解输入, 还能基于海量参数生成高质量的输出, 与研究人员进行交互, 进行知识总结、想法生成、方法论建议等, 作为研究人员的助手^[2]。

在强大的语言生成能力之外, 当今大语言模型

正在向多模态数据理解和生成发展, 如代码、图表、图像、视频等。例如, 较早的**CLIP(contrastive language-image pre-training)**模型基于文字-图像数据集训练, 已经实现从文字到图像的转化^[3]。而现今, **ChatGPT-4o**大语言模型能做到代码、图表、图像、声音等多模态数据的理解和生成, 而**Sora**模型则将大模型的多模态能力扩展到了视频生成领域^[4-5]。这使得大语言模型的应用范畴远不止于文本问答^[6-7]。例如, 大语言模型的代码生成能力使得大模型能调用外部函数和库, 让大模型能够控制更多资源, 解决具有一定难度的研究问题, 如化学实验领域大语言模型**Coscientist**, 金属有机框架生成大语言模型**ChatMOF**等^[8-10]。

然而, 现有的大语言模型更像是一个“全科医生”而不是“专科医生”, 其在处理垂直领域任务时, 生成的结果存在幻觉(**hallucination**)问题。幻觉指的是大语言模型在回答时生成不符合事实或不符合输入提示词的内容^[11]。这种问题的根本原因在于大语言模型的训练机制。大语言模型依赖于海量的通用文本数据进行训练, 其目标是最大化上下文中的词语预测概率, 因此模型倾向于生成与训练数据中常见模式相符的内容, 而不是对特定领域知识的深刻理解。当面对不熟悉或特定领域的任务时, 模型可能会根据其训练中掌握的通用知识进行推测, 这就容易导致内容不准确或与事实不符的“幻觉”产生^[11]。对于专业领域研究, 错误的信息会导致严重的研究偏差, 所以避免幻觉对于搭建专业领域大语言模型尤为重要^[12]。

1.3 用**RAG(retrieval-augmented generation)**架构解决大语言模型的幻觉问题

为避免大语言模型的“幻觉”问题, 常用的方

法包括微调(fine-tuning)和检索增强生成 RAG(retrieval-augmented generation)。微调方法虽然能够训练针对特定领域的大语言模型，但微调很可能带来过拟合、灾难性遗忘等问题，需要特制领域数据库构建、层冻结、层级学习率衰减等复杂手段优化，并且很难保证整体运行稳定性^[13-15]。此外，若要对模型的微调训练数据库大规模更新，该模型就要重新微调，因此微调方法的数据库更新尤为困难^[16]。换言之，微调方法更适合静态的数据库。但是，电池研究是一个尚在增速的研究领域，偏向静态的领域微调模型难以胜任。

相比之下，采用 RAG 架构构建电池领域专业大语言模型更加简洁、经济、高效。大语言模型在零样本(zero-shot)输入时容易产生幻觉，但若在提示词中加入相关的背景知识，让大语言模型依据这些知识组织回答则能明显降低甚至消除幻觉。RAG 技术的核心原理在于，根据用户输入“检索”相关背景知识，并将其“增强”以形成提示词，进而输

入大语言模型，以生成无幻觉的回答^[17]。RAG 技术的基本架构(naive RAG)主要由索引(indexing)、检索(retrieval)和生成(generation)过程组成^[18]。该架构依赖于与用户输入相关的背景知识，而索引过程则是构建背景知识向量数据库的过程。它将不同格式的数据统一化，随后分块并使用嵌入模型(embedding model)储存入向量数据库。检索过程则是从数据库中提取相关知识的过程。该过程利用嵌入模型将用户询问向量化，随后与向量数据库比较相似度，检索出达到相似度阈值的数据。生成过程则将检索到的数据与用户输入整合，形成最终的提示词，提供大语言模型以生成输出^[17]。这种以大模型为中心的 RAG 架构，不仅保留了大语言模型的强大解析力、出色的鲁棒性和处理多模态数据的能力，还利用增强数据集增加了垂直领域问答的精准性和针对性。另外，模型与数据库在 RAG 架构中的相对独立性还降低了更新数据集的难度，这对于快速发展的电池研究领域来说尤为重要。

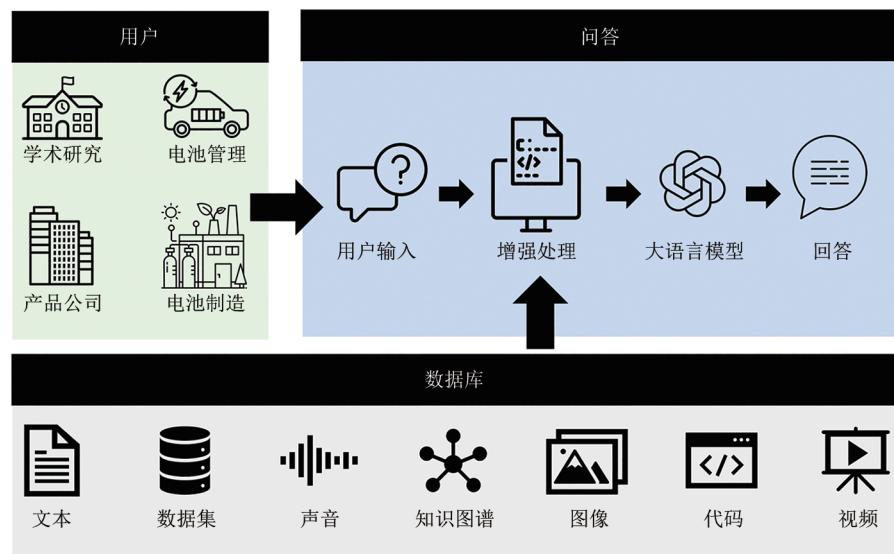


图 1 RAG 架构流程
Fig. 1 RAG pipeline

当然，上文提到的 RAG 初级架构还远不足以搭建一个成熟的领域机器人。例如，初级 RAG 架构的检索过程易受用户输入的质量影响，很可能出现查找偏差、遗漏重要信息。另外，在检索过程中，检索得到的数据可能包含大量重复或不连贯信息，直接作为大语言模型的提示词很可能使得大语言模型只做简单复述，或是产生本不存在的联系，导致幻觉^[18]。针对这些问题，RAG 架构衍生出了大量优化

方法，例如假设文档嵌入(hypothetical document embedding)，多查询检索(multi-query)等流程上的优化，以及模块化 RAG 等架构上的优化^[18-19]。此外，近期利用知识图谱(knowledge graph)构建增强数据库已成为 RAG 领域的新兴方向。知识图谱主要由实体(entities)、关系(relationships)和相应的属性(attributes)构成，利用知识图谱构建 RAG 架构的增强数据库，不仅能够显著提升模型对复杂知识的理

解能力, 还可以增强其在多领域知识间的联结与推理能力, 弥补传统语言模型在处理复杂语义和推理任务中的不足^[20]。渐趋完备的优化体系让大语言模型RAG架构能够胜任广泛的问答工作, 大大拓展了大语言模型RAG架构的使用场景。

2 大语言模型RAG架构在电池领域的具体应用

2.1 电池材料设计

RAG架构大大减少了大语言模型回答的幻觉, 能更好发挥大语言模型的生成能力优势, 辅助电池材料设计。电池材料设计研究主要关注晶体结构、分子设计、原子缺陷等材料的物理和化学特性, 涉及材料的合成方法、结构和材料性质的数据^[20-29]。这些数据或分布在结构化的数据库中, 或以非结构化形式散布在描述性文本中, 而材料设计需要将这些多源异构的数据整理成结构化数据, 进而决定材料设计的不同参数^[30]。利用大语言模型的理解和生成能力可以整合这些信息, 并基于此与用户交互。例如, Thik等^[31]利用大语言模型生成基于非结构化文本的材料合成配方(*cooking recipe*)。他们利用ChatGPT-4读取文献中论述材料合成配方的片段, 将其中的实验步骤和实验参数结构化呈现, 描绘了利用大语言模型将非结构化数据转化为结构化数据的概念。但是, 他们的实验范围只围绕单一文段, 并未呈现大体量文段下大语言模型的知识总结和联结能力, 也没有构建能生成总结或建议的大语言模型问答机。后来的研究更有效地利用了大语言模型的生成能力, 扩充了大语言模型在材料设计中的使用场景。例如, 在Deb等^[32]的研究中, 采用ChatGPT进行了八项材料设计相关的任务, 包括生成特定物质的CIF(crystallographic information file), 生成用于DFT计算的输入文件等复杂任务。可是, 由于幻觉消除机制的缺失, 这些任务都需要多轮手工迭代才能完成, 且结果受提示词甚至是用户询问时间影响较大。而Zhao等^[2]的研究引入RAG架构减少幻觉, 搭建RAG架构的大语言模型Battery-GPT, 极大增加了问答结果的精确性和可信度。他们以电池快充为例, 通过文本嵌入技术将文本数据转化为向量表示, 搭建电池快充数据库; 将用户提问向量化, 基于余弦相似度在快充数据库中检索相关信息, 再通过Battery-GPT生成精确的

答案。通过逐层增加输入提示词的精确度, Zhao等^[2]展示了Battery-GPT在应对强专业性垂直领域问题时回答的精确性和可信度, 如图2中展示的负极材料设计问答任务。对于材料设计问题, Battery-GPT能总结数据库中前人的研究结果, 给出带有引用的回答, 并据此给出建议, 更好地发挥大语言模型的生成能力优势。

除了文本、数据库等文字信息之外, 电池材料设计还包含诸多图谱信息, 这些信息也可以被整合到RAG架构中。这些图谱包括扫描电子显微镜(SEM)和透射电子显微镜(TEM)图像, 以及X射线衍射(XRD)和拉曼光谱等谱学数据, 能够提供电池材料在微观层次上的结构信息^[33-35]。结合这些图谱, 神经网络(neural network)技术可以被广泛用来预测材料性质, 如通过微观结构预测材料离子电导率等, 极大促进了微观表征技术的发展与材料构效关系的判断^[36-38]。而相关神经网络架构及其学习结果可以结合到RAG架构的数据库中, 在多模态大语言模型收到用户的显微图像或图谱输入时, 给出相关研究文献, 甚至给出参考性预判或是生成预测代码, 辅助用户进行基于图谱的性质预测。此外, 随着近年多模态大模型的兴起, 整合多模态大模型的RAG架构也能应用于电池材料的图谱分析, 相关内容将在展望部分进行详细论述。

然而, 尽管RAG架构在减少幻觉现象和增强生成内容可信度方面表现, 其在电池材料设计领域的应用仍然存在一些不足。首先, 是最新研究信息获取不及时的问题。在数据源更新不及时或知识库未能及时扩充的情况下, RAG生成的结果可能会基于过时的数据, 给出时效性低的回答。其次, 由于大语言模型主要依赖自然语言处理(natural language processing, NLP)技术, RAG在进行词向量生成时更倾向于专注字面意思, 而无法充分蕴含其中的深层次知识。这意味着在处理复杂的科学概念或专业术语时, 模型可能无法完全捕捉到其深层含义, 进而影响生成结果的准确性和可靠性。最后, RAG架构在向量检索过程中, 可能无法总是匹配到最合适的向量, 这种匹配问题会进一步影响生成内容的质量, 尤其是在电池研究这样需要精确和高度相关的科学数据的情况下。

2.2 电池单元设计和制造

电池单元的研究包括科研端的电池单元设计和

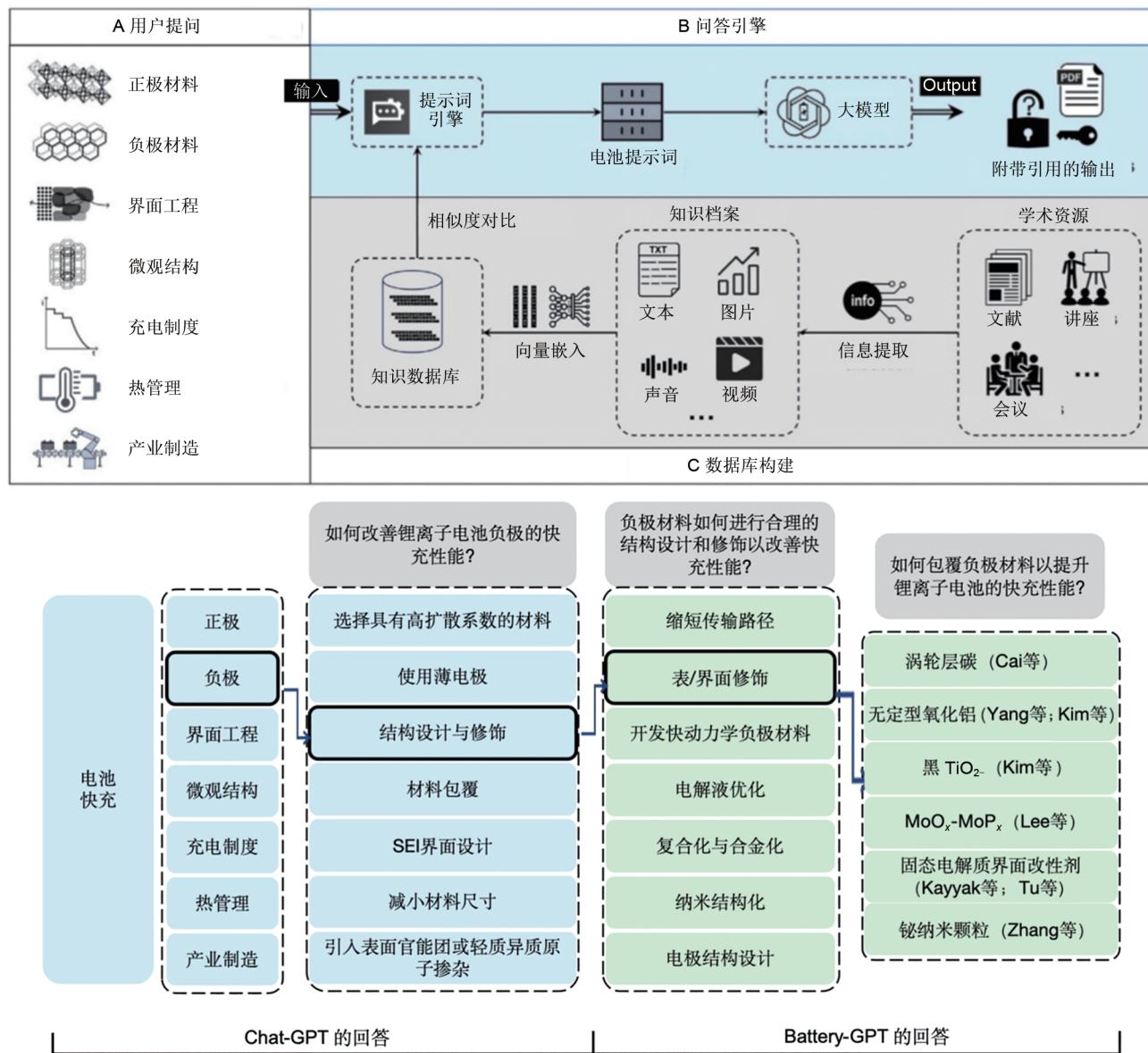


图2 RAG 架构实例：电池快充^[2]
Fig. 2 Example of RAG: fast-charging^[2]

产业端的电芯制造。电芯设计领域的数据特点和电池材料设计相似，均为多源异构，需要整合和分析形成电芯设计需要的参数^[39-41]。RAG 架构能够从这些多种数据源中检索出与电芯设计相关的最新研究成果和技术细节，并与用户交互；或是综合各类数据，提出优化能量密度、功率密度和安全性的参考方案或计算模型。RAG 在电芯设计方面的应用逻辑与材料设计相似，在此不再赘述。

大语言模型 RAG 架构的潜能还能发挥在电芯制造的产业应用中。其一，利用 RAG 架构可以搭建电芯制造认知助手(cognitive assistants)，为产

业人员提供科研数据接口，桥接科研和产业的鸿沟。认知助手是一种人机交互系统，旨在通过人与认知助手的紧密交互让人获取知识，加强人解决复杂问题的能力，而非替代人^[42]。Freire 等^[43]的研究提出了基于大语言模型构建认知助手，利用大语言模型的生成能力进行知识问答、任务委派、工人训练以及生产线参数调节等工作。此外，现今大语言模型的多模态能力能够让产业人员与认知助手的交互扩展到声音识别、操作图像识别，甚至是增强现实(augmented reality)等场景，而非仅仅局限于语言交互^[44]。对这些认知助手进行相关产业数据增

强, 搭建RAG架构的认知助手, 能够大大增加问答的可信度。电芯制造是科研与产业联系极为紧密的领域, 通过构建与产业人员交互的认知助手, 多模态大语言模型的RAG架构可以很好地助推科研成果转化。

其二, 大语言模型RAG架构可以辅助电芯制造产业管控。诚然, 针对特殊生产线的定制神经网络学习模型有着大语言模型难以媲美的预测精度, 是生产线调控、预警等高精度工作的不二之选^[45-47]。但大语言模型的强大文本处理能力与生成能力是定制神经网络学习模型所不具有的, 可以基于对设备手册、维保记录、行业报告等文本数据的分析辅助管控。例如, Zhou等^[48]基于产品质量缺陷相关文本构建因果质量知识图(**causal quality-related knowledge graph**), 搭建因果知识增强大语言模型**CausalKGPT**, 能对文本形式的质量问题描述提供因果分析, 追溯质量问题可能的原因。虽然该方法在追溯诸如材料性质参数、化学元素等更细粒度的微观影响因素上可能存在不足, 但它在利用大语言模型文本解析力的同时加入了生成语言的逻辑性, 为构建RAG架构的电芯制造产业管控模型提供了借鉴。未来可以增加知识库的广度, 构建更稳健的生产线管控建议RAG模型; 还可以结合现今的多模态大模型, 构建基于图像的知识库, 实现针对质检图片、生产线监控图片等的RAG模型解读和问答。

RAG架构尽管在电芯设计和制造中的应用展现了很大的潜力, 但仍存在一些显著的局限性。首先, RAG架构依赖于高质量和广泛覆盖的知识库。然而与材料研发领域相比, 电芯制造领域的公开信息和语料相对较少, 使专业化数据库的构建变得困难。在材料研发中, 海量的论文和研究数据为RAG架构提供了丰富的知识源, 使其能够生成更为精准和深入的设计建议。相比之下, 在电芯制造领域, 信息内容更多集中于产品相关数据, 而生产线的运行数据涉及产业隐私, 信息来源有限。这种数据稀缺性使得建立专业数据库变得极为困难, 进而限制了RAG架构的发挥, 导致生成的建议可能不够全面或不够精确。

2.3 电动交通和电网的电池管理系统

电池管理系统指对电池单元或电池组进行实时监控和状态估计, 并基于此调控电池的工作条件,

确保电池的安全性和效率的系统^[49-52]。电池管理系统广泛运用于电网和电动交通系统中, 其主要功能包括电池状态参数管理、热管理和安全性保障, 确保电池在安全范围内高效运行^[53-57]。在电网应用中, 电池管理系统还需处理复杂的电力调度任务, 结合调峰、调频等手段提高电网的灵活性和可靠性, 保障电池系统和电网系统的高效连接^[55]。在电动交通中, 电池管理系统则需要确保高能量密度、高功率密度和极端条件下的电池安全性, 注重安全范围内的更高性能^[56]。

电网和电动交通系统都是庞大且复杂的系统, 相应的电池管理系统设计需要综合考虑电池本身和全系统的状态参数, 需要跨领域的知识支撑, 如数学、机械工程、电化学等。大语言模型RAG架构可以作为跨领域的知识专家系统, 为电池管理系统研发提供更多维度的思路。例如, 在Buehler的研究中, 他利用本构知识图搭建了具有更强知识连接能力的RAG系统。问答实验的结果表明, 相较于普通文本嵌入增强的RAG架构, 由本体知识图增强的RAG架构展示了更高的知识点关联性, 在单一领域的问答中体现了对概念的更深理解和更强的发散能力, 在交叉领域问答中显著减少了幻觉, 并提出了更深刻的见解^[58]。在本构知识图之外, 基于规则的知识图、基于随机游走算法的知识图等同样展现了良好的连接能力和推理能力, 能够用来搭建RAG系统的增强数据库^[58]。在电池研发过程中, 大语言模型RAG架构能作为电池管理系统设计知识问答机, 依据知识图谱, 提出更具深度的跨领域见解, 为电池管理系统的设计提供创新思路。

此外, 电池管理系统需要高效分析大量参数, 降低调控延迟, 维护系统稳定运行。例如电池组层面的充电状态(**state of charge, SOC**)、健康状态(**state of health, SOH**)、功能状态(**state of function, SOF**)分析, 系统层面的负载管理、电力调度管理等^[59-63]。凭借其对大体量多模态数据的高效分析能力和推理能力, 大语言模型RAG架构能综合系统数据做出预测, 为系统参数调节提供参考^[64]。借助它的知识联结和推理能力, 大语言模型RAG架构可以识别出潜在的故障模式和异常情况, 实现早期预警和故障诊断。而通过检索组件对数据库的快速查找, 大语言模型RAG架构还能实现实时数据分析, 辅助电池管理系统的实时监控和动态优化, 提

高系统的响应速度和决策准确性。

尽管 RAG 架构在电动交通和电网的电池管理系统中显示出许多潜力，其实际应用仍存在一些重要的局限性。首先，电池管理系统相关论文中涉及的复杂数学公式和计算模型往往难以通过自然语言处理的方式进行准确处理和理解，导致 RAG 架构在解读和应用这些公式时可能出现偏差。同时也面临着材料研发具有类似的挑战，由于电池管理系统技术发展迅速，研究进展日新月异，在数据库追踪和整合最新研究成果时可能存在滞后，影响生成内容的时效性和准确性。RAG 架构在向量检索过程中，可能无法总是匹配到最合适的向量，这时就会不可避免地产生幻觉使生成的回答不够完整和准确。

2.4 RAG 架构在电池技术的应用的异同

RAG 架构在电池技术的应用可以从微观层面的材料设计、器件层面的电池单元设计与制造，以及系统层面的电池管理系统三个维度来分析。在这三个领域中，RAG 架构均发挥了强大的跨学科整合能力，能够从化学、物理学、机械工程、电化学等不同领域中提取相关知识，为电池技术提供支持。RAG 架构在电池技术的应用中展现出显著的不同点，这些差异主要体现在信息来源的丰富性、数据处理的复杂性以及跨尺度信息的整合上。首先，在材料研发领域，作为电池技术的前端阶段，该领域论文和专利信息相对丰富，RAG 架构能够利用大量的文字性描述和构效关系的总结来提炼有价值的信息。然而，在设计制造和电池管理系统中，公开信息和语料的稀缺使得构建专业化的数据库尤为困难。其次，文献中的数据类型也有所不同。材料设计中的数据多为描述性文字，而这些文字可以被 RAG 架构有效提炼和整合；然而，电池管理系统相关应用中大量涉及的数学公式和计算模型则不容易通过自然语言处理的方式来处理和理解，导致 RAG 架构在这些场景中的应用受到限制。最后，材料设计的信息主要集中在微观层面，而电池单元的设计与制造以及电池管理系统则需要跨越宏观和微观尺度，整合多种层次的信息。

尽管如此，RAG 架构在推动电池技术创新中的潜力依然不容忽视。RAG 架构在电池技术的三个维度，从微观材料设计到器件制造再到系统管理，均展现了显著的应用潜力。然而，这些应用的有效性在很大程度上依赖于数据的质量和知识库的完善程

度。未来研究应进一步优化 RAG 在动态环境中的应用，提升其在不同维度上的实用性和可靠性。

3 展 望

3.1 多模态 RAG 在电池领域的应用

多模态大模型 (multimodal large language models, MM-LLMs) 近年来在人工智能领域取得了显著进展^[64]。这些模型能够处理和生成涉及多种模态的信息，如文本、图像、视频和音频，极大扩展了传统语言模型的应用范围。通过不同模态信息的高度融合，MM-LLMs 可以实现更加全面的理解和更丰富的内容生成，使得它们在许多复杂任务中表现出色。在此基础上，RAG 架构能进一步提升多模态大模型的能力。如图 4 所示，RAG 架构结合了检索和生成的优势，通过检索相关信息来增强生成的准确性和可靠性，尤其适用于处理涉及多个模态的复杂数据集。多模态 RAG 在各个领域都有广泛的应用潜力，电池材料学就是其中一个典型的应用领域。

3.1.1 图像信息的多模态 RAG 应用

电池领域涉及大量的图像数据，如扫描电子显微镜(SEM)图像、透射电子显微镜(TEM)图像以及电池结构的显微图像等。这些图像用于分析电池材料的微观结构、形貌变化以及材料在不同工作条件下的行为。通过多模态 RAG 技术，系统可以从数据库中检索出与当前图像相似的实验结果或案例，并生成解释文本，帮助研究人员快速理解材料特性。此外，多模态 RAG 技术还可以将这些图像与其他模态的数据(如实验报告或材料描述)进行结合，生成更加详尽的分析报告。这种方法不仅提高了材料研究的效率，还为优化电池设计提供了宝贵的参考信息。

3.1.2 光谱信息的多模态 RAG 应用

光谱信息在电池材料的表征中扮演着重要角色，常见的包括 X 射线光电子能谱(XPS)、拉曼光谱、核磁共振波谱(NMR)等。这些光谱数据能够揭示材料的化学组成、电子结构和表面特性。多模态 RAG 技术可以通过分析这些光谱数据，与数据库中存储的相关文献或实验数据比对，生成对光谱特征的解释。此外，RAG 还能够将光谱数据与其他模态信息(如图像或文本)结合，帮助研究人员全面理解材料特性。例如，将光谱数据与扫描电子显微

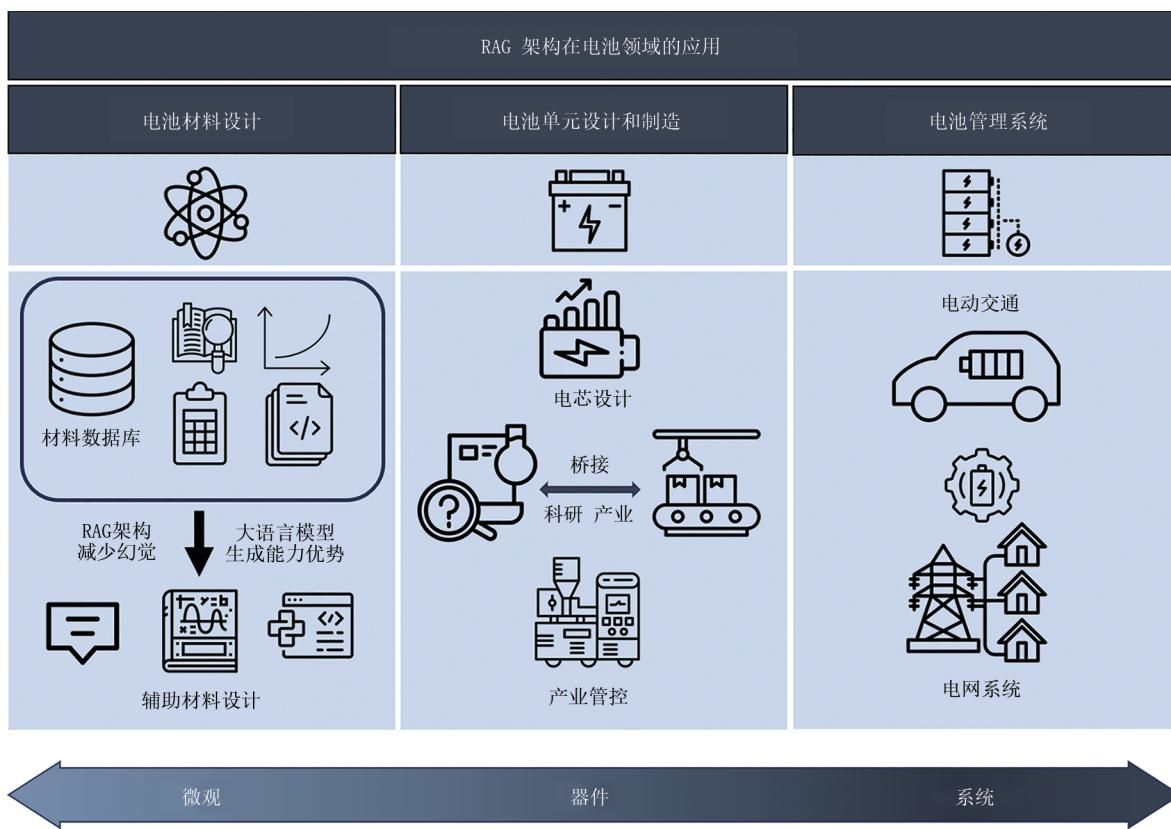


图3 RAG架构在电池领域的应用
Fig. 3 The application of RAG in the battery field

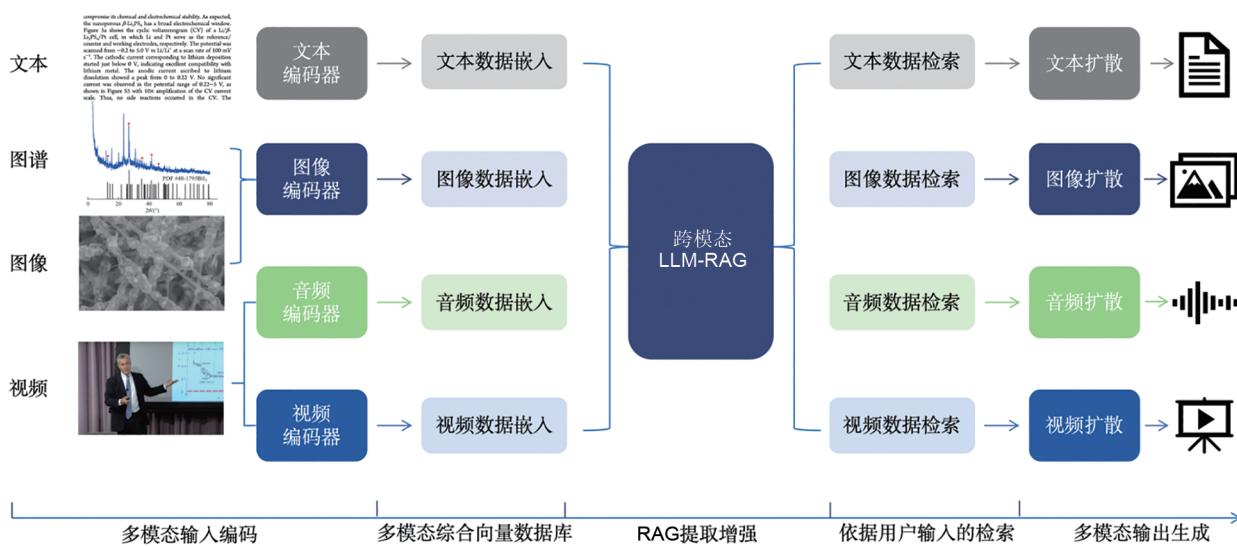


图4 RAG架构多模态应用展望
Fig. 4 The outlook for multimodal applications using the RAG architecture

镜(SEM)图像结合,可以揭示材料在化学成分和微观结构之间的关系,进而优化材料性能。

3.1.3 跨模态信息整合: 多模态RAG综合应用

多模态RAG技术的优势在于其能够整合不同类

型的信息,提供更加全面和精确的分析。在电池领域应用多模态RAG技术的关键在于如何有效整合跨模态信息。这不仅仅是将图片和文字简单对应,还需要理解图像与图像之间、图谱与文字之间的深层

次关联。例如，通过多模态 RAG 技术，可以将电化学测试曲线与材料显微结构图像关联起来，分析不同条件下材料结构变化对电池性能的影响，从而提供更加精准的电池设计建议。多模态 RAG 技术在电池研究中的应用，不仅能加速实验数据的分析和理解，还能帮助研究人员发现新的材料设计思路，从而推动电池技术的进步。这一技术的应用将为电池材料设计、性能优化和故障分析等方面带来革命性的变化，助力电池领域迈向新的高度。

3.2 RAG 技术在电池研究中的其他应用展望

3.2.1 回答的可解释性

不论在科研领域还是产业领域，大语言模型 RAG 架构的回答都需要极强的可解释性，以确保回答的可信度。例如，回答应标注引述出处，说明检索增强的信息如何影响输出，以及给出可信度和不确定性的计算^[65]。当下已经有诸多针对大语言模型 RAG 架构性能的评估方式，如 RGB、RECALL 等评估基准，以及 EM、MRR 等评分机制^[17,66-67]。未来的研究应致力于构造更加透明可视化的 RAG 架构，最大化回答的可解释性。

3.2.2 隐私数据保护

在电池技术研究中，数据的隐私和安全性至关重要。首先，对于封装完毕的大语言模型 RAG 架构，攻击者可能通过成员资格推断攻击 (membership inference attack) 的方式批量获取增强数据库中的数据^[68]。另外，用户的输入数据可能涉及专有材料配方、商业秘密和敏感实验结果，如果泄露将带来巨大的风险。如何制定严格的数据加密和访问控制措施应当是未来 RAG 架构研究要解决的问题。

3.2.3 系统的智能运维

由于电池系统运行和维护关乎系统安全性，当下仍需大量人力投入，大语言模型只能充当辅助作用。但随着 RAG 架构的透明化、可解释化，大语言模型 RAG 架构可以承担更多的智能决策任务，作为决策中心调控外部资源，进行参数优化、能量分配优化、异常情况处理等任务，减少人力投入，增强系统效率。

4 结 论

通过大语言模型 RAG 架构与大语言模型微调的对比，展示了 RAG 架构经济、简洁且高效的特

点，证明其更适用于研究基数大、增长快的电池领域。然后，通过对近期文献的综述和发散，展现了 RAG 结构大语言模型在电池材料设计、电池单元设计和制造、电动交通与电网的电池管理系统三个层面的应用。最后，展望了多模态 RAG 技术，探讨了大语言模型 RAG 架构在知识问答、科研建议与产业辅助之上的潜在应用，展现了大语言模型 RAG 架构的广阔前景。大语言模型 RAG 架构在电池领域的应用尚刚刚起步。未来，通过 RAG 的流程优化以及与其他架构的融合，大语言模型 RAG 架构有望成为电池研究人员的重要助手，显著加快电池研究。

参 考 文 献

- [1] VASWANI A, SHAZEER N, PARMER N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.
- [2] ZHAO S, CHEN S H, ZHOU J Y, et al. Potential to transform words to Watts with large language models in battery research[J]. Cell Reports Physical Science, 2024, 5(3): 101844. DOI: 10.1016/j.xrpp.2024.101844.
- [3] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. 2021: 2103.00020.[2024-06-30]. <https://arxiv.org/abs/2103.00020v1>
- [4] OpenAI. Hello GPT-4o[EB/OL]. [2024-06-30]. <https://openai.com/index/hello-gpt-4o/>
- [5] OpenAI. SORA[EB/OL]. [2024-06-30]. <https://openai.com/sora>
- [6] BUEHLER M J. Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design[J]. ACS Engineering Au, 2024, 4(2): 241-277. DOI: 10.1021/acsengineeringau.3c00058.
- [7] 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. DOI: 10.1038/s42256-024-00832-8.
- [8] BOIKO D A, MACKNIGHT R, KLINE B, et al. Autonomous chemical research with large language models[J]. Nature, 2023, 624(7992): 570-578. DOI: 10.1038/s41586-023-06792-0.
- [9] KANG Y, KIM J. ChatMOF: An artificial intelligence system for predicting and generating metal-organic frameworks using large language models[J]. Nature Communications, 2024, 15(1): 4705. DOI: 10.1038/s41467-024-48998-4.
- [10] ZHENG Z L, ZHANG O F, BORGES C, et al. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis[J]. Journal of the American Chemical Society, 2023, 145(32): 18048-18062. DOI: 10.1021/jacs.3c05819.
- [11] JI Z W, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[EB/OL]. 2022: 2202.03629.[2024-06-30]. <https://arxiv.org/abs/2202.03629v7>
- [12] 吴思远, 王雪龙, 肖睿娟, 等. 基于大型语言模型的工具对电池研究

- 的机遇与挑战[J]. 储能科学与技术, 2023, 12(3): 992-997. DOI: 10.19799/j.cnki.2095-4239.2023.0071.
- [15] WU S Y, WANG X L, XIAO R J, et al. Problem and perspective for battery researcher based on large language model[J]. Energy Storage Science and Technology, 2023, 12(3): 992-997. DOI: 10.19799/j.cnki.2095-4239.2023.0071.
- [13] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(13): 3521-3526. DOI: 10.1073/pnas.1611835114.
- [14] TINN R, CHENG H, GU Y, et al. Fine-tuning large neural language models for biomedical natural language processing[J]. Patterns, 2023, 4(4): 100729. DOI: 10.1016/j.patter.2023.100729.
- [15] MOSBACH M, ANDRIUSHCHENKO M, KLAKOW D. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines[EB/OL]. 2020: 2006.04884.[2024-06-30]. <https://arxiv.org/abs/2006.04884v3>
- [16] KIM Y, OH J, KIM S, et al. How to fine-tune models with few samples: Update, data augmentation, and test-time augmentation [EB/OL]. 2022: 2205.07874. [2024-06-30]. <https://arxiv.org/abs/2205.07874v3>
- [17] CHEN J W, LIN H Y, HAN X P, et al. Benchmarking large language models in retrieval-augmented generation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16): 17754-17762. DOI: 10.1609/aaai.v38i16.29728.
- [18] GAO Y F, XIONG Y, GAO X Y, et al. Retrieval-augmented generation for large language models: A survey[EB/OL]. 2023: 2312.10997.[2024-06-30]. <https://arxiv.org/abs/2312.10997v5>
- [19] EIBICH M, NAGPAL S, FRED-OJALA A. ARAGOG: Advanced RAG output grading[EB/OL]. 2024: 2404.01037. [2024-06-30]. <https://arxiv.org/abs/2404.01037v1>
- [20] PAN S R, LUO L H, WANG Y F, et al. Unifying large language models and knowledge graphs: A roadmap[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3580-3599. DOI: 10.1109/TKDE.2024.3352100.
- [21] ZHANG J, SUTING W, ZHAOXIANG W, et al. Solid electrolyte interphase (SEI) on graphite anode correlated with thermal runaway of lithium-ion batteries[J]. Energy Storage Science and Technology, 2023, 12(7): 2105-2118.
- [22] WANG A P, KADAM S, LI H, et al. Review on modeling of the anode solid electrolyte interphase (SEI) for lithium-ion batteries [J]. NPJ Computational Materials, 2018, 4: 15. DOI: 10.1038/s41524-018-0064-0.
- [23] WANG X X, NI L, XIE Q X, et al. Highly electrocatalytic active amorphous Al₂O₃ in porous carbon assembled on carbon cloth as an independent multifunctional interlayer for advanced lithium-sulfur batteries[J]. Applied Surface Science, 2023, 618: 156689. DOI: 10.1016/j.apsusc.2023.156689.
- [24] LAI G M, JIAO J Y, FANG C, et al. The mechanism of Li deposition on the Cu substrates in the anode-free Li metal batteries[J]. Small, 2023, 19(3): e2205416. DOI: 10.1002/smll.202205416.
- [25] WANG F, SUN Y, CHENG J. Switching of redox levels leads to high reductive stability in water-in-salt electrolytes[J]. Journal of the American Chemical Society, 2023, 145(7): 4056-4064. DOI: 10.1021/jacs.2c11793.
- [26] HAN X, GU L H, SUN Z F, et al. Manipulating charge-transfer kinetics and a flow-domain LiF-rich interphase to enable high-performance microsized silicon-silver-carbon composite anodes for solid-state batteries[J]. Energy & Environmental Science, 2023, 16(11): 5395-5408. DOI: 10.1039/D3EE01696J.
- [27] CHEN G J, YU L W, GAN Y H, et al. Reinforcing the stability of cobalt-free lithium-rich layered oxides via Li-poor Ni-rich surface transformation[J]. Journal of Materials Chemistry A, 2024. DOI: 10.1039/d4ta01403k.
- [28] LI J, ZHOU M S, WU H H, et al. Machine learning-assisted property prediction of solid-state electrolyte[J]. Advanced Energy Materials, 2024, 14(20): 2304480. DOI: 10.1002/aenm.202304480.
- [29] MERCHANT A, BATZNER S, SCHOENHOLZ S S, et al. Scaling deep learning for materials discovery[J]. Nature, 2023, 624(7990): 80-85. DOI: 10.1038/s41586-023-06735-9.
- [30] XU J, XIAO R J, LI H. ESM cloud toolkit: A copilot for energy storage material research[J]. Chinese Physics Letters, 2024, 41(5): 054701. DOI: 10.1088/0256-307x/41/5/054701.
- [31] THIK J, WANG S W, WANG C H, et al. Realizing the cooking recipe of materials synthesis through large language models[J]. Journal of Materials Chemistry A, 2023, 11(47): 25849-25853. DOI: 10.1039/D3TA05457H.
- [32] DEB J, SAIKIA L, DIHINGIA K D, et al. ChatGPT in the material design: Selected case studies to assess the potential of ChatGPT [J]. Journal of Chemical Information and Modeling, 2024, 64(3): 799-811. DOI: 10.1021/acs.jcim.3c01702.
- [33] LIU Y, LI C, LI C X, et al. Porous, robust, thermally stable, and flame retardant nanocellulose/polyimide separators for safe lithium-ion batteries[J]. Journal of Materials Chemistry A, 2023, 11(43): 23360-23369. DOI: 10.1039/d3ta05148j.
- [34] 凌仕刚, 许洁茹, 李泓. 锂电池研究中的EIS实验测量和分析方法 [J]. 储能科学与技术, 2018, 7(4): 732-749. DOI: 10.12028/j.issn.2095-4239.2018.0092.
- LING S G, XU J R, LI H. Experimental measurement and analysis methods of electrochemical impedance spectroscopy for lithium batteries[J]. Energy Storage Science and Technology, 2018, 7(4): 732-749. DOI: 10.12028/j.issn.2095-4239.2018.0092.
- [35] LIU X H, HUANG J Y. In situ TEM electrochemistry of anode materials in lithium ion batteries[J]. Energy & Environmental Science, 2011, 4(10): 3844-3860. DOI: 10.1039/C1EE01918J.
- [36] KITAHARA A R, HOLM E A. Microstructure cluster analysis with transfer learning and unsupervised learning[J]. Integrating Materials and Manufacturing Innovation, 2018, 7(3): 148-156. DOI: 10.1007/s40192-018-0116-9.
- [37] ZHANG Y, LIN X Y, ZHAI W B, et al. Machine learning on microstructure-property relationship of lithium-ion conducting oxide solid electrolytes[J]. Nano Letters, 2024, 24(17): 5292-

5300. DOI: 10.1021/acs.nanolett.4c00902.
- [38] KUSCHE C, RECLIK T, FREUND M, et al. Large-area, high-resolution characterisation and classification of damage mechanisms in dual-phase steel using deep learning[J]. PLoS One, 2019, 14(5): e0216493. DOI: 10.1371/journal.pone.0216493.
- [39] WANG Z H, ZHOU X Z, ZHANG W G, et al. Parameter sensitivity analysis and parameter identifiability analysis of electrochemical model under wide discharge rate[J]. Journal of Energy Storage, 2023, 68: 107788. DOI: 10.1016/j.est.2023.107788.
- [40] 黄晟贤,徐会升,王起鹏,等.冲击荷载下圆柱型动力锂离子电池的响应特性研究[J/OL].储能科学与技术.[2024-05-05]. <https://doi.org/10.19799/j.cnki.2095-4239.2024.0274>.
- HUANG S, XU H, WANG Q, et al. Study on the Response characteristics of cylindrical power lithium-ion batteries under impact load[J/OL]. Energy Storage Science and Technology. [2024-05-05]. <https://doi.org/10.19799/j.cnki.2095-4239.2024.0274>.
- [41] WU X K, SONG K F, ZHANG X Y, et al. Safety issues in lithium ion batteries: Materials and cell design[J]. Frontiers in Energy Research, 2019, 7: 65. DOI: 10.3389/fenrg.2019.00065.
- [42] OAKLEY J. Intelligent cognitive assistants [EB/OL].[2024-06-30]. <https://www.src.org/program/ica/>
- [43] KERNAN FREIRE S, FOOSHERIAN M, WANG C F, et al. Harnessing large language models for cognitive assistants in factories[C]// Proceedings of the 5th International Conference on Conversational User Interfaces. ACM, 2023: 1-6. DOI: 10.1145/3571884.3604313.
- [44] CHEN H S, HOU L, WU S Z, et al. Augmented reality, deep learning and vision-language query system for construction worker safety[J]. Automation in Construction, 2024, 157: 105158. DOI: 10.1016/j.autcon.2023.105158.
- [45] PERES R S, JIA X D, LEE J, et al. Industrial artificial intelligence in industry 4.0 - Systematic review, challenges and outlook[J]. IEEE Access, 2874, 8: 220121-220139. DOI: 10.1109/ACCESS.2020.3042874.
- [46] GUO N L, CHEN S H, TAO J, et al. Semi-supervised learning for explainable few-shot battery lifetime prediction[J]. Joule, 2024, 8(6): 1820-1836. DOI: 10.1016/j.joule.2024.02.020.
- [47] TRIVEDI C, BHATTACHARYA P, PRASAD V K, et al. Explainable AI for industry 5.0: Vision, architecture, and potential directions [J]. IEEE Open Journal of Industry Applications, 2024, 5: 177-208. DOI: 10.1109/OJIA.2024.3399057.
- [48] ZHOU B, LI X Y, LIU T Y, et al. CausalKGPT: Industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing [J]. Advanced Engineering Informatics, 2024, 59: 102333. DOI: 10.1016/j.aei.2023.102333.
- [49] LIN T Z, CHEN S H, HARRIS S J, et al. Investigating explainable transfer learning for battery lifetime prediction under state transitions[J]. eScience, 2024: 100280. DOI: 10.1016/j.esci.2024.100280.
- [50] 朱伟杰, 史尤杰, 雷博. 锂离子电池储能系统BMS的功能安全分析与设计[J]. 储能科学与技术, 2020, 9(1): 271-278. DOI: 10.19799/j.cnki.2095-4239.2019.0177.
- ZHU W J, SHI Y J, LEI B. Functional safety analysis and design of BMS for lithium-ion battery energy storage system[J]. Energy Storage Science and Technology, 2020, 9(1): 271-278. DOI: 10.19799/j.cnki.2095-4239.2019.0177.
- [51] SHEN M, GAO Q. A review on battery management system from the modeling efforts to its multiapplication and integration[J]. International Journal of Energy Research, 2019, 43(10): 5042-5075.
- [52] RAHIMI-EICHI H, OJHA U, BARONTI F, et al. Battery management system: An overview of its application in the smart grid and electric vehicles[J]. IEEE Industrial Electronics Magazine, 2013, 7(2): 4-16. DOI: 10.1109/MIE.2013.2250351.
- [53] LIU Y, CHEN S H, LI P Y, et al. Status, challenges, and promises of data-driven battery lifetime prediction under cyber-physical system context[J]. IET Cyber-Physical Systems: Theory & Applications, 2024. DOI: 10.1049/cps2.12086.
- [54] LU T, ZHAI X A, CHEN S H, et al. Robust battery lifetime prediction with noisy measurements via total-least-squares regression[J]. Integration, 2024, 96: 102136. DOI: 10.1016/j.vlsi.2023.102136.
- [55] LAUDER M T, SUTHAR B, NORTHROP P W C, et al. Battery energy storage system (BESS) and battery management system (BMS) for grid-scale applications[J]. Proceedings of the IEEE, 2014, 102(6): 1014-1030. DOI: 10.1109/JPROC.2014.2317451.
- [56] MISHRA S, SWAIN S C, SAMANTARAY R K. A Review on battery management system and its application in electric vehicle [C]// 2021 International Conference on Advances in Computing and Communications (ICACC). IEEE, 2021: 1-6. DOI: 10.1109/ICACC-202152719.2021.9708114.
- [57] WANG J, FENG X N, YU Y Z, et al. Rapid temperature-responsive thermal regulator for safety management of battery modules[J]. Nature Energy, 2024, 9: 939-946. DOI: 10.1038/s41560-024-01535-5.
- [58] CHEN X J, JIA S B, XIANG Y. A review: Knowledge reasoning over knowledge graph[J]. Expert Systems with Applications, 2020, 141: 112948. DOI: 10.1016/j.eswa.2019.112948.
- [59] XIONG R, LI L L, TIAN J P. Towards a smarter battery management system: A critical review on battery state of health monitoring methods[J]. Journal of Power Sources, 2018, 405: 18-29. DOI: 10.1016/j.jpowsour.2018.10.019.
- [60] ZHAO X Z, SUN B X, ZHANG W G, et al. Error theory study on EKF-based SOC and effective error estimation strategy for Li-ion batteries[J]. Applied Energy, 2024, 353: 121992. DOI: 10.1016/j.apenergy.2023.121992.
- [61] XIONG R, SUN Y, WANG C X, et al. A data-driven method for extracting aging features to accurately predict the battery health [J]. Energy Storage Materials, 2023, 57: 460-470. DOI: 10.1016/j.ensm.2023.02.034.
- [62] WANG W T, YANG K Y, ZHANG L S, et al. An end-cloud collaboration approach for online state-of-health estimation of lithium-ion batteries based on multi-feature and transformer[J].

- Journal of Power Sources, 2024, 608: 234669. DOI: 10.1016/j.jpowsour.2024.234669.
- [63] 吴思远, 李泓. 发展基于“语义检测”的低参数量、多模态预训练电池通用人工智能模型[J]. 储能科学与技术, 2024, 13(4): 1216-1224. DOI: 10.19799/j.cnki.2095-4239.2024.0092.
WU S Y, LI H. Developing a general pretrained multimodal battery model with small parameters based on semantic detection [J]. Energy Storage Science and Technology, 2024, 13(4): 1216-1224. DOI: 10.19799/j.cnki.2095-4239.2024.0092.
- [64] WU S Q, FEI H, QU L G, et al. NExT-GPT: Any-to-any multimodal LLM[EB/OL]. 2023: 2309.05519. <https://arxiv.org/abs/2309.05519v3>
- [65] 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. DOI: 10.1039/d3dd00113j.
- [66] BRUCKHAUS T. RAG does not work for enterprises[EB/OL]. 2024: 2406.04369. [2024-06-30]. <https://arxiv.org/abs/2406.04369v1>
- [67] LIU Y, HUANG L Z, LI S C, et al. RECALL: A benchmark for LLMs robustness against external counterfactual knowledge[EB/OL]. 2023: 2311.08147. [2024-06-30]. <https://arxiv.org/abs/2311.08147v1>
- [68] ANDERSON M, AMIT G, GOLDSTEEN A. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation[EB/OL]. 2024: 2405.20446.[2024-06-30]. <https://arxiv.org/abs/2405.20446v2>