

基于人机协作的多智能体科学假设生成

陈子阳¹ 赵翔¹ 赵润豪¹ 倪子淇² 叶益聪²

¹(国防科技大学大数据与决策实验室 长沙 410003)

²(国防科技大学空天科学学院 长沙 410003)

(chenziyangnudt@nudt.edu.cn)

Multi-Agent Scientific Hypothesis Generation Based on Human-Machine Collaboration

Chen Ziyang¹, Zhao Xiang¹, Zhao Runhao¹, Ni Ziqi², and Ye Yicong²

¹(Laboratory for Big Data and Decision, National University of Defense Technology, Changsha 410003)

²(College of Aerospace Science and Engineering, National University of Defense Technology, Changsha 410003)

Abstract With the explosive growth of scientific literature and the continuous deepening of research fields, researchers face significant information processing challenges when attempting to formulate novel scientific hypotheses. Although large language models (LLMs) possess considerable potential for data processing and knowledge integration, they remain limited in their ability to generate original and insightful scientific hypotheses. Existing research predominantly emphasizes utilizing LLMs to expedite and refine established theories and technologies, often overlooking the initial stages of scientific inquiry where novel hypotheses are proposed and new theories are developed—a stage vital to scientific advancement. This study, grounded in the principles of divergent and convergent thinking from the theory of structured intelligence, proposes an innovative human-in-the-loop multi-agent framework (HILMA) for the reliable generation of scientific hypotheses. HILMA framework incorporates a real-time, systematic knowledge retrieval enhancement mechanism, dynamically integrating the latest research advancements to construct citation network subgraphs, providing LLMs with comprehensive and up-to-date scientific knowledge surveys. Additionally, the framework enhances hypothesis generation through a multi-agent argumentation approach that simulates the scientific peer review process, while also leveraging the intuition and expertise of human experts to further refine and diversify the generated hypotheses. A series of human-machine evaluations has shown that this method demonstrates significant advantages over existing baselines in generating high-quality scientific hypotheses and holds promise as a key facilitator for driving technological innovation.

Key words large language models (LLMs); scientific hypothesis generation; multi-agent; human-machine collaboration; theory of structural intelligence

摘要 随着科学文献数量的快速增长和研究领域的不断深化,科研人员在提出创新性科学假设时面临巨大的信息处理挑战。尽管大语言模型 (large language models, LLMs) 在数据处理和知识整合方面展现出巨大潜力,但它们在生成具有创新性和深度的科学假设方面仍存在许多不足。目前的研究主要集中在如何利用 LLMs 加速已有理论和技术的推进和完善,而忽视了科学研究从无到有的初始阶段,这一阶段

收稿日期: 2024-06-20; 修回日期: 2024-12-31

基金项目: 国家自然科学基金项目 (U23A20296, 62272469); 湖南省科技创新计划项目 (2023RC1007)

This work was supported by the National Natural Science Foundation of China (U23A20296, 62272469) and the Science and Technology Innovation Program of Hunan Province (2023RC1007).

通信作者: 赵翔 (xiangzhao@nudt.edu.cn)

涉及新假设的提出和新理论的构建,是科学进步的关键.基于结构智力理论中的发散思维和收敛思维,提出了一种创新的人机协作多智能体框架(human-in-the-loop multi-agent framework, HILMA),以实现可靠的初始科学假设生成.该框架结合实时系统化的知识检索增强机制,通过动态整合最新科研进展,构建引文网络子图,为LLMs提供前沿和完备的科研知识综述.同时,通过多智能体辩论方法模拟科学同行评审过程,并且结合人类专家的直觉和专业知識,进一步优化和精炼生成的假设,增强科学假设的多样性和论证深度.一系列人机评估表明,与现有基线相比,HILMA在生成高质量科学假设方面展现出显著优势,有望成为推动科技创新的关键工具.

关键词 大语言模型;科学假设生成;多智能体;人机协作;结构智力理论

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440552

CSTR: 32373.14.issn1000-1239.202440552

科学研究在推动社会发展和技术进步中起着至关重要的作用.科学假设的生成是科研过程中的一个基础环节,它为实验设计和理论探索提供了方向.然而,随着科学文献数量的急剧增加和研究领域的不断深化,研究人员在提出科学假设时面临着巨大的挑战.一方面,研究人员需要从大量的文献中提取有价值的信息,这是一个耗时且劳力密集的过程^[1];另一方面,创新的科学假设往往需要跨学科的知识 and 深入的洞察力^[2],而这些难以通过传统方法迅速实现.

在这种背景下,人工智能(artificial intelligence, AI),尤其是大语言模型(large language models, LLMs),因其强大的数据处理能力和信息整合能力,成为科学研究中的一种重要工具. LLMs通过在大量的文本数据上进行训练,能够理解和生成复杂的文本信息,表现出在不同任务上处理和生成知识的能力^[3].例如, GPT-4等模型不仅在文本生成上表现出色,还在诸如编程^[4]、法律^[5]和生物医药^[6]等专业领域展示了其应用潜力.在科学研究领域, LLMs已被用于各种任务,包括药物发现、医疗诊断和材料性能预测^[7],为研究人员提供了强大的支持,特别是在快速浏览和整理大量科学文献时显示出其独特优势.例如, LLMs能够自动识别文献中的关键概念和关联,从而加速文献综述的编写过程.

在当前的科学研究与工程实践中,大部分研究关注于如何使用LLMs将已有的理论或技术从初步阶段快速推进到成熟阶段^[8],以此提升科学研究的效率,即“1~100”的过程.这包括一些过程繁杂但对创新要求较低的任务,如自动化文献总结、科学文章撰写以及代码实现等.尽管这一阶段的研究对技术发展和应用普及至关重要,但它依赖于已有的框架和理论基础.然而,科学探索从无到有的初始阶段,即“0~1”的过程,在许多情况下被忽视.这一阶段涉及

到新假设的提出、初步概念的形成以及新理论的构建.正是这些活动定义了科学研究的前沿性和创新性,提出好的科学假设是科学进步的启动器和基石^[9].

科学假设生成指的是在科学研究的前期阶段,系统地整合现有知识并探索新的理论路径,以提出具有创新性的主张与思路的过程.尽管LLMs在科学研究中提供了强大的支持,它们在科学假设生成的应用上还是存在明显的不足和局限.首先,LLMs依赖于大量预训练数据,这些数据无法及时反映科学的最新进展,而学术界的知识更新非常迅速,尤其是在如计算机、生物、材料等快速发展的领域,导致模型在生成假设时依赖于过时或不完整的信息^[10],无法站在科学前沿的“巨人肩膀”上进行有效创新.其次,虽然LLMs能生成看似复杂的科学内容,但这些内容在科学的准确性和深度上显得不足,缺乏真正的创新性和深度,倾向于重复训练数据中的模式,而不是提出创新的科学理论或方法^[11].这些挑战主要源于LLMs的固有局限性和科学研究的复杂需求.这也表明,尽管LLMs为科学假设的生成提供了新的可能,但要实现其在科学研究中的有效应用,以及生成具备研究深度与创新性的假设,仍需深入探索和改进.

科学假设决定着整个研究的质量和方向,高质量的科学假设应当能够在已有研究的基础上进一步深化和创新.在这个步骤中,如何让LLMs产生创新,则成为最关键的问题. Guilford^[12]在结构智力理论中提出,创新产生于思维的发散和收敛过程之中.发散思维是指从一个问题或主题出发,能够生成多种可能的答案或获得解决方案的能力.这种思维模式强调的是想象力、创造力和多样性的生成.收敛思维则是指针对特定问题找到具体、正确答案的能力.这种思维模式更多关注逻辑推理、精确度和效率.科学研究的创新思路就来源于对研究问题的发散思考和收

敛具化. 发散过程需要在一定信息基础上, 对问题和解决方案进行扩展与探索, 这一过程通常通过交流、讨论和辩论来激发; 而收敛过程则需要深入理解问题并对特定方向进行深入探索, 而这依赖于人类专家具备的高阶洞见能力. 这些正是 LLMs 在创新过程中面临的挑战, 导致发散的角度不够广泛、收敛的方向不够精准和可靠.

为了解决 LLMs 在生成科学假设中存在的问题, 本文基于结构智力理论, 提出了基于人机协作的多智能体框架(human-in-the-loop multi-agent framework, HILMA). 如图 1 所示, HILMA 框架包括基于引文网络的文献系统化检索增强、知识增强型 LLMs 研究想法生成、人机协作的多智能体辩论迭代 3 个模块. 首先, 针对 LLMs 在知识动态更新上的不足这一问题, 本文通过检索增强的方法为 LLMs 注入最新的科研相关知识, 在无需对模型进行昂贵的重新训练或复杂的微调的情况下, 通过低成本的上下文提示来弥补 LLMs 在动态知识更新上的不足, 从而使其在生成科学假设时具备最新且完备的科学研究进展. 其次, 通过引入基于多智能体辩论的假设迭代增强方法, 以辩论方式模拟人类的发散思维, 激发 LLMs 的能力, 使用发散思维来产生尽可能多的初步假设. 在该框架中, 不同的智能体角色负责从多个角度审视和辩论初始假设, 通过丰富的讨论和批评性反馈, 增加假设的多样性和深度. 这种方法模拟了科学界的同行评审过程, 提升了假设的科学性和实用性. 最后, 为了进一步提升 LLMs 的专业性和针对性, HILMA 框架结合了人类专家的高阶洞察力和模型的数据处理

能力. 通过在辩论过程中实时引入人类学者的直觉和专业知识, 以指导模型的假设生成过程, 并通过收敛思维从中选择最有潜力的方案进行进一步深化和精炼. 这种协作旨在最大化利用了 LLMs 潜力和人类的洞察力, 以生成更符合实际研究需求的科学假设. 这一综合策略的应用, 能够显著提升 LLMs 在科学研究领域中的应用效果, 使得生成的假设在质量和创新性方面满足科学研究的要求.

综上所述, 本文的主要贡献包括 4 个方面:

1) 系统地分析了 LLMs 在处理科学创造性问题时所面临的局限性, 揭示了其在发散思维和收敛思维等方面的不足;

2) 提出了科学文献的结构化组织查询方法, 通过对文献引文网络的自适应组织和构建, 有效地整合了最新的科学研究成果, 确保为 LLMs 提供全面且前沿的知识;

3) 受结构智力理论的启发, 提出了人机协作的多智能体辩论框架, 可在多智能体和人类的协作下, 模仿人类的发散思维和收敛思维过程, 有效地产生和精炼科学假设, 提升其质量与创新性;

4) 基于人类和模型的系列评估, 实验结果证明了所提 HILMA 框架的优越性, 相较于现有 LLMs 驱动的基线模型, HILMA 生成的科学假设在创新性、实用性、可行性等方面均有显著提升.

1 相关工作

本节聚焦于 3 个与本文研究密切相关的研究方

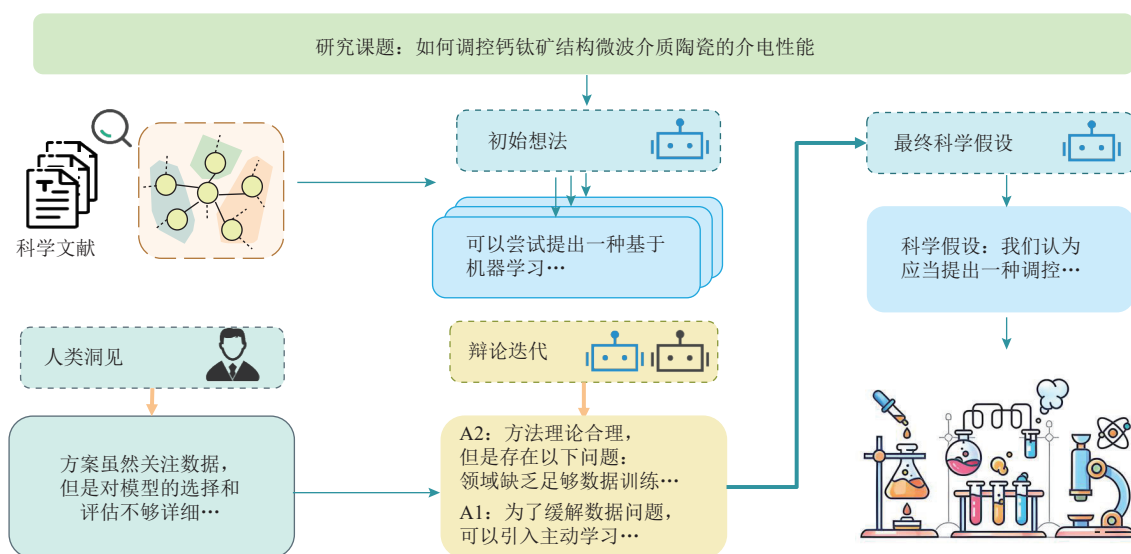


Fig. 1 HILMA framework

图 1 HILMA 框架

向: 科学假设生成、知识增强的 LLMs, 以及人机协作的智能系统. 这些领域的发展为本文研究奠定了理论基础, 提供了前沿技术.

1.1 科学假设生成

基于人工智能的科学假设生成是 AI4Science 领域的核心问题, 旨在利用 AI 辅助或自动化科学探索的初始阶段^[13-14]. 已有研究主要集中在研发能够从现有数据中预测新科学现象的算法和模型^[15]. 近年来, 随着机器学习技术的蓬勃发展, 尤其是深度神经网络在众多数据密集型任务上的成功应用, 科学假设生成的研究也扩展到更复杂的科学问题求解中. LLMs 是这一趋势的新发展, 如何利用其参数化知识提出合理有用的科学假设, 成为了一个备受关注的问题^[16-17]. 例如, Shojaei 等人^[18]研究借助 LLMs 的编程能力有效发现科学方程. Lu 等人^[19]提出了关于 AI 科学家的构想, 其通过设定初始模板, 促使 LLMs 生成众多研究思路. 与此同时, 将这些思路与现有文献中的方法进行对比, 进而筛选出评分较高的研究思路, 以应用于后续的实验中. 尽管在科学研究的求解和应用中对 LLMs 的探索不断深入^[20], 但它们生成的科学假设在创新性和深度方面仍存在较大不足.

本文研究基于结构智力理论中的发散思维和收敛思维, 提出人机协作的多智能体框架. 通过结合实时的系统化知识检索增强机制, 动态整合最新科研进展, 为模型提供前沿完备的信息基础. 同时, 通过多智能体辩论和人机协作, 进一步优化和精炼生成的假设, 增强假设的多样性和论证深度.

1.2 知识增强的 LLMs

LLMs 的生成阶段存在幻觉^[21-22]、有害性^[23]、事实性^[24]和缺乏长时记忆^[25]等问题, 知识增强被认为是克服这些局限性的有效方法^[26]. 知识增强的 LLMs 首先使用外部检索器从特定的知识源(如百科、图谱和数据库等)检索相关的结构化和非结构化知识; 然后将检索到的知识作为 LLMs 的外部上下文信息, 引导 LLMs 生成以知识为基础的回答^[27].

已有研究关注于检索器和阅读器的优化, 研究人员使用稀疏检索器, 如 BM25^[28]和 TF-IDF^[29]进行相关性计算并检索. 然而, 稀疏方法在提取文本内容的语义特征方面存在不足^[30]. 为解决该问题, 研究人员提出基于语言模型的密集检索方法, 通过将文档和查询编码为稠密向量, 有效地表示文本内容的语义特征^[31-32]. 一些近年来研究探索 LLMs 作为检索器的性能, Shen 等人^[33]证明 LLMs 可以在多个基准数据集上作为零样本检索器使用; 而 Ma 等人^[34]提出了

一种利用 LLMs 的 Listwise Reranker, 在不使用任务特定训练数据的情况下实现了强大的重新排序效果; Sun 等人^[35]探究了 LLMs 的相关性排名, 发现经过引导的 LLMs 能够达到最先进的监督方法的性能.

知识增强模型能够准确捕捉和应用专业知识, 在需要深度领域知识的科学研究中尤为重要. Jeong 等人^[36]在生物医学领域引入知识增强的 LLMs, 通过检索特定领域的文档和让 LLMs 自我反思, 能够生成准确且有解释性的答案, 有效地支撑了生物医学领域的发展. 此外, 一些研究尝试实时更新模型的知识库, 以保证其输出的时效性和准确性^[37]. 通过知识增强, 不仅提高了 LLMs 在特定领域内的表现, 还能够适应快速变化的科研环境, 生成与当前研究前沿对齐的内容.

1.3 人机协作的智能系统

人机协作的智能系统强调智能系统与人类的互补性, 旨在将人类的创造力和决策能力与机器的计算能力和数据处理速度相结合^[38], 协同完成特定任务. 传统研究主要集中于提高人与机器人、AI 系统等智能体交互的效率, 以满足人类需求^[39]. LLMs 的兴起标志着该领域的重大转变, 人类的反馈和推理在增强智能体能力方面的作用日益彰显, 能够显著提升智能体的表现. 近年来的研究采用启发式规则或可学习的算法引导智能体寻求人类的帮助^[40-41]. 此外, 研究者开始重视探究特定的引导提示, 以激励基于 LLMs 的智能体主动寻求人类的输入, 从而在这些协作系统中构建更具互动性和协作性的应用^[42-43]. Feng 等人^[44]设计了一种通用可学习的方法, 通过直接规划的形式, 实现人类与 LLMs 之间的高效协作. Dhillon 等人^[45]的研究表明, 通过提供 AI 辅助, 可以显著提升人类用户的写作质量和效率. 李戈等人^[46]探讨了 LLMs 在人机协同软件开发与演化中的应用及其带来的挑战, 强调了人在软件开发与演化中的主导地位 and 可信保障的重要性. 在需求工程领域, 靳东明等人^[47]提出了 ChatModeler 框架, 通过 LLMs 和人类的协同合作, 优化了需求获取和建模过程. 该框架利用模型自动处理任务, 减轻人类的负担, 同时能根据反馈进行调整, 提高需求模型的质量和交互效率.

在科学研究领域, 人机协作系统具备重要意义. 复杂科学问题的研究要求同时具备深厚的专业知识和高效的信息处理能力, 而人类与 LLMs 在这一过程中各自具备独特的优势. 通过人机协作, 能够提高科学假设的质量和創新性, 共同创造出符合科学标准且创新性强的科学假设.

2 人机协作的多智能体辩论框架

本文研究提出了一种新的科学假设生成框架 HILMA, 旨在通过人机协作和多智能体技术, 从海量的科学文献中提取和生成有价值的科学假设. 首先, HILMA 通过关键词查找与研究主题相关的文献, 并且基于文献引用关系网络进行拓展和筛选, 确保输入数据的质量和相关性; 然后, 基于系统化的文献信息, 使用 LLMs 进行深入探索与分析, 生成初步的科学假设; 最后, 利用人机协作的多智能体系统进行科学假设迭代, 模拟创新孕育的思维发散和收敛过程, 最终实现科学假设生成的准确性和创新性.

2.1 基于引文网络的文献系统化检索增强方法

为了给 LLMs 提供系统化的科学知识, 本节提出了基于引文网络的文献系统化检索增强方法. 通过对文献引文网络的自适应组织和构建, 有效地整合了最新和广泛的科学研究成果, 系统性地梳理不同研究点的研究脉络, 确保为 LLMs 提供的知识具备前沿性和系统性. 方法主要分为 2 个部分: 1) 自顶向下的子图引文网络构建. 从相关研究关键词出发, 查找相关中心文献, 并且基于中心文献深入挖掘, 构建子图引文网络. 2) 自底向上的文献网络综述生成. 从各个子图引文网络出发, 汇总每个子图网络的研究脉络和现状, 形成高阶的子图研究综述. 二者相结合, 能够实现对特定领域文献的深入挖掘和梳理, 为 LLMs 提供系统可靠的文献知识来源, 图 2 展示了基于引文网络的文献系统化检索增强流程.

2.1.1 自顶向下的子图引文网络构建

为了在科学文献领域进行系统化检索增强, 采用基于引文网络的方法, 旨在构建一个结构化的文献组织和查询系统. 本节详细介绍了自顶向下的子图引文网络构建过程, 包括关键词检索、文献筛选和

网络拓展等步骤.

首先, 利用 Semantic Scholar API^[48] 进行核心关键词的检索, 以获取与研究内容相关文献的数字对象唯一标识符 (digital object unique identifier, DOI). 这一步骤的关键在于选择能够准确代表研究领域核心的关键词, 以确保检索结果的准确性和全面性. 由于检索结果数量庞大且杂乱, 需要对文献进行筛选, 保留具有重要性和影响力的中心文献. 这一步骤借助文献的被引次数、文献类型以及文献来源来进行评估和排序. 具体地, 定义中心文献集合为 $\{c_1, c_2, \dots, c_m\}$, 其中 m 表示中心文献的数量.

基于确定的中心文献, 将从引文、被引和相关文献 3 个维度进行子图网络的拓展与构建. 具体地, 通过文献的引用关系、被引关系、相关关系来构建基于特定中心文献 c 的子图引文网络, 表示为 $G_c = (V_c, E_c)$, 其中 V 表示网络的节点集合, E 表示边集合. 引用网络、被引网络、相关网络的构建过程可以通过以下公式表示:

$$E_{\text{cite}} = \{(v_i, v_c) | v_i \in V_{\text{cite}}, \text{文献} i \text{ 引用 } c\}, \quad (1)$$

$$E_{\text{cited}} = \{(v_i, v_c) | v_i \in V_{\text{cited}}, \text{文献} i \text{ 被 } c \text{ 引用}\}, \quad (2)$$

$$E_{\text{related}} = \{(v_i, v_c) | v_i \in V_{\text{related}}, \text{文献} i \text{ 与 } c \text{ 相关}\}, \quad (3)$$

$$G_c = (V_{\text{cite}} \cup V_{\text{cited}} \cup V_{\text{related}}, E_{\text{cite}} \cup E_{\text{cited}} \cup E_{\text{related}}). \quad (4)$$

通过以上公式, 能够从一个中心文献构建出一个结构化、多维度的文献网络, 该网络以中心文献为核心, 囊括了与研究内容密切相关的文献以及研究的发展脉络和趋势. 为进一步的研究提供了深入的信息基础和理论支持.

2.1.2 自底向上的文献网络综述生成

尽管经过筛选, 所构建的每个子图引文网络仍然较为庞大 (数十篇). 考虑到 LLMs 上下文窗口长度以及计算效率, 本文采取自底而上的文献网络综述生成方法, 将每个子图引文网络中的文献进行汇总, 让 LLMs 生成该研究的综述.

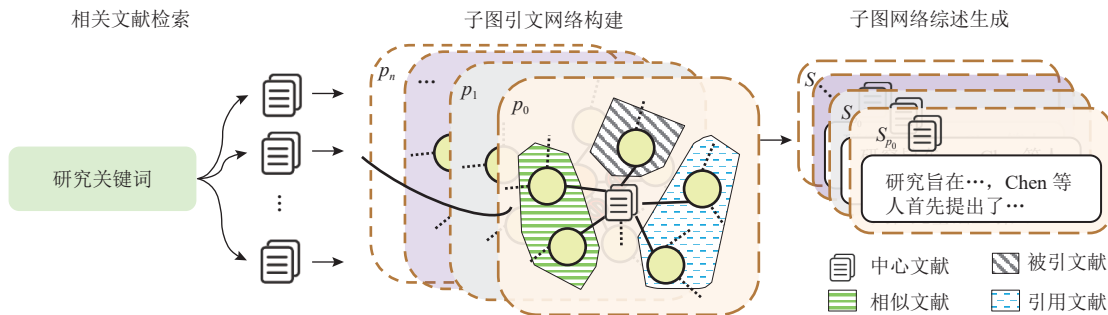


Fig. 2 Systematic literature retrieval enhancement process based on citation network

图 2 基于引文网络的文献系统化检索增强流程

对于每个子图引文网络 $G_i = (V_i, E_i)$, 其每个节点表示一篇文献, LLMs 将利用节点中的文献文本信息, 结合其对该研究领域的深入理解, 撰写研究综述. 这些综述将涵盖该子图引文网络所涉及的研究主题、研究方法、研究成果等方面. 在撰写综述的过程中, LLMs 将考虑文献之间的相互关系, 以及其在研究领域中的重要性和影响力.

经由上述过程, 每个子图引文网络 G_i 都拥有一个对应研究综述 s_i , 其中 $i \in \{1, 2, \dots, m\}$, 这些综述将为 LLMs 提供科学假设生成和辩论的信息基础, 避免不必要的计算和重复查询. 同时, 这些综述还可以为用户提供全面而系统的研究概览, 促进其对后续假设生成的理解.

通过自顶向下的子图引文网络构建和自底向上的文献网络综述生成, 建立了一个结构化、系统化的知识库, 囊括了完善的具体文献和不同子领域的文献综述, 为后续的假设生成和验证提供坚实的基础.

2.2 知识增强的研究想法生成

在构建了基于引文网络的文献系统化检索增强机制后, 本节将探讨如何应用知识增强型 LLMs 来生成具有深度和创新性的研究想法. 该过程旨在整合结构化和系统化的文献知识, 直接融入模型的推理和生成能力中, 以克服传统 LLMs 生成创新科学假设的局限性.

首先, 利用自顶向下的子图引文网络构建方法, 得到了围绕特定中心文献组织的高度结构化的文献子集. 这些子集形成了多维度的信息, 为 LLMs 提供了丰富的背景知识和上下文信息. 每个子集都是围绕其核心文献构建的. 通过分析每个子集的研究综述 $\{s_1, s_2, \dots, s_m\}$, LLMs 能够从研究意义、潜在的研究空间、未来发展趋势等角度进行评分, 从而选择出最具研究潜力和价值的子集. 该选择过程可以形式化表示为

$$\arg \max(score([s_1, s_2, \dots, s_m])), \quad (5)$$

其中 $score$ 表示 LLMs 对综述的评分.

选定研究目标后, 将对应的目标子图 G_i 中的文献进行进一步的结构化组织, 使 LLMs 能够充分获取该领域的发展脉络与研究现状. 在此基础上, LLMs 将综合这些文献中的知识和其自身的语言生成能力, 生成全面而有效的研究想法. 该想法生成过程可以用以下公式表示:

$$h_0 = LLM(P, G_i, s_i), \quad (6)$$

其中 h_0 为初始生成的科学假设, P 为指令提示 (prompt), G_i 为选定的目标引文子图, s_i 为 G_i 的研究

综述.

通过这种知识增强的方法, LLMs 生成的研究提案不仅具有更高的可信度和创新性, 而且能够提升研究提案的整体质量. 知识增强型 LLMs 不仅可以加速研究想法的生成过程, 而且为科学发现的第一阶段提供了更高效的支持.

2.3 人机协作的多智能体辩论迭代

在严谨的科学研究探索中, 创新性假设的孵化是一个动态且迭代的过程, 要求不断质疑、修正与深化. 本节介绍一种创新的人机协作策略, 通过多智能体辩论迭代机制, 旨在实现科学假设的深度精炼与优化. 此方法的核心在于利用来自不同背景和专业知识的智能体 (包括 LLMs 和人类专家) 的多样化视角, 通过辩论和批判性思考来迭代地精炼和增强初步生成的研究想法, 确保每个假设都经过全面的考量和验证.

多智能体辩论迭代的过程设计如下:

1) 角色分配. 每个智能体被分配一个特定角色, 包括假设提出者、审阅者和中立分析师. 这些角色帮助结构化辩论, 使每个智能体可以从不同的角度审视问题.

2) 初始假设生成. 基于 2.2 节知识增强型 LLMs 的输出, 首先形成一组初步的科学假设.

3) 开放式辩论. 智能体之间进行开放式辩论, 每个智能体根据其角色提出支持或反对假设的论点. 此阶段旨在揭示假设的潜在弱点和未考虑的变量. 人类专家在过程中参与并且引导智能体的辩论方向.

4) 证据集成. 辩论中提出的有效点将用于修改或强化假设. 该过程包括请求更多引文子图、重新分析已有数据或引入新的科学理论. 人类专家在此过程中扮演仲裁者的角色, 适时引导辩论的深度与广度, 同时根据辩论进程中的关键发现, 向知识增强型 LLMs 提出数据查询或模型调整的需求, 以获取更多支持或反驳假设的证据.

5) 迭代循环. 每次辩论后都对假设进行评估和调整, 直到达到预定的科学严格性和创新性标准. 每回合辩论后, 整合智能体汇总辩论要点, 与人类专家共同评估假设的改进空间, 从而迭代生成更加成熟与精细的假设版本. 这个过程可以多次迭代.

多智能体辩论迭代的主要目的是利用集体智慧来提高假设的质量和可靠性. 每个假设都经过详尽的挑战和防御, 从而确保其在逻辑上的健全性和在实证基础上的坚实性, 有助于揭示那些可能未被单一智能体注意到的新的研究方向或方法论问题.

在研究框架中,人机协作是关键的一步,它结合了人类专家的创造性思维和高阶决策能力与 LLMs 的高效数据处理和整合的能力.此环节的核心目的是充分发挥人类与 LLMs 智能体各自的优势,共同创造出符合科学标准且创新性强的科学假设.在这一过程中,人类专家负责提供方向性指导、深入分析和复杂决策,这包括指出未被发现的假设漏洞、提供额外的专业知识或重新定义问题的边界,而多智能体则在大数据处理、假设迭代和模式提取方面发挥作用.人机协作模式不仅加速了科学假设的生成过程,还增强了假设的适用性和准确性,确保生成的科学假设既有深度又有广度,同时能够适应复杂多变的科研环境.

2.4 人机协作科学假设生成平台

如图 3 所示,本文研究开发了一个人机协作科学假设生成平台,旨在提升科研人员在科学假设生成过程中的效率和准确性.平台由功能区和对话区组成.在功能区,用户可以选择 LLMs 基座和文献检索源,并控制智能体之间的对话进程及汇总最终结论.在对话区,用户可通过对话框输入想要探索的研究主题或科学问题,平台能够根据输入自主开展系统化的文献检索、增强总结,并生成初步的科学假设.



Fig. 3 Human-machine collaboration hypothesis generation platform

图 3 人机协作假设生成平台

平台允许研究人员实时监控和管理多智能体协同生成的科学假设,并在多智能体对话过程中介入和调整.实时监控功能不仅让研究人员能够全面了解假设生成的进展,还允许他们对生成的假设进行实时评估和调整,从而确保最终生成的假设更加符合实际科研需求.

3 实 验

本节将描述数据集、模型、评估设置和实施细节.

3.1 实验设置

通过 Semantic Scholar 的 API 接口获取实时的学术文献信息.在 HILMA 框架中,使用 Qwen-Max 作为基座 LLMs,版本为 qwen-max-0403, top_k 设置为 0.8, 温度 (temperature) 设置为 0.85. 本文以材料学科为例,让 20 名材料专业的研究生就各自的研究领域提出 5 个真实的研究问题,共获得 100 个不同的研究主题.

基于 100 个材料科学的研究问题,使用 3.2 节提到的基线模型生成对应的科学假设.每个基线共获得 100 个科学假设用于评测,共计 600 个生成的科学假设用于评估.

3.2 基线模型

由于科学假设生成是一个全新任务,尚无直接可用的对比基线.因此,将完整的 HILMA 框架与下列通用基线以及消融变体进行比较:

1) ChatGPT^[49]. ChatGPT 是由 OpenAI 开发的 LLMs,能够生成连贯且具有创造性的文本,广泛应用于对话系统、内容创作和文本生成任务.本基线使用 ChatGPT 直接生成科学假设.

2) CoT^[50]. CoT (chain-of-thought) 是一种提示工程方法,旨在引导 LLMs 逐步思考,以生成更加可靠的回答.在实验中,本基线通过在提示词中加入逐步思考的指令,引导 LLMs 进行科学假设生成.

3) ICL^[51]. ICL (in-context learning) 是一种通过上下文提示进行学习和生成的技术,依赖于提供给模型的输入上下文,以便生成相关的输出.在实验中,使用人工撰写的科学假设样例作为上下文提示,让模型参考这些示例生成科学假设.

4) RAG. RAG (retrieval-augmented generation) 是一种结合检索和生成的技术,通过在生成过程中动态检索相关文献来增强模型的生成能力.通过主题关键词实时检索相关的科研论文摘要,选取前 10 篇提供给 LLMs 作为辅助参考.这一方法能够在一定程度上缓解 LLMs 知识更新不及时的问题,提供更为新颖和前沿的科研信息.

5) Multi-agent. 多智能体系统通过多个具有不同角色和功能的智能体协同工作来实现复杂任务.

本研究设计了多智能体框架基线,其中不同的智能体负责不同的任务,包括假设提出者、批评者或中立分析师.通过智能体之间的自动化协作和辩论,迭代生成科学假设.本文方法旨在最大化利用集体智慧,提高假设的科学性和创新性.

3.3 评估设置

鉴于科学假设生成是一个新任务,没有已知基准

可以衡量其生成质量. 因此, 采用基于模型的自动评估与人类评估相结合的方法, 来验证实验基准模型.

1) 基于模型的评估. 参考最近使用 LLM 判断输出文本质量的范式^[52-53], 使用 GPT-4 来判断生成的科学假设的质量. 以 5 个不同的标准衡量科学假设的质量, 然后要求评估模型对每个标准的生成思路进行 5 点李克特量表评分^[54]. 表 1 中提供了专家制定的用于引导评估的详细标准和提示.

2) 基于人类的评估. 类似于基于模型的评估, 同

Table 1 5-Point Likert Scale Used in Model Evaluation

表 1 模型评估中使用的 5 点李克特评分量表

评估指标	评分	描述
创新性	1	没有任何新颖之处, 重复已有的研究成果.
	2	有少量新颖之处, 但大部分内容是已有知识的延伸.
	3	有一些新的观点或方法, 但总体仍基于已有的框架.
	4	提出了较为新颖的观点或方法, 有显著的创新点.
	5	极具创新性, 提出了全新的观点或方法, 可能引发领域内的重大变革.
实用性	1	完全没有实际应用价值, 难以转化为实际应用.
	2	有一定的理论价值, 但实际应用价值有限.
	3	有一定的实际应用潜力, 但需进一步研究和验证.
	4	具备较高的实际应用价值, 具有较好的转化前景.
	5	极具实际应用价值, 能够迅速转化为具体应用并产生显著效益.
可行性	1	完全不可行, 技术或资源方面难以实现.
	2	可行性较低, 面临较大的技术或资源挑战.
	3	有一定的可行性, 但需要克服一些技术或资源障碍.
	4	基本可行, 技术和资源要求在可控范围内.
	5	完全可行, 现有技术和资源可以支持其实现.
数据支持	1	完全缺乏数据支持, 无法验证假设.
	2	数据支持有限, 无法充分验证假设.
	3	有一定的数据支持, 但不完全, 需进一步数据验证.
	4	有较充分的数据支持, 能够验证假设的大部分内容.
	5	数据支持非常充分, 能够全面验证假设.
理论基础	1	缺乏理论基础, 无法与现有科学理论相一致.
	2	理论基础薄弱, 与现有科学理论存在较大矛盾.
	3	有一定的理论基础, 但与现有科学理论存在一些不一致.
	4	具有较好的理论基础, 能够与现有科学理论相结合.
	5	理论基础非常坚实, 与现有科学理论完全一致, 并有可能推动理论的发展.
整体评价	1	整体评价很差, 假设缺乏科学价值和实际意义.
	2	整体评价较差, 假设存在较多不足之处.
	3	整体评价一般, 假设有一定价值但需改进.
	4	整体评价较好, 假设具备较高的科学价值和实际意义.
	5	整体评价非常好, 假设具有重要的科学价值和实际意义.

一个主题下不同基线模型生成的科学假设将被两两配对, 让人类评注者在 2 个隐去了模型信息的回答之间进行成对比较, 选择质量更高的科学假设. 为了实现这一目标, 本研究开发了在线评估平台, 如图 4 所示, 标注者登录平台后, 平台会自动推送针对同一个主题的匿名答案对, 评注者通过点击按钮即可实现标注, 能够协同自动保存并统计评注者的偏好选择结果. 为了保证人类评估的质量, 专家评注者均为熟悉该领域的硕士研究生与博士研究生, 且至少发表过 1 篇学术论文, 评估过程共由 5 位专家评注者进行.



Fig. 4 The constructed artificial online evaluation platform

图 4 构建的人工在线评测平台

3.4 实验结果与分析

为了比较本文模型与基准方法的效果差异, 对生成的科学假设数据进行了基于 LLMs 的评分测试. 表 2 展示了不同基线模型李克特量表的 GPT-4 评分结果.

Table 2 Comparison of Different Baselines

表 2 不同基线模型对比

模型	创新性	实用性	可行性	数据支持	理论基础	总体评价
ChatGPT	3.29	3.41	2.80	2.27	3.27	3.07
CoT	3.81	3.98	2.94	2.37	3.64	3.32
ICL	4.09	4.07	3.21	2.60	3.72	3.70
RAG	4.10	3.94	3.18	2.65	3.80	3.61
Multi-agent	4.51	4.20	3.13	2.46	4.08	3.96
HILMA (本文)	4.60	4.25	3.40	2.80	4.20	4.10

注: 黑体数值表示最优值.

首先, 从基线模型的整体评分来看, LLMs 生成的科学假设在实用性和理论基础方面取得了相对较高的分数, 这主要归因于 LLMs 在海量文本上的预训练能够掌握大量基础理论和常识知识. 然而, 在创新性、可行性和数据支持 3 个评分维度上效果不佳. 这是因为 LLMs 局限于已经见过的重复知识和模式, 缺乏创新性思维, 并且容易提出不切实际的想法, 缺乏

可行性和理论支持。

其次,通过对比 CoT, ICL, ChatGPT 的表现,可以发现通过引导 LLMs 逐步思考和提供具体的科学假设样例,能够有效地提升生成假设的质量,这说明通用的 LLMs 能力提升方法对科学假设任务也是适用的;通过检索最新的相关文献, RAG 方法能够显著提升生成的科学假设质量,这得益于模型能够获取最新的科研进展,有效地提升了创新性和数据支撑评分,使得生成的科学假设具备前沿性和有效性。此外,得益于多智能体迭代讨论, Multi-Agent 取得了更优的模型评分,说明通过最大化利用集体智慧,能够提高假设的科学性和创新性。

本文的 HILMA 框架显著优于所有的基线模型,创新性评分达到了 4.6 分,总评分是所有基线中唯一突破 4 分的模型,达到了“具备较高的科学价值和实际意义”的标准。实验结果表明通过充分利用人类与 LLMs 智能体各自的优势,有助于创造出符合科学标准且创新性强的科学假设。此外,实验结果突出了人机协作的重要性,在这一过程中,人类专家负责提供方向性指导、深入分析和复杂决策,而 LLMs 则通过引文网络构建、多智能体迭代等不断优化假设,增强假设的适用性和准确性,确保生成的科学假设既有深度也有广度。

为了进一步验证人机协作的重要性,首先使用 HILMA 框架生成了未经讨论的初始假设,然后基于此进行人机协作,迭代生成最终的科学假设,以对比二者的评分差异。如图 5 所示,通过人机协作迭代能够在所有指标上显著提升模型的生成效果。这是因为人类专家能够提供方向性指导,帮助 LLMs 进行复杂决策和启发新思路。在人机迭代过程中,一些细节错误和假设不合理的地方也能够被及时发现和改正,进一步提升假设的整体质量。

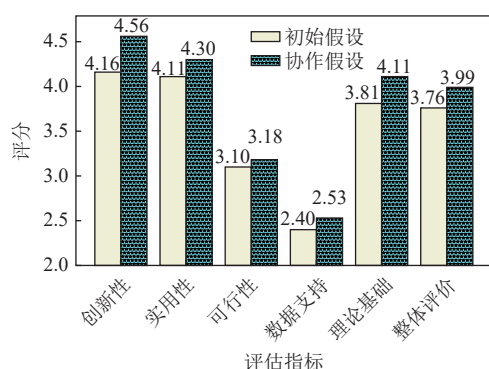


Fig. 5 Comparison of initial hypotheses and hypotheses after human-machine collaboration

图 5 初始假设与人机协作后的假设对比

图 6 展示了基于子图引文网络生成的科学假设与基于 LLMs 本身的科学假设的雷达图对比。经过了子图引文网络能够显著增强 LLMs 生成科学假设的创新性与总体评价,表明子图引文网络能够为 LLMs 提供最新的科学文献与研究进展,帮助 LLMs 从最新研究中探索创新思路与前沿方法。

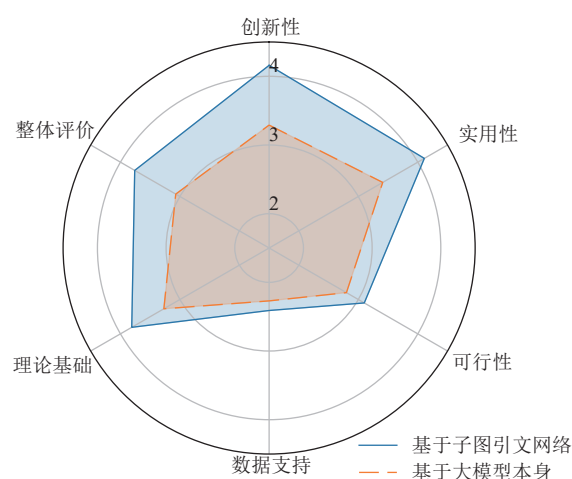


Fig. 6 Impact of subgraph citation network introduction on generation quality

图 6 子图引文网络的引入对生成质量的影响

3.5 多步辩论评估

为验证多轮辩论迭代的效果,开展了多轮评估实验。基于相同的初始假设,在不同的迭代轮次终止并生成科学假设,以观察其质量随迭代轮次的变化。如图 7 所示,经过多轮迭代, Multi-agent 和 HILMA 方法的整体质量均有所提升,这是因为在多轮辩论中,假设中存在的明显缺陷和不足能够被发现和纠正。

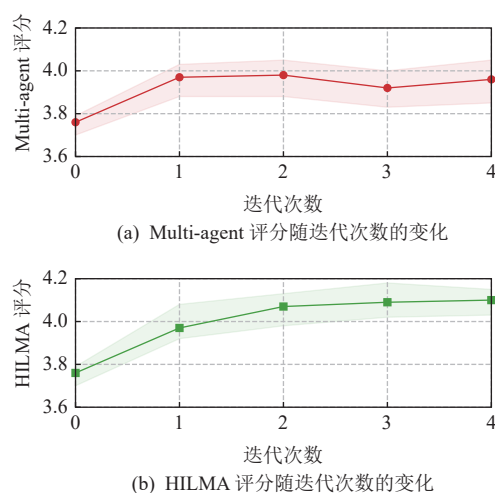


Fig. 7 Impact of iteration count on human-machine collaboration

图 7 迭代次数对人机协作的影响

然而,注意到 Multi-agent 方法在迭代 2 轮后质量有所提升,但在后续轮次中出现了停滞甚至下降的现象.这主要是由于多智能体缺乏对科研的深刻理解,在讨论中容易集体跑偏,过度关注无关细节,忽略了科学假设的核心内容.相比之下, HILMA 方法的假设质量随着迭代次数的增加表现出更高的稳定性.人类专家的高阶洞见能够及时纠正讨论中的偏差,指出问题所在,不断提升科学假设的总体质量.这突显了人机协作的重要性.

3.6 不同基座模型评估

表 3 展示了 HILMA 在开源模型 Llama-3(Llama3-70B-Instruct^[55])、Qwen-72B(Qwen1.5-72B-Chat^[56])和闭源模型 GPT-3.5(GPT-3.5-Turbo^[57])、通义千问(Qwen-Max^[58])上的实验效果.结果表明,生成的假设质量与模型本身的能力密切相关,基座模型越强,生成的假设质量越高.其中, Qwen-Max 表现出最佳的效果.

Table 3 Comparison of Different Large Language Models

表 3 不同大语言模型对比

基座模型	创新性	实用性	可行性	数据支持	理论基础	整体评价
GPT-3.5	4.05	4.01	3.12	2.45	3.69	3.57
Llama-3	4.08	4.08	3.07	2.28	3.87	3.62
Qwen-72B	4.17	4.00	3.14	2.51	3.96	3.77
Qwen-Max	4.60	4.25	3.40	2.80	4.20	4.10

3.7 人类评估结果

对基线模型进行人工评估,图 8 展示了不同模型初始假设与人机协作后假设的对比情况.热力图展示了在人工评测中,两两模型评测的胜率.

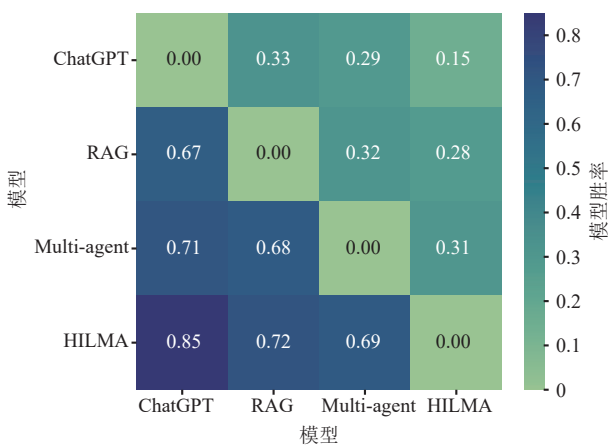


Fig. 8 Heatmap of model win rates based on human evaluation

图 8 基于人工评估的模型胜率热力图

如图 8 所示, HILMA 框架生成假设在与其他模型的对比中胜率显著,说明其整体质量优于其他基

线模型.在初始假设的对比中, ChatGPT, RAG, Multi-agent, HILMA 的质量依次递增,其中 HILMA 框架的初始假设质量最高.这与 3.4 节的模型评估结果一致,进一步验证了 HILMA 框架在科学假设生成中的优越性和潜力.

3.8 案例分析

表 4 展示了以“氮化硅陶瓷”为主题词, ChatGPT, RAG, Multi-agent, HILMA 生成的科学假设.其中, ChatGPT 的假设提出了结合材料科学和电子工程的设计思路,但缺乏对如何实现这种结合的具体细节. RAG 通过检索增强的方法,提出了微系统的概念,但同样缺少具体的实施步骤和技术细节,使得这些假设在转化为研究方案时存在一定难度.

相比之下, Multi-agent 和 HILMA 生成的科学假设则更加具体且具有可操作性. Multi-agent 详细描述了化学合成、电化学技术、量子化学计算等实验方法,并提供了清晰的研究路径. HILMA 的假设则更为全面和深入,它提出了一种多级结构的氮化硅陶瓷复合材料设计,详细阐述了每一层级材料设计的方法,给出了预期的微观结构和性能提升的理论解释.这些详细的方法和步骤使这一假设更容易转化为实际的科研实践,具备较高的可操作性和可实施性,更符合使用者的预期.

4 结 论

在当今科学研究不断发展的背景下,创新性假设的生成变得尤为重要.然而,面对海量科学文献和复杂的跨学科知识体系,传统的科研方法在信息处理与知识整合方面面临巨大挑战.本文提出了一种基于人机协作的多智能体科学假设生成框架 HILMA,通过结合结构智力理论中的发散思维和收敛思维,有效地提升了科学假设的生成质量和创新性.实验结果表明,与现有的基线模型相比, HILMA 框架在生成高质量科学假设方面展现出了显著的优势.这不仅在理论上验证了结构智力理论在科学研究中的应用价值,更在实践中证明了人机协作在科学创新中的潜力.

HILMA 结合了多智能体和人类协作,具备广泛的可扩展性,但仍然存在一些不足,主要体现在系统化检索效率与人机协调难度上.为了让 LLMs 具备充足且系统化的知识, HILMA 需要对大量文献进行检索、筛选和总结,这一过程涉及多次的文献 API 检索和 LLMs 的调用,耗时较长,可能影响用户体验;在人

Table 4 Cases of Scientific Hypothesis Generation Taking “Silicon Nitride Ceramics” as the Research Topic
表 4 以“氮化硅陶瓷”为研究主题的科学假设生成案例

模型	科学假设
ChatGPT	我建议氮在硅氮化陶瓷的研究中, 结合材料科学和电子工程的领域. 利用电子工程中传感器技术和材料科学中的纳米结构设计原理, 设计并研发一种新型的硅氮化陶瓷传感器. 该传感器可用于高温、高压或腐蚀性环境下的测量, 并具有更高的灵敏度和精准度. 这一想法受到材料科学中硅氮化陶瓷的高温稳定性和电子工程中传感器技术的启发.
RAG	我提议开发一种集成的硅氮化物微系统, 该系统利用辐射热交换和优化的热响应时间来提高传感器对温度变化的敏感性. 受上述关于自由站立硅氮化物膜热耦合特性的研究启发, 我们可以设计微机械谐振器, 其中辐射热交换被刻意管理和利用, 以增强其作为温度传感器的性能. 通过精确控制膜的尺寸、厚度和发射率, 可以调整辐射热交换在总热耦合中的贡献. 此外, 结合高真空环境中的实验验证, 我们可以创建一个模拟平台来预测和优化这些微系统的热噪声和频率噪声特性. 这样, 我们不仅可以提高现有传感器的性能, 还能为未来的温度感应和热管理应用开辟新途径.
Multi-agent	我提议结合氮化硅陶瓷的研究和实验, 探索其在腐蚀抑制领域的应用潜力. 氮化硅陶瓷具有优异的耐磨、耐腐蚀和高温稳定性, 因此可以作为一种新型的腐蚀抑制材料, 特别适用于极端环境下的金属保护. 首先, 可以通过化学合成方法, 制备不同形貌和结构的氮化硅陶瓷颗粒或涂层, 并使用表征技术 (如 SEM, TEM, XRD) 对其进行详细的形貌和结构分析. 这有助于优化氮化硅陶瓷材料的合成方法, 从而提高其实际应用价值. 其次, 利用电化学技术 (如 EIS 和 PDP), 评估氮化硅陶瓷材料对金属表面腐蚀的抑制效果, 并研究其与金属表面的电化学反应动力学和机理. 这将有助于深入理解氮化硅陶瓷材料的腐蚀抑制机理, 从而提高其实际应用效果. 进一步, 结合量子化学计算方法 (如 DFT), 预测氮化硅陶瓷材料在不同腐蚀介质下的吸附行为和电子转移机制, 以深入理解其腐蚀抑制机理. 这有助于优化氮化硅陶瓷材料的设计和應用范围, 提高其实际应用效果. 最后, 探索氮化硅陶瓷与其他腐蚀抑制剂 (例如有机分子或无机纳米材料) 的复合应用, 以期获得协同效应, 提高金属表面的腐蚀防护性能. 这将为腐蚀抑制领域提供新的思路和方法, 推动相关交叉研究的进一步探索. 这一系列综合研究将有助于拓展氮化硅陶瓷材料在腐蚀抑制领域的应用范围, 为开发新型高效的金属腐蚀抑制策略提供新的理论和实践基础. 期待这一综合性研究能够为相关领域的科学研究和工程实践带来积极的影响.
HILMA (本文)	我提议, 通过设计一种多级结构的氮化硅陶瓷复合材料, 该材料同时整合了微纳尺度的增强相和智能响应型腐蚀抑制剂, 以实现超耐腐蚀、自适应修复及显著提高机械性能的革新目标. 此复合材料的核心设计包括以下几个层次: 1) 基层优化. 首先, 对氮化硅基层进行成分与微观结构的双重优化, 采用 Y_2O_3 稳定的 $\alpha-Si_3N_4$ 为主相, 结合溶胶-凝胶法制备的超细晶粒 ($< 500\text{ nm}$), 以确保材料本体的高强度和高韧性. 通过梯度分布的晶粒尺寸设计, 从表面到内部逐步增大晶粒, 以获得更好的表面耐磨性及整体抗裂纹扩展能力. 2) 微米级增强相分散. 在基层内均匀分散碳化硅 (SiC) 晶须或短纤维作为第 2 相增强材料, 利用其高模量和优异的界面粘结性, 进一步提升复合材料的断裂韧性和耐磨性. 通过控制其体积分数和定向排列, 可以有效引导裂纹偏转和桥接, 减少裂纹扩展. 3) 纳米级智能抑制剂包覆. 在表面微结构层, 将智能响应型有机抑制剂 (如温敏或 pH 敏感型聚合物包裹的苯并咪唑衍生物) 通过原位聚合合法包覆在纳米粒子 (如 SiO_2) 表面, 随后嵌入到氮化硅表面层. 这些智能抑制剂在遇到腐蚀介质时能自动释放活性分子, 迅速封闭陶瓷表面的微裂纹或缺陷, 同时在介质条件恢复时重新吸附回纳米载体, 实现自适应的腐蚀防护. 4) 梯度功能涂层. 最外层采用 CVD 技术沉积一层梯度功能涂层, 从富含智能抑制剂的表面逐渐过渡到富含耐磨增强颗粒 (如类金刚石碳) 的底层, 旨在结合化学抑制与物理阻挡的双重防护机制, 进一步提升耐腐蚀与耐磨性能. 通过综合运用先进材料科学、纳米技术和智能材料设计理念, 此多级结构氮化硅陶瓷复合材料有望在极端环境下展现前所未有的耐久性和可靠性, 特别适用于航空航天、深海探索及高级能源系统等对材料性能要求极高的领域. 实验验证将涵盖从材料合成、微观结构表征、腐蚀行为评估到机械性能测试的全方位研究, 以全面评估该创新方案的有效性和潜在应用价值.

机协作上, 当前市面上可用的 LLMs 均为助理模型, 存在过度倾向于满足人类指令的问题, 因此, 人类的微小意见可能对整个协作的走向和最终生成内容产生较大影响, 进而影响科学假设的完整性. 因此, 如何更好地均衡人类和多智能体之间的协作需要进一步的探索研究. 此外, 虽然 HILMA 是通用的科学假设生成框架, 但是受限于原始数据和评测条件, 目前只在材料科学领域验证了 HILMA 的效果, 并且缺少实验结果的支撑, 需要结合材料合成实验来进一步验证科学假设的可行性和效果.

未来, 拓展 HILMA 框架至不同的学科领域将成为重点探索的方向. 鉴于不同学科在研究内容和方式上存在较大差异, 单一框架难以在所有学科均取得理想效果. 从数据视角来看, 医学影像、电子信号、语音学等在数据处理方面对图像、信号、音视频等有着巨大需求的专业, 单纯的文本增强生成往往难以达成良好的成效. 探索如何利用多模态 LLMs 实现

多模态的系统化检索增强, 是一个值得深入研究的方向. 多元的数据能够进一步提升假设生成的可靠性和完整性.

此外, 伴随各个学科的理论和技术不断相互交叉与融合, 学科交叉已成为显著趋势. 如何针对不同领域构建专家智能体, 针对同一课题分配多领域角色以实现学科交叉讨论, 也将是一个值得深入探究的重要内容. 通过进一步优化人机协作机制和智能体辩论策略, 科学研究的效率和创新性有望得到进一步提升, 为解决复杂的科学问题和推动技术进步提供新的思路和方法.

作者贡献声明: 陈子阳提出了算法思路 and 实验方案, 并撰写论文; 赵翔提出指导意见, 完成论文的修改和校对; 赵润豪参与了文献调研和论文修订工作; 倪子淇完成了文献调研、实验数据收集和论文修订工作; 叶益聪提出指导意见.

参 考 文 献

- [1] Bornmann L, Haunschild R, Mutz R. Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases[J]. *Humanities and Social Sciences Communications*, 2021, 8(1): 1–5
- [2] Rothenberg A. The Janusian process in scientific creativity[J]. *Creativity Research Journal*, 1996, 9(2/3): 207–231
- [3] Birhane A, Kasirzadeh A, Leslie D, et al. Science in the age of large language models[J]. *Nature Reviews Physics*, 2023, 5(5): 277–280
- [4] Fakhoury S, Naik A, Sakkas G, et al. LLM-based test-driven interactive code generation: User study and empirical evaluation[J]. arXiv preprint, arXiv: 2404.10100, 2024
- [5] Wu Yiquan, Zhou Siying, Liu Yifei, et al. Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 12060–12075
- [6] Thirunavukarasu A J, Ting D S, Elangovan K, et al. Large language models in medicine[J]. *Nature Medicine*, 2023, 29(8): 1930–1940
- [7] Liu Yiheng, Han Tianle, Ma Siyuan, et al. Summary of ChatGPT-related research and perspective towards the future of large language models[J]. *Meta-Radiology*, 2023, 1(2): 1–14
- [8] Meyer J G, Urbanowicz R J, Martin P C, et al. ChatGPT and large language models in academia: Opportunities and challenges[J]. *BioData Mining*, 2023, 16(1): 20–31
- [9] Walsh E, Anders K, Hancock S, et al. Reclaiming creativity in the era of impact: Exploring ideas about creative research in science and engineering[J]. *Studies in Higher Education*, 2013, 38(9): 1259–1273
- [10] Chen Ziyang, Li Dongfang, Zhao Xiang, et al. Temporal knowledge question answering via abstract reasoning induction[J]. arXiv preprint, arXiv: 2311.09149, 2023
- [11] Ziems C, Held W, Shaikh O, et al. Can large language models transform computational social science?[J]. *Computational Linguistics*, 2024, 50(1): 237–291
- [12] Guilford J P. The structure of intellect[J]. *Psychological Bulletin*, 1956, 53(4): 267–293
- [13] Wang Hanchen, Fu Tianfan, Du Yuanqi, et al. Scientific discovery in the age of artificial intelligence[J]. *Nature*, 2023, 620(7972): 47–60
- [14] Baek J, Jauhar S K, Cucerzan S, et al. ResearchAgent: Iterative research idea generation over scientific literature with large language models[J]. arXiv preprint, arXiv: 2404.07738, 2024
- [15] Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: A preliminary study using GPT-4[J]. arXiv preprint, arXiv: 2311.07361, 2023
- [16] Majumder B P, Surana H, Agarwal D, et al. Data-driven discovery with large generative models[J]. arXiv preprint, arXiv: 2402.13610, 2024
- [17] Qi Biqing, Zhang Kaiyan, Li Haoxiang, et al. Large language models are zero shot hypothesis proposers[J]. arXiv preprint, arXiv: 2311.05965, 2023
- [18] Shojaee P, Meidani K, Gupta S, et al. LLM-SR: Scientific equation discovery via programming with large language models[J]. arXiv preprint, arXiv: 2404.18400, 2024
- [19] Lu C, Lu Cong, Lange R T, et al. The AI scientist: Towards fully automated open-ended scientific discovery[J]. arXiv preprint, arXiv: 2408.06292, 2024
- [20] Li Yunxin, Hu Baotian, Shi Haoyuan, et al. VisionGraph: Leveraging large multimodal models for graph theory problems in visual context[J]. arXiv preprint, arXiv: 2405.04950, 2024
- [21] Ji Ziwei, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. *ACM Computing Surveys*, 2023, 55(12): 1–38
- [22] Shuster K, Spencer P, Chen M, et al. Retrieval augmentation reduces hallucination in conversation[C]//Proc of the 2021 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 3784–3803
- [23] Wang Mengru, Yao Yunzhi, Xi Zekun, et al. Safety analysis of large model content generation based on knowledge editing[J]. *Journal of Computer Research and Development*, 2024, 61(5): 1143–1155 (in Chinese)
(王梦如, 姚云志, 习泽坤, 等. 基于知识编辑的大语言模型内容生成安全分析[J]. *计算机研究与发展*, 2024, 61(5): 1143–1155)
- [24] Wang Cunxiang, Liu Xiaozhe, Yue Yuanhao, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity[J]. arXiv preprint, arXiv: 2310.07521, 2023
- [25] Xu Xinchao, Gou Zhibin, Wu Wenquan, et al. Long time no see! Open-domain conversation with long-term persona memory[C]//Proc of the 60th Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 2639–2650
- [26] Gao Yunfan, Xiong Yun, Gao Xinyu, et al. Retrieval-augmented generation for large language models: A survey[J]. arXiv preprint, arXiv: 2312.10997, 2023
- [27] Lewis L, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Proc of the 34th Conf on Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 9459–9474
- [28] Robertson S E, Zaragoza H. The probabilistic relevance framework: BM25 and beyond[J]. *Foundations and Trends in Information Retrieval*, 2009, 3(4): 333–389
- [29] Wu H, Luk P W P, Wong K F, et al. Interpreting TF-IDF term weights as making relevance decisions[J]. *ACM Transactions on Information Systems*, 2008, 26(3): 1–37
- [30] Guo Jiafeng, Cai Yinqiong, Fan Yixing, et al. Semantic models for the first-stage retrieval: A comprehensive review[J]. *ACM Transactions on Information Systems*, 2022, 40(4): 1–42
- [31] Bruch S, Gai S, Ingber A. An analysis of fusion functions for hybrid retrieval[J]. *ACM Transactions on Information Systems*, 2023, 42(1): 1–35
- [32] Li Hang, Mourad A, Zhuang Shengyao, et al. Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls[J]. *ACM Transactions on Information Systems*, 2023, 41(3): 1–40
- [33] Shen Tao, Long Guodong, Geng Xiubo, et al. Large language models are strong zero-shot retriever[J]. arXiv preprint, arXiv: 2304.14233, 2023

- [34] Ma Xueguang, Zhang Xinyu, Pradeep R, et al. Zero-shot listwise document reranking with a large language model[J]. arXiv preprint, arXiv: 2304.14233, 2023
- [35] Sun Weiwei, Yan Lingyong, Ma Xinyu, et al. Is ChatGPT good at search? Investigating large language models as re-ranking agents[C]// Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 14918–14937
- [36] Jeong M, Sohn J, Sung M, et al. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models[J]. arXiv preprint, arXiv: 2401.15269, 2024
- [37] Mousavi S M, Alghisi S, Riccardi G. Is your LLM outdated? Benchmarking LLMs & alignment algorithms for time-sensitive knowledge[J]. arXiv preprint, arXiv: 2404.08700, 2024
- [38] Wang Z, Choi D, Xu Shenyu, et al. Putting humans in the natural language processing loop: A survey[J]. arXiv preprint, arXiv: 2103.04044, 2021
- [39] Wu Xingjiao, Xiao Luwei, Sun Yixuan, et al. A survey of human-in-the-loop for machine learning[J]. *Future Generation Computer Systems*, 2022, 135: 364–381
- [40] Cai Zefan, Chang Baobao, Han Wenjuan. Human-in-the-loop through chain-of-thought[J]. arXiv preprint, arXiv: 2306.07932, 2023
- [41] Mehta N, Teruel M, Sanz P F, et al. Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback[C]//Proc of the 62nd Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 1306–1321
- [42] Huang Wenlong, Xia Fei, Xiao T, et al. Inner monologue: Embodied reasoning through planning with language models[J]. arXiv preprint, arXiv: 2207.05608, 2022
- [43] Wang Xingyao, Wang Zihan, Liu Jiateng, et al. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback[J]. arXiv preprint, arXiv: 2309.10691, 2023
- [44] Feng Xueyang, Chen Zhiyuan, Qin Yujia, et al. Large language model-based human-agent collaboration for complex task solving[J]. arXiv preprint, arXiv: 2402.122914, 2024
- [45] Dhillion P S, Molaei S, Li Jiaqi, et al. Shaping human-AI collaboration: Varied scaffolding levels in co-writing with language models[J]. arXiv preprint, arXiv: 2402.11723, 2024
- [46] Li Ge, Peng Xin, Wang Qianxiang, et al. Challenges from LLMs as a natural language based human-machine collaborative tool for software development and evolution[J]. *Journal of Software*, 2023, 34(10): 4601–4606 (in Chinese)
(李戈, 彭鑫, 王千祥, 等. 大语言模型: 基于自然交互的人机协同软件开发与演化工具带来的挑战[J]. *软件学报*, 2023, 34(10): 4601–4606)
- [47] Jin Dongming, Jin Zhi, Chen Xiaohong, et al. ChatModeler: A human-machine collaborative and iterative requirements elicitation and modeling approach via large language models[J]. *Journal of Computer Research and Development*, 2024, 61(2): 338–350 (in Chinese)
(靳东明, 金芝, 陈小红, 等. ChatModeler: 基于大语言模型的人机协作迭代式需求获取和建模方法[J]. *计算机研究与发展*, 2024, 61(2): 338–350)
- [48] Ai2. Semantic Scholar API [EB/OL]. [2024-05-17]. <https://www.semanticscholar.org/product/api>
- [49] OpenAI. ChatGPT [EB/OL]. [2024-05-17]. <https://chat.openai.com>
- [50] Wei J, Wang Xuezhi, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proc of the 36th Conf on Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 24824–24837
- [51] Min S, Lyu X, Holtzman A, et al. Rethinking the role of demonstrations: What makes in-context learning work[C]// Proc of the 2022 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2022: 11048–11064
- [52] Zheng Lianmin, Chiang W, Sheng Ying, et al. Judging LLM-as-a-Judge with MT-Bench and chatbot arena[J]. arXiv preprint, arXiv: 2306.05685, 2023
- [53] Fu Jinlan, Ng S K, Jiang Zhengbao, et al. Gptscore: Evaluate as you desire[J]. arXiv preprint, arXiv: 2302.04166, 2023
- [54] Joshi A, Kale S, Chandel S, et al. Likert scale: Explored and explained[J]. *British Journal of Applied Science & Technology*, 2015, 7(4): 396–403
- [55] Meta. Llama3-70B-Instruct[EB/OL]. [2024-05-10]. <https://hugging-face.co/meta-llama>
- [56] Alibaba. Qwen1.5-72B-Chat[EB/OL]. [2024-05-10]. <https://hugging-face.co/Qwen/Qwen1.5-72B-Chat>
- [57] OpenAI. GPT-3.5-Turbo[EB/OL]. [2024-05-10]. <https://platform.openai.com/docs/models/gpt-3-5-turbo>
- [58] Alibaba. Qwen-Max API[EB/OL]. [2024-05-10]. <https://help.aliyun.com/zh/dashscope/developer-reference>



Chen Ziyang, born in 1999. PhD candidate. Member of CCF. His main research interests include natural language processing, knowledge graph, and large language models.

陈子阳, 1999年生. 博士研究生. CCF会员. 主要研究方向为自然语言处理、知识图谱、大语言模型.



Zhao Xiang, born in 1986. PhD, professor. Distinguished member of CCF. His main research areas include graph data management and analysis, knowledge graphs, and big data knowledge engineering.

赵翔, 1986年生. 博士, 教授. CCF杰出会员. 主要研究方向为图数据管理与分析、知识图谱、大数据知识工程.



Zhao Runhao, born in 2002. Master candidate. His main research interests include natural language processing, knowledge graph question answering, and data fusion.

赵润豪, 2002年生. 硕士研究生. 主要研究方向为自然语言处理、知识图谱问答、数据融合.



Ni Ziqi, born in 2000. PhD candidate. Her main research interests include materials informatics and energy storage ceramics.

倪子淇, 2000 年生. 博士研究生. 主要研究方向为材料信息学、储能陶瓷.



Ye Yicong, born in 1985. PhD, professor. His main research interests include materials informatics and special metal materials.

叶益聪, 1985 年生. 博士, 教授. 主要研究方向为材料信息学、特种金属材料.