基础学科交叉特征及热点主题识别

——以 Web of Science 大语言模型主题论文为例 *

王方媛　徐慧婷
（上海图书馆，上海　200031）

摘　要：［目的 / 意义］本文聚焦大模型在基础学科的应用及其引起的学科交叉现象，为以大模型为代表的人工智能技术推动基础学科交叉融合提供理论支持与实践指引。［方法 / 过程］先从 Web of Science 核心合集的大模型主题研究论文中筛选出生命科学、材料科学、数学、化学、物理学和地球科学六类基础学科的论文，再选择涉及两种及以上基础学科的论文作为研究对象，构建学科交叉主题分析框架。采用 BERTopic 主题建模方法，结合多值邻接矩阵构建学科交叉网络，分析学科交叉特征。基于主题强度、影响力和关注度指标，采用熵权法计算综合主题热度，识别热点主题。［结果 / 结论］六类基础学科具有高度学科交叉性，其中化学、物理学与材料科学的交叉融合最为显著。本研究识别出“蛋白质序列预测与基因分析”等八个主题及“量子自旋与相变理论”“蛋白质序列预测与基因分析”“材料设计与化学合成”三个热点主题。

关键词：大语言模型　基础学科　BERTopic　学科交叉　主题模型
分类号：TP3-05
DOI：10.31193/SSAP.J.ISSN.2096-6695.2025.03.04

0 引 言

科学智能（AI for Science，AI4S）是运用人工智能技术推动科学研究和知识发现的新范式[1]。2024 年，诺贝尔物理学奖和化学奖均颁发给了人工智能（Artificial Intelligence，AI）领域的科学家[2]。2025 年，智源研究院发布的《2025 十大 AI 技术趋势》提出了“科学的未来：AI4S 驱动科学研究范式变革”为十大 AI 技术趋势之首[3]。AI 技术能辅助研究者生成假设、设计实验、处理实验数据，在科学发现中的应用日益广泛[4]。

* 本文系 2024 年度上海市重点智库课题“人工智能驱动科研范式变革研究及对上海布局 AI for Science 的启示”的研究成果之一。
［作者简介］王方媛，女，助理研究员，研究方向为文献计量与科学评价、信息服务与情报分析，Email：fywang@libnet.sh.cn；徐慧婷，女，助理研究员，研究方向为产业与前沿技术、科研范式变革，Email：htxu@libnet.sh.cn（通讯作者）。

近年来，大型语言模型（Large Language Models，LLMs），也称“大语言模型”或“大模型”，以前所未有的速度发展，成为科学智能（AI4S）领域的研究前沿。在大模型时代，科学智能（AI4S）的赋能效果整体上要好于小模型时期[5]。大模型的核心优势在于其卓越的自然语言处理（Natural Language Processing，NLP）能力，在文本分类、命名实体识别和情感分析等任务中表现出色。在高性能算力支持下，大模型正重塑科学研究，成为推动科研突破和解决实际问题的强劲动力。

基础学科作为科学研究与技术进步的根基，在支撑应用研究、构建理论体系及提供方法论框架方面发挥着不可替代的作用。近年来，随着大模型技术在生命科学、材料科学、数学、化学、物理学与地球科学等典型基础学科中的深入应用，原有的研究范式、问题域界限及知识结构正经历深刻重塑[6]。不同于传统以“基础学科如何驱动大模型发展”为核心的问题域，本研究聚焦大模型在基础学科中的应用及其所引发的学科交叉融合现象，识别热点主题。

学科交叉前沿识别已成为科技战略情报的重要环节，在推动关键领域科研创新方面发挥着全局性、前瞻性和战略性支撑作用[7]。在科学智能（AI4S）背景下，大模型的应用范围广泛且影响深远，有效推动了学科交叉研究的发展[8]。因此，本文通过系统梳理大模型主题相关文献，构建基础学科交叉主题分析框架，识别大模型在基础学科应用中所引发的学科交叉融合趋势与研究热点，旨在为科学智能（AI4S）发展趋势研判及学科交叉协同创新提供理论支持与实践指引。

1　相关研究

1.1　大模型相关研究

国内外关于大模型的研究，一方面关注大模型在科学智能（AI4S）中的关键作用，另一方面则聚焦于大模型在各领域中的实际应用。大模型正成为科学智能（AI4S）发展的关键引擎，重构科研基础设施，推动知识自动化。李国杰[9]提出智能化科研是科研方法的重大变革，不仅要关注科学智能（AI4S），更要关注大语言模型（LLMs）。苏新宁等[10]提出基于大模型的大规模研究平台正成为不可或缺的基础设施。苏莉雯等[11]提出将大语言模型构建为知识工厂，并探索面向科学家的知识自动化服务方式，有望成为推动高效科学智能（AI4S）的关键支撑力量。此外，大模型在教育研究、数字学术服务与科研流程等领域的应用日益深化，展现出推动学术创新的新潜力。刘泽民等[12]聚焦大语言模型在教育研究中的应用，基于“问题—方法—过程”框架，系统分析其在推动研究范式转型中的作用、潜在风险与应对策略。孙坦等[13]提出图书馆数字学术服务正迎来平台化重构机遇，基于大语言模型与人工智能生成内容（AIGC）的融合应用，图书馆可通过自主建设、第三方接入与嵌入式升级三种路径，打造契合科研需求的智能化服务体系。Filimonov[14]讨论了大语言模型在加速科研进程中的作用，特别是在快速处理科研文献、实验数据分析及生成研究假设中的应用。Tranchero 等[15]提出了“生成性 AI 实验”框架，认为大模型能够帮助学者快速验证理论，尤其是在不确定性高的环境下，对战略决策过程进行模拟和优化。

1.2　学科交叉相关研究

“学科交叉”（有的文献称为“跨学科”，英文术语为 interdisciplinary）概念最早由美国哥伦比亚大学心理学家 Woodworth 提出，指的是跨越单一学科界限，融合两个及以上学科的实践活

动［16］。学科交叉研究强调不同学科之间的整合、交叉与协作，旨在通过融合多领域的知识与方法，探索并解决复杂问题［17］。知识挖掘研究特别是主题识别研究，是从事跨学科研究的主要目的［18］。

学科交叉主题识别研究主要采用引文分析和内容分析两种方法。引文分析通过文献的引用关系揭示知识流动模式，侧重研究外部结构特征。Chi 等［19］通过共被引网络分析跨文化关系研究的知识结构，识别核心文献与研究主题的演化趋势，并构建该领域的知识框架。张艺蔓等［20］提出了基于引文耦合强度来量化学科交叉度的分析方法，发现与情报学联系最为紧密的三个学科为计算机科学、经济学及新闻与传媒学。然而引文分析方法存在无法直接识别交叉学科主题及结果滞后的问题，内容分析法则通过直接计量和分析文献主题，更注重挖掘文献内部特征。邓君等［21］从学科与主题两大基线出发，采用网络科学的方法对数字人文领域中的学科交叉状况进行分析，采用 BERTopic 主题模型进行主题识别。Wang 等［22］提出了一个基于 BERTopic 的跨学科主题识别与演变分析框架，并用图书馆与信息科学领域的数据集进行了实证研究。齐世杰等［23］提出了一种结合学科多样性和学科凝聚性的论文学科交叉性计算方法，并在作物智能育种领域进行了实证分析。

2 研究设计与方法

2.1 研究设计

面对重大复杂问题，学科交叉融合是实现突破的关键［24］。本研究关注为大模型技术发展提供支撑的基础学科，聚焦大模型在基础学科的应用及其引起的学科交叉现象，分析学科交叉特征，识别热点主题。从数据采集与预处理、基础学科交叉网络构建、主题识别和热点主题识别四个环节，构建研究框架（图 1）。

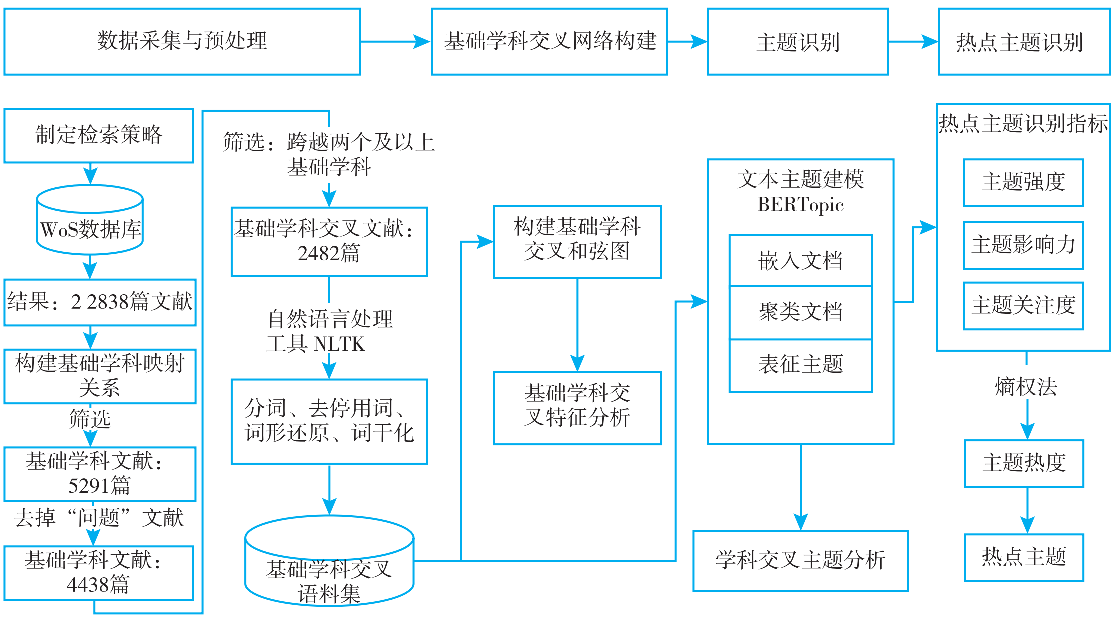

图 1　研究框架

2.2　数据采集与预处理

将 Web of Science（WoS）核心合集作为数据源，结合 Large Language Models、LLMs 等制定检索策略。为确保检索结果的全面性，采用的检索式为“TS=（large language model*）OR TS=（LLM*）”，利用通配符“*”以覆盖“large language model”的多种变体（如 models、modeling 等）及“LLM”相关缩写的扩展形式（如 LLMs、LLM-based 等），文档类型仅限于研究论文，共检索到 2 2838 篇文献。

在学科选择上，以每篇文献在 WoS 数据库中的“Research Area”信息作为学科分类依据[21，25]，参考已有研究成果[26]，确定研究所涉及的基础学科为生命科学、材料科学、数学、化学、物理学和地球科学六类。考虑到 WoS 中的“Research Area”分类和基础学科分类之间的差异，将两者之间做了人工映射。WoS 核心合集中关于大模型主题的研究论文的学科信息共涉及 147 个“Research Area”分类，从中提取与六类基础学科相关的分类共 38 个，构建学科映射关系（图 2）。

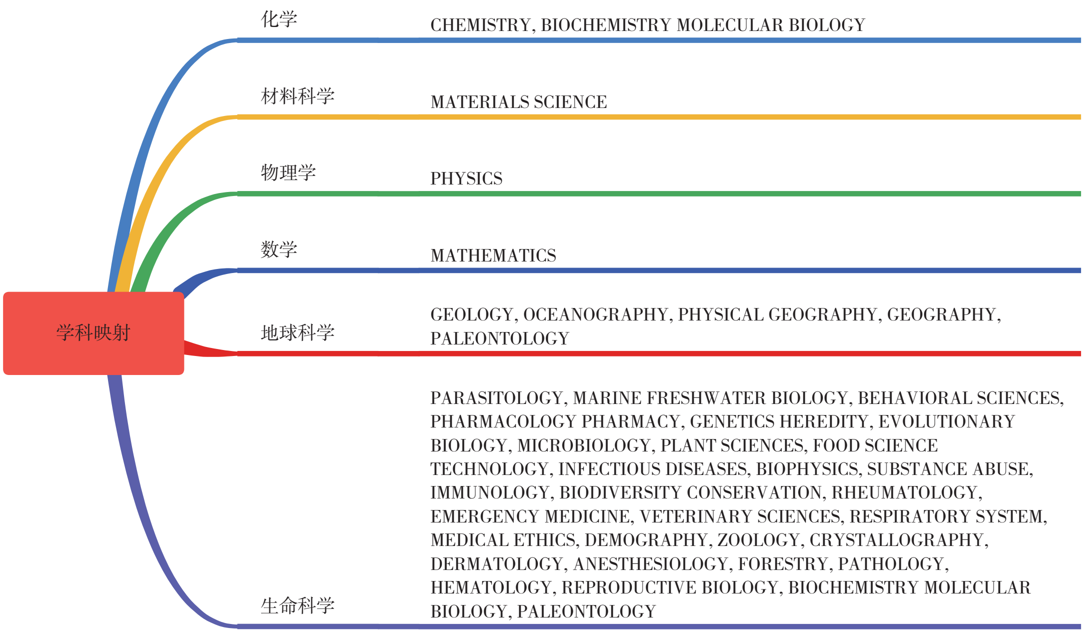

图 2　学科映射关系

根据上述学科映射关系，在 2 2838 篇文献中，匹配到上述六类基础学科文献共 5291 篇。采用手动检查方法从以下两方面进行文献验证，一是排除撤稿、重复发表、作者信息缺失、摘要信息缺失、非研究性质（如研究综述、专题序言、新闻报道、书评、资讯、倡议等）的文献，二是检查文献标题、摘要和关键词，判断是否符合研究主题。经检验，去除不合要求的 853 篇文献后，基础学科交叉文献共计 4438 篇。从中选择跨越两个及以上基础学科的 2482 篇文献为基础学科交叉语料集文献。

本研究对主题建模的语料集文献进行了数据清洗。首先，语料集由文献标题、关键词和摘要合并生成，剔除缺失关键信息、存在格式异常或重复记录的文献数据，保证数据完整性与一致性。其次，采用自然语言处理工具 NLTK（Natural Language Toolkit）对语料集进行分词处理，并

结合 NLTK 提供的英文停用词表去除频繁出现但语义承载能力较弱的通用功能词。为进一步统一词汇表达、去除语义冗余，引入词形还原与词干提取技术，对不同形式的词项进行归一化处理。最后，构建标准化语料集为主题建模提供数据基础。

2.3 基础学科交叉网络构建

借助多值邻接矩阵来呈现学科间的交叉关系[27]。若一篇论文涉及多个学科，便反映它是多学科交叉融合的成果。当某篇论文同时属于学科 i 和学科 j，则 A_{ij} 取值为 1。若有 w 篇论文同时属于学科 i 和学科 j，则 A_{ij} 取值为 w_{ij}。在此基础上，构建学科交叉加权网络 G[28]。其中，N 是节点集合，E 是边集合，W 是权重集合。k_i 代表学科 i 与多少个其他学科存在交叉关系，计算方法如公式（1）所示：

$$k_i = \sum_{j=1}^{N} A_{ij} + \sum_{j=1}^{N} A_{ji} \tag{1}$$

式（1）中，k_i 的值越大，意味着学科 i 与其他学科的交叉关系越多。权重 w_{ij} 的值越大，表明学科 i 与学科 j 之间的交叉融合程度越高。本研究将综合运用上述指标，深入剖析大模型主题论文的基础学科交叉情况。

2.4 主题识别方法

采用 BERTopic[29-30] 主题模型对基础学科交叉语料集的文本进行主题识别。与经典 LDA 主题模型相比，BERTopic 能自动确定主题数量，且在词间关系和语义信息处理上优于 LDA，更适用于科技论文的主题建模[31-32]。这种方法包含三个核心阶段。首先，文档嵌入阶段利用 BERT 模型将文本转换为词向量。其次，文档聚类阶段使用 UMAP 技术对高维嵌入数据执行降维，并借助 HDBSCAN 算法划分出语义相近的文档集群。最后，主题表征阶段应用 c-TF-IDF 方法优化 TF-IDF 技术，以评估各类别中词汇的权重，并利用最大边际相关性策略挑选出最能代表各主题的核心词汇。

采用的 BERTopic 主题建模的参数设置如下。第一，文本嵌入模型选用“all-MiniLM-L6-v2”[33]，其是被广泛应用的英文文本嵌入模型[34-35]。第二，初始化 UMAP，将投影后的维数（n_components）设置为 2，即通过 UMAP 算法将高维文本嵌入表示降维至二维空间[36]。评估点与点之间距离用余弦相似性度量。将最小距离参数（min_dist）从默认的 0.1 降低至 0.05，以实现更紧密的嵌入效果[37]。第三，聚类方式用 HDBSCAN[38]，最小聚类规模（min_cluster_size）设定为 10。同时，将最小样本数（min_samples）设定为 15。第四，主题提取，将主题数量（nr_topics）设置为“auto”模式，BERTopic 模型会根据聚类结果自动调整并生成适合的主题数量[22]。

2.5 热点主题识别方法

综合运用主题强度、主题影响力和主题关注度三个指标，借助熵权法客观确定各指标权重[39]，并加权得到热点主题识别的综合指标，据此精准识别热点主题。

主题强度反映了某一研究主题在文档集合中的重要性，通过该主题文献数量占总文献数量的比例来衡量[40]。计算方法如公式（2）所示：

$$I_j = \frac{n_j}{m} \tag{2}$$

式（2）中，I_j 表示第 j 个主题的主题强度，m 为总文献数量，n_j 为第 j 个主题的文献数量。

学术文献的被引次数是衡量其科学影响力的关键指标，而文献下载量则能及时反映其受关注的程度[7]。为避免主题文献数量对衡量结果的干扰，采用主题的平均被引量和平均下载量来评估其影响力和关注度。主题影响力通过该主题文献的总被引量除以其文献数量的结果来衡量。计算方法如公式（3）所示：

$$C_j = \frac{\sum_{i=1}^{n_j} c_{ij}}{n_j} \tag{3}$$

式（3）中，C_j 表示第 j 个主题的主题影响力，c_{ij} 表示第 j 个主题中第 i 篇文献的被引量，n_j 为第 j 个主题的文献数量。

主题关注度通过研究主题下所有文献的总下载量除以该主题文献数量的结果来衡量。计算方法如公式（4）所示：

$$D_j = \frac{\sum_{i=1}^{n_j} d_{ij}}{n_j} \tag{4}$$

式（4）中，D_j 表示第 j 个主题的主题关注度，d_{ij} 表示第 j 个主题的第 i 篇文献的下载量，n_j 为第 j 个主题的文献数量。

运用上述方法，可计算出第 j 个主题的主题强度 I_j、主题影响力 C_j 和主题关注度 D_j。随后，借助熵权法确定主题强度、主题影响力和主题关注度的权重 w_I、w_C、w_D，并加权得到热点主题识别的综合指标——主题热度 H_j，计算方法如公式（5）所示：

$$H_j = w_I \times I_j + w_C \times C_j + w_D \times D_j \tag{5}$$

依据综合指标主题热度 H_j 来识别各个阶段的热点主题，将综合指标值高于平均值的主题认定为热点主题。

3　研究分析

3.1　基础学科交叉特征分析

根据基础学科在文献中的共现关系，构建一个多值邻接矩阵，据此生成无向加权基础学科交叉网络，通过和弦图展示基础学科交叉态势（图 3）。

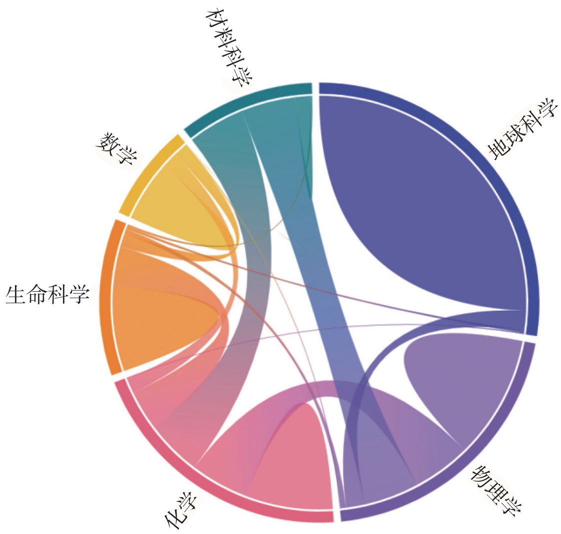

图 3　基础学科交叉和弦图

在图 3 中，每段圆弧代表大模型主题研究的一个基础学科方向，圆弧长度反映该基础学科方向的文献数量，圆弧越长，文献数量越多。圆弧间的连接弦表示基础学科间的交叉关系，连接弦的宽度代表两个基础学科共有的文献数量，弦越宽，基础学科交叉程度越显著。从各基础学科的连接弦宽度来看，化学与物理学、材料科学与物理学、材料科学与化学之间的交叉连接最为显著，位列前三，反映出化学、物理学与材料科学三类基础学科在大模型主题研究中的协同度较高，彼此之间形成了紧密的交叉融合关系。这种密切关联源于三类学科在研究对象、方法体系和应用场景上的高度契合，特别是在材料设计、分子建模与性能预测等前沿交叉领域，大模型在其中起到了促进知识整合与方法迁移的桥梁作用。此外，生命科学与化学的交叉也较为密切。相较之下，地球科学和数学与其他基础学科的交叉较少，显示其在当前研究中的相对独立性。这些基础学科交叉网络的可视化程度，揭示了大模型主题研究的基础学科融合态势，为学科交叉研究提供重要的参考。

3.2　基础学科交叉主题分析

运用 BERTopic 模型对语料集进行主题建模，共识别出八个主题。为深入了解各主题的内容及不同研究主题的偏好，分别提取八个主题中最具代表性的前 10 个高分主题词（图 4）。

在图 4 中，主题序号按照各主题包含文献数量多少进行降序排列，其中“蛋白质序列预测与基因分析”主题的文献数量最多，而“隐私保护与网络检测”主题的文献数量最少。八个主题分别涵盖了多个基础学科交叉领域，包括生物医学、药物开发、材料设计等。每个主题的研究不仅凸显了大模型在各基础学科中的广泛应用，还展示了深度学习等技术在基础学科研究中的重要作用。其中，“蛋白质序列预测与基因分析”主题以其包含的文献数量之多而最为显著。该主题的高分主题词包括“protein”“sequence”“gene”“analysis”“methods”“prediction”“results”“biological”“genome”“molecular”，研究聚焦于结合大模型技术提高蛋白质序列预测的准确性，以深入理解基因表达的机制及与各种生物过程和疾病的关联。此外，“气候变化与海洋研究”“材料能量特性与热分解”“材料设计与化学合成”“地震与火星地质研究”“疾病治疗与风险分析”“量子自旋与相变理

论”“隐私保护与网络检测”七个研究主题，均体现了大模型在推动科学发现、加速技术创新中的重要作用。

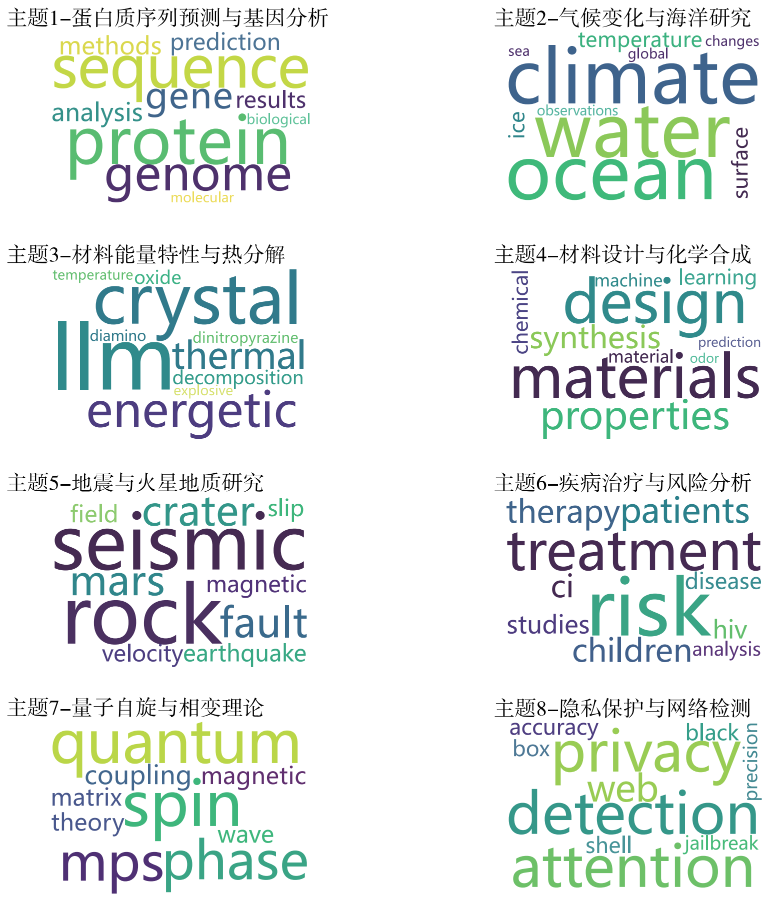

图4　八个研究主题的词云图

本研究进一步考察了八个研究主题与六类基础学科的关系。若某篇文献既属于基础学科 d，又属于研究主题 t，则在基础学科 d 与研究主题 t 之间记为权重为 1 的关联关系。若 n 篇文献同时属于基础学科 d 和研究主题 t，则其关联权重记为 n。在 BERTopic 主题建模中，每个研究主题会以一定概率分配不同文献，即一篇文献可能与多个研究主题相关，因此本研究将每篇文献分配到概率最高的研究主题中，从而得到基础学科与研究主题的关联关系（图 5）。

图 5 呈现了基础学科与研究主题的关联关系。左侧为基础学科，右侧为研究主题。端点位置反映文献数量，位置越高，文献数量越多。连线宽度则表示基础学科与研究主题共有的文献数量，宽度越大，文献重合度越高。大模型在多个基础学科的应用研究逐渐得到广泛关注，呈现出基础学科交叉融合的趋势。不同研究主题涉及的基础学科各有侧重，并在大模型应用的推动下实现了理论与应用的紧密结合。

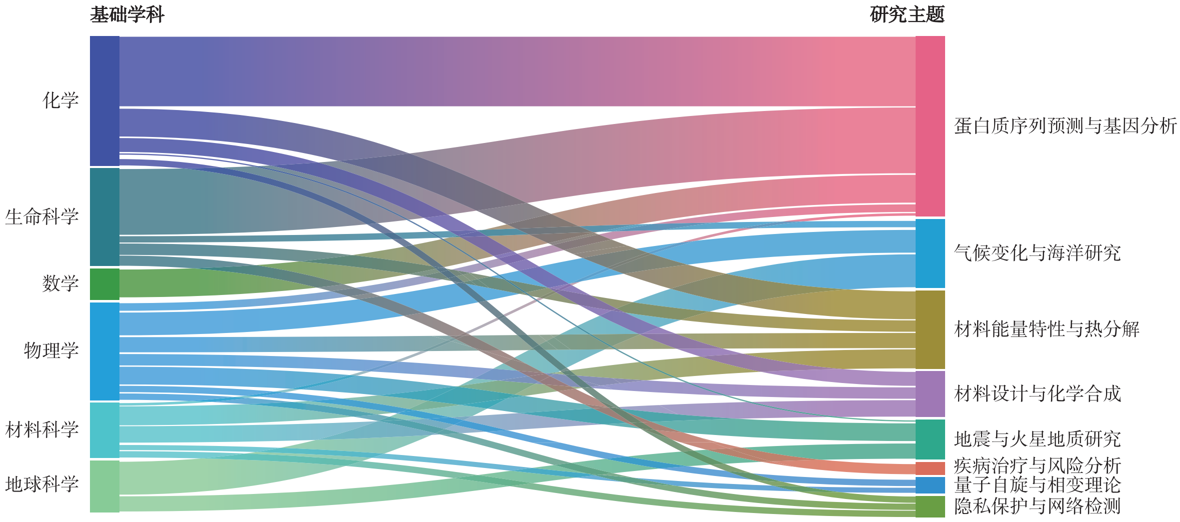

图 5　基础学科与研究主题关联关系

在“蛋白质序列预测与基因分析”主题中，化学、生命科学和数学作为主要的基础学科，分别从分子结构、基因功能与数学建模的角度提供了理论支持。蛋白质序列的准确预测依赖于数学统计模型的优化，而基因编辑技术通过化学反应机制和生命科学对基因功能的理解，推动了精准医疗的发展。

“气候变化与海洋研究”主题涉及地球科学、物理学和生命科学。地球科学为气候变化提供了历史演变与全球背景的基础。物理学通过大气物理和辐射传输模型深化了对气候变化的理解。生命科学则从生态学角度探讨气候变化对海洋生态系统及生物多样性的影响。

在“材料能量特性与热分解”主题方面，化学、材料科学、物理学和生命科学是主要的基础学科。化学为热分解反应机制提供了理论框架。材料科学深入研究材料在高温条件下的稳定性与能量转化特性。物理学通过热力学和量子力学的原理揭示了材料的微观热行为。而生命科学则关注生物材料的降解过程及其生态影响。

“材料设计与化学合成”主题主要依赖于化学、材料科学和物理学。化学为新材料的合成与反应机制提供了基础。材料科学通过分子设计与性能优化推动新型材料的开发。物理学则通过量子物理和固态物理的研究，深入分析材料的微观结构与性质，为材料创新提供了理论支持。

“地震与火星地质研究”主题涉及地球科学和物理学。在地震研究方面，地球科学通过探测地球内部结构与分析震波传播机制，深化了对地震发生及其过程的理解；物理学为震波的传播和能量传递，提供了理论框架和实验支持。火星地质研究通过类地地质过程的比较，为行星地质演化提供了新的视角与研究方法。

在“疾病治疗与风险分析”主题中，生命科学和化学发挥了重要作用。生命科学通过对疾病生物学机制的深入研究，为疾病预防与治疗提供了理论支持。而化学通过生物标志物识别与药物合成，推动了疾病诊断与治疗的发展。

“量子自旋与相变理论”主题涉及物理学和材料科学。物理学为量子自旋和相变提供了量子力学的基础理论支持。材料科学通过设计新型量子材料，控制物质的相变行为，推动了量子技术

的发展与应用。

“隐私保护与网络检测”主题涉及化学、材料科学和物理学。在隐私保护方面，化学与材料科学为加密技术提供了材料保障，尤其是在量子加密技术的应用中起到了关键作用。物理学通过信号处理与量子通信，提升了网络安全与隐私保护水平。

综上所述，大模型技术在各个领域的广泛应用，不仅推动了基础学科的交叉与融合，也促进了基础学科研究从理论到实践的不断发展。这些基础学科之间的紧密合作，为大模型的进一步发展提供了机遇与挑战。

3.3　基础学科交叉热点主题识别

根据前文热点主题识别方法，先依次计算各研究主题的主题强度、主题影响力和主题关注度，再根据熵权法计算各项的权重分别为 w_I=0.405、w_C=0.399、w_D=0.1962，加权计算出各研究主题的主题热度（表 1）。

表 1　八个研究主题的主题热度

主题序号	研究主题	主题强度	主题影响力	主题关注度	主题热度
1	蛋白质序列预测与基因分析	0.364	54.079	13.453	24.361
2	气候变化与海洋研究	0.170	15.292	17.569	9.617
3	材料能量特性与热分解	0.160	15.574	45.213	15.150
4	材料设计与化学合成	0.092	20.714	42.971	16.733
5	地震与火星地质研究	0.089	11.176	20.382	8.494
6	疾病治疗与风险分析	0.052	32.800	8.950	14.862
7	量子自旋与相变理论	0.039	67.267	24.667	31.691
8	隐私保护与网络检测	0.034	3.000	14.077	3.973

八个主题的主题热度平均值为 15.610。本文将综合指标值高于平均值的主题判定为热点研究主题，识别出三个热点主题，分别为“量子自旋与相变理论”（主题热度 31.691）、“蛋白质序列预测与基因编辑”（主题热度 24.361）和“材料设计与化学合成”（主题热度 16.733）。

近年来，大模型技术特别是深度学习（DL）、生成对抗网络（GAN）、量子神经网络（QNN）和量子变分算法（VQE），在“量子自旋与相变理论”主题中的应用取得了显著进展。深度学习被广泛用于量子自旋系统的相变识别，能够从大规模数据中有效识别不同量子相态和相变点[41]。生成对抗网络在量子态重建和相变检测中表现突出，能在噪声较大的实验数据中生成精确量子态并揭示相变特征[42]。量子神经网络结合量子计算和神经网络优势，为解决高维量子多体问题提供了新方案，特别是在量子自旋系统的能量优化和相变模拟中取得突破[43]。量子变分算法成功模拟了复杂量子自旋链的相变行为，提升了计算精度[44]。深度生成模型，如变分自编码器（VAE），在量子临界现象的研究中表现出色，能够自动识别量子系统中的相变特征，并在非平衡系统中进行有效预测[45]。

大模型技术也为“蛋白质序列预测和基因分析”主题提供了新的工具和方法。2023 年，AlphaFold2 在蛋白质折叠问题上取得了新的突破，尤其是在跨物种蛋白质结构预测和蛋白质复

合物预测方面，精度接近实验数据[46]。2024 年，谷歌 DeepMind 发布了 AlphaFold3 模型，该模型能够高精度预测蛋白质、DNA、RNA 及配体等生命分子的结构及其相互作用[47]。此外，研究者结合图神经网络和深度学习，开发了新的蛋白质—蛋白质相互作用预测工具，揭示了细胞内复杂的蛋白质网络[48]。在基因分析领域，CRISPR 基因编辑技术的精度不断提高。2023 年，新的 CRISPR 变种优化了靶向准确性和减少了脱靶效应，尤其在 RNA 靶向编辑中取得了突破性进展[49]。2024 年，深度学习与 CRISPR 的结合，使得 AI 能够优化编辑靶点设计，预测编辑效果，从而提高了基因治疗的成功率和安全性[50]。

大模型技术还推动了“材料科学与化学合成”研究的创新。在化学合成方面，浙江大学莫一鸣团队构建了由大语言模型 GPT-4 驱动的反应开发框架（LLM-RDF），通过六个自主开发的智能体完成化学合成开发流程中的关键任务，展现了极强的自主研究与决策能力[51]。在新材料发现方面，机器学习算法结合材料基因组数据库和高通量计算，成功预测并合成出新型钙钛矿太阳能电池材料、高性能储能材料等[52]。在催化剂设计领域，AI 辅助催化剂设计优化了工业过程的能源效率，推动绿色化学发展[53]。在材料表征领域，深度学习算法在电子显微镜图像分析、X 射线衍射谱图解析等方面取得突破，实现了纳米尺度材料结构的快速表征与精确分析[54]。

值得注意的是，大模型技术尽管在上述三个主题展现出强大的学科交叉融合和创新潜力，但其在应用过程中仍存在诸多局限性，包括但不限于：大模型面临“幻觉”问题[55]；训练数据中固有的偏差会传导至预测结果，导致蛋白质结构或基因编辑靶点选择出现系统性偏差[56]；当前主流大模型缺乏实时更新能力，其知识库的时效性滞后于基础学科前沿发展，难以有效支撑快速演进的科研需求[57]。

4 结 语

以大模型为代表的新一代人工智能技术在科学研究的广泛应用，正推动新一轮的学科交叉融合。本研究基于 BERTopic 主题建模与网络科学方法，以 WoS 核心合集中跨越两个及以上基础学科的 2482 篇文献为研究样本，系统识别了大模型主题研究中的基础学科交叉特征与热点主题。一方面，构建了可量化的基础学科交叉分析框架，有助于理解不同基础学科间的融合机制；另一方面，提出了数据驱动的研究主题识别方法，为探索未来科研方向提供了方法论支持。

本研究仍存在一定局限性。首先，数据来源主要依赖 WoS 核心合集数据库，未能全面覆盖相关研究成果，未来可结合 Scopus 等多源数据提升覆盖度。其次，尽管 BERTopic 具有较强的语义分析能力，但会受文本预处理、参数设置等因素影响，未来可引入 BERT+TF-IDF、LDA2Vec 等方法优化主题建模效果。最后，当前研究主要聚焦物理学等六类基础学科，未来可进一步拓展至计算机科学、工程科学、医学等其他领域，进一步深化研究大模型在基础学科的应用及其引发的学科交叉融合。

【参考文献】

［1］Zhao A P, Li S, Cao Z, et al. AI for science: predicting infectious diseases［J］. Journal of Safety Science and Resilience, 2024, 5(2): 130–146.

［2］Nobel Prize. Nobel prizes 2024［EB/OL］.［2025–02–13］. https://www.nobelprize.org/all-nobel-prizes-2024/.

［3］智源研究院. 2025十大AI技术趋势［R/OL］.［2025–08–07］北京：智源研究院，2025.https://hub.baai.ac.cn/view/42526.

［4］Wang H, Fu T, Du Y, et al. Scientific discovery in the age of artificial intelligence［J］. Nature, 2023, 620(7972): 47–60.

［5］Miao Q, Wang F–Y. AI4S Based on Parallel Intelligence ［M/OL］// Artificial Intelligence for Science (AI4S) . Springer, Cham, 2024: 1–19［2025–02–13］. https://doi.org/10.1038/s41586-023-06221-2.

［6］李伦，刘梦迪. 人工智能驱动的科学研究范式革命：态势与未来［J］. 探索与争鸣，2024（10）：143–151，180.

［7］白如江，张亚辉，张玉洁，等. 基于引文——主题双重测度的交叉前沿识别研究［J］. 现代情报，2024，44（10）：27–40，63.

［8］Chai X, Zhang M, Tian H. AI for science: practice from Baidu paddlePaddle［C］//2024 Portland International Conference on Management of Engineering and Technology (PICMET), Portland, or, USA, 2024:1–12.

［9］李国杰. 智能时代呼唤新的科研方法［J］. 科技导报，2024，42（10）：40–45.

［10］苏新宁，吕先竞. 人工智能赋能人文社会科学研究方法变革［J］. 西华大学学报（哲学社会科学版），2025，44（1）：1–10，121.

［11］苏莉雯，吴杨. 生成式人工智能在口腔医学的应用潜力与挑战［J］. 口腔医学研究，2024，40（1）：11–17.

［12］刘泽民，陈向东. 人工智能科学（AI4S）引发的范式变革——大语言模型视角下教育研究的问题、方法与过程［J］. 远程教育杂志，2024，42（5）：23–34.

［13］孙坦，张智雄，周力虹，等. 人工智能驱动的第五科研范式（AI4S）变革与观察［J］. 农业图书情报学报，2023，35（10）：4–32.

［14］Filimonov V Y. Large language models and their role in modern scientific discoveries［J/OL］. Philosophical Problems of IT & Cyberspace, 2024, 25(1): 42–57［2025–08–07］. https://doi.org/10.17726/philIT.2024.1.3.

［15］Tranchero M, Brenninkmeijer C–F, Murugan A, et al. Theorizing with large language models［R/OL］.［2025–02–13］. National Bureau of Economic Research, 2024. https://www.matteotranchero.com/pdf/Tranchero%20et%20al_2024_LLMs.pdf.

［16］刘仲林. 交叉科学时代的交叉研究［J］. 科学学研究，1993，11（2）：9–16.

［17］Wu S, Lin M, Ji M, et al. Exploring core knowledge in interdisciplinary research: insights from topic modeling analysis［J/OL］. Applied Sciences, 2024, 14(21): 10054［2025–08–07］. https://doi.org/10.3390/app142110054.

［18］章成志，吴小兰. 跨学科研究综述［J］.情报学报，2017，36（5）：523–535.

［19］Chi R, Young J. The interdisciplinary structure of research on intercultural relations: a co–citation network analysis study［J/OL］. Scientometrics, 2013, 96: 147–171［2025–08–07］. https://doi.org/10.1007/s11192-012-0894-3.

［20］张艺蔓，李秀霞，韩牧哲. 基于引文耦合的情报学学科结构时序分析［J］. 情报杂志，2015，34（3）：100–106.

［21］邓君，滕玥. 数字人文中的学科交叉与主题偏好研究——基于WoS的跨维度分析［J］.图书情报工作，2024，68（13）：3–14.

［22］Wang Z, Chen J, Chen J, et al. Identifying interdisciplinary topics and their evolution based on BERTopic ［J/OL］. Scientometrics, 2024,129: 7359-7384［2025-08-07］. https://doi.org/10.1007/s11192-023-04776-5.

［23］齐世杰，串丽敏，赵静娟，等. 基于主题模型的领域新兴交叉主题识别研究——以作物智能育种为例［J］.数字图书馆论坛，2024，20（9）：38-47.

［24］霍朝光，王晓玉，但婷婷，等. 学科交叉视域下中美两国学者学科交叉性测度与对比分析——以Scopus数据库为例［J］. 情报资料工作，2025，46（2）：91-102.

［25］Vera-baceta MA, Thelwall M, Kousha K. Web of science and scopus language coverage［J/OL］. Scientometrics, 2019, 121(3): 1803-1813［2025-08-07］. https://doi.org/10.1007/s11192-019-03264-z.

［26］Xu Y, Liu X, Cao X, et al. Artificial intelligence: a powerful paradigm for scientific research［J/OL］. The Innovation, 2021, 2(4): 100179［2025-08-07］. https://doi.org/10.1016/j.xinn.2021.100179.

［27］王慧，梁兴柱，张绪，等. 基于邻接矩阵优化和负采样的图卷积推荐［J］. 计算机应用研究，2024，41（12）：3628-3633.

［28］Abasi R. Research on the knowledge graph of educational management science in China［D］. Shanghai: East China Normal University, 2021.

［29］Lee Y-G, Kim S. A comparative study on topic modeling of LDA, Top2Vec, and BERTopic Models using LIS Journals in WoS［J］. Journal of the Korean Society for Library Information Science, 2024, 58(1): 5-30.

［30］Uğuz S, Tülü Ç N. Topic modeling analysis in the field of large language models with BERTopic (2020-2024)［C］//2024 Innovations in intelligent Systems and Applications Conference (ASYU). Ankara, Turkiye, 2024: 1-6.

［31］陈明红，梁萱，陈海华. 主题模型在科技论文中的应用对比：以信息资源管理领域为例［J］. 图书馆论坛，2025，45（6）：104-113.

［32］Liu Y, Wan F. Unveiling temporal and spatial research trends in precision agriculture: a BERTopic text mining approach［J/OL］. Heliyon, 2024,10:e36808［2025-08-07］. https://doi.org/10.1016/j.heliyon.2024.e36808.

［33］Wang X. EAD: effortless anomalies detection, a deep learning based approach for detecting outliers in English textual data［J/OL］. PeerJ Computer Science, 2024, 10: e2479［2025-08-07］. https://doi.org/10.7717/peerj-cs.2479.

［34］Park Y, Shin Y. Adaptive Bi-encoder model selection and ensemble for text classification［J/OL］. Mathematics, 2024, 12(19): 3090［2025-08-07］. https://doi.org/10.3390/math12193090.

［35］S á nchez-franco M J, Calvo-mora A, Periáñez-cristobal R. Clustering abstracts from the literature on Quality Management (1980-2020)［J/OL］. Total Quality Management Business & Excellence, 2022, 34(7-8): 959-989［2025-08-07］. https://doi.org/10.1080/14783363.2022.2139674.

［36］任嵘嵘，项李梅，徐文哲，等. 基于BERTopic方法的新能源汽车技术机会发现［J］. 情报杂志，2025，44（5）：147-155.

［37］杨思洛，吴丽娟. 基于BERTopic模型的国外信息资源管理研究进展分析［J］. 情报理论与实践，2024，47（2）：189-197.

［38］AlqurshiL F, Ahmad I. A data-driven multi-perspective approach to cybersecurity knowledge discovery through topic modelling［J/OL］. Alexandria Engineering Journal, 2024, 107: 374-389［2025-08-07］. https://doi.org/10.1016/j.aej.2024.07.044.

［39］胡泽文，韩雅蓉，王梦雅. 基于LDA-Word2vec的图书情报领域机器学习研究主题演化与热点主题识别［J］. 现代情报，2024，44（4）：154-167.

［40］许菱，郑婧然，王耀刚，等. 基于LDA算法的关键共性技术识别研究——以智能纺织领域为例［J/OL］. 现代纺织技术：1-12［2025-07-29］. http://kns.cnki.net/kcms/detail/33.1249.TS.20250319.1552.012.html.

［41］Chung M, Huang G, Mcculloch I, et al. Deep learning of phase transitions for quantum spin chains from correlation aspects [J/OL]. Physical Review B, 2023,107(21): 214451 [2025-07-29]. https://doi.org/10.1103/PhysRevB.107.214451.

[42] Wu X D, Cong S. GAN-Based quantum state estimation [C]//2023 42nd Chinese Control Conference (CCC),Tianjin, China, 2023: 8165-8170.

[43] Hafeez M A, Munir A, Ullah H. H-QNN: a hybrid quantum-classical neural network for Improved Binary Image Classification [J/OL]. AI, 2024, 5(3): 1462-1481 [2025-07-29]. https://doi.org/10.3390/ai5030070.

[44] Tilly J, Chen H, Cao S, et al. The variational quantum eigensolver: a review of methods and best practices [J/OL].Physics Reports, 2022, 986: 1-128 [2025-07-29]. https://doi.org/10.48550/arXiv.2111.05176.

[45] Frohnert F, Van Nieuwenburg E. Explainable representation learning of small quantum states [J/OL]. Machine Learning: Science and Technology, 2024, 5(1): 015001 [2025-07-29]. https://doi.org/10.48550/arXiv.2306.05694.

[46] Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2 [J/OL]. Nature Communications, 2022, 13(1): 1265 [2025-07-29]. https://doi.org/10.1038/s41467-022-28865-w.

[47] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3 [J/OL].Nature, 2024, 630(8016): 493-500 [2025-07-29]. https://doi.org/10.1038/s41586-024-07487-w.

[48] Zhou Z, Yin Y, Han H, et al. ProAffinity-GNN: a novel approach to Structure-Based Protein-Protein binding affinity prediction via a curated data set and graph neural networks [J/OL]. Journal of Chemical Information Modeling, 2024, 64(23): 8796-8808 [2025-07-29]. https://doi.org/10.1021/acs.jcim.4c01850.

[49] Yang H, Patel D J. Structures, mechanisms and applications of RNA-centric CRISPR-Cas13 [J/OL]. Nature Chemical Biology, 2024, 20(6): 673-688 [2025-07-29]. https://doi.org/10.1038/s41589-024-01593-6.

[50] Villiger L, Joung J, Koblan L, et al. Crispr technologies for genome, epigenome and transcriptome editing [J/OL].Nature Reviews Molecular Cell Biology, 2024, 25(6): 464-487 [2025-07-29]. https://doi.org/10.1038/s41580-023-00697-6.

[51] Ruan Y, Lu C, Xu N, et al. An automatic end-to-end chemical synthesis development platform powered by large language models [J/OL]. Nature Communications, 2024, 15(1): 10160 [2025-07-29]. https://doi.org/10.1038/s41467-024-54457-x.

[52] Mohammed H, Mia M F, Wiggins J, et al. Nanomaterials for energy storage systems-a review [J/OL]. Molecules, 2025, 30(4): 883 [2025-07-29]. https://doi.org/10.3390/molecules30040883.

[53] Xu Y, Wang H, Zhang W, et al. AI-Empowered catalyst discovery: a survey from classical machine learning approaches to large language models [J/OL]. ArXiv,2025 [2025-07-29]. https://doi.org/10.48550/arXiv.2502.13626.

[54] Yu J, Wang Z, Saksena A, et al. 3D deep learning for enhanced atom probe tomography analysis of nanoscale microstructures [J/OL].Acta Materialia, 2024, 278: 120280 [2025-07-29]. https://doi.org/10.1016/j.actamat.2024.120280.

[55] Sanu E, Amudaa T, Bhat P, et al. Limitations of Large Language Models [C]//2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2024: 1-6.

[56] Ruan W, Lyu Y, Zhang J, et al. Large language models for bioinformatics [J/OL]. ArXiv, 2025 [2025-07-29]. https://arxiv.org/abs/2501.06271.

[57] Zhang N, Yao Y, Tian B, et al. A comprehensive study of knowledge editing for large language models [J/OL]. ArXiv, 2024 [2025-07-29]. https://doi.org/10.48550/arXiv.2401.01286.

王方媛，徐慧婷 . 基础学科交叉特征及热点主题识别——以 Web of Science 大语言模型主题论文为例［J］.
文献与数据学报，2025，7（3）：049-063.

Identification of Interdisciplinary Characteristics and Hot Topics of Fundamental Disciplines: Take Papers of Large Language Models Theme in Web of Science as an Example

Wang Fangyuan　Xu Huiting

(Shanghai Library, Shanghai 200031, China)

Abstract:［**Purpose/Significance**］This paper focuses on the application of Large Language Models (LLMs) in fundamental disciplines and the interdisciplinary phenomena by them, providing theoretical support and practical guidance for the promotion of interdisciplinary integration in fundamental disciplines by artificial intelligence technologies represented by LLMs.［**Method/Process**］Six fundamental disciplines such as life sciences, materials science, mathematics, chemistry, physics and earth sciences are choosed from the papers on LLMs topics in the core collection of Web of Science(WoS).Then, papers in two or more fundamental disciplines are selected as the research objects to construct an interdisciplinary topic analysis framework. The BERTopic topic modeling method is adopted and combined with the multi-valued adjacency matrix to build the interdisciplinary network to analysis interdisciplinary characteristics.Based on the indicators of topic intensity, influence and attention, the entropy weight method is adopted to calculate the comprehensive topic heat scores to identify hot topics.［**Result/Conclusion**］Six Fundamental disciplines have a high degree of interdisciplinary nature,with the most significant cross-disciplinary integration occurring among chemistry, physics, and materials science. The research identified eight topics such as “Protein Sequence Prediction and Gene Analysis” and three hot topics including “quantum spin and phase transition theory,” “protein sequence prediction and gene analysis,” and “materials design and chemical synthesis”.

Keywords: Large Language Models (LLMs); Foundational disciplines; BERTopic; Interdisciplinary; Topic model

（本文责编：魏　进）