·情报方法与技术创新·

基于大语言模型的科学问题自动生成研究

周 凝[1] 闵 超[1*] 范 涛[2] 刘雨萱[3] 张 雯[1,4] 袁勤俭[1]
（1. 南京大学信息管理学院，江苏 南京 210023；2. 南京财经大学公共管理学院，江苏 南京 210023；
3. 南京大学社会学院，江苏 南京 210023；4. 江苏省社会科学院，江苏 南京 210004）

摘 要：［目的/意义］科学问题是科学研究的起点，决定了科学研究的深度、广度及其影响。探索一种从海量的科技文献中自动生成科学问题的方法对提高科研选题效率具有重要意义。［方法/过程］本文提出了一种利用大语言模型从科技文献中自动生成科学问题的方法（AGMSQ）。首先，将科学问题划分为描述性、解释性、方法性、评价性和规范性五类；其次，根据科学问题的类型和结构，设计输入要素组合，由“未来工作句子”（FWS）中提取的关键要素三元组和领域扩展搜索主题构成；最后，利用参数微调的大语言模型 ChatGPT-4、ChatGPT-3.5、Claude3 Sonnet 和 Gemini Pro 根据输入要素组合生成科学问题。［结果/结论］利用自然语言处理领域的 FWS 数据集进行方法性问题的生成，根据专家评估的结果，模型生成的科学问题在清晰度、原创性、可行性、价值上均有良好的表现，其中 Claude3 Sonnet 生成效果最好。研究证明了大语言模型在科学问题生成方面的能力，为科学问题自动生成的研究提供了新思路。

关键词：科学问题；自动生成；大语言模型；AI for Science；自然语言处理
DOI：10.3969/j.issn.1008-0821.2026.03.001
〔中图分类号〕G255.51；TP18 〔文献标识码〕A 〔文章编号〕1008-0821（2026）03-0003-15

Automatic Generation of Research Questions Based on Large Language Models

Zhou Ning[1] Min Chao[1*] Fan Tao[2] Liu Yuxuan[3] Zhang Wen[1,4] Yuan Qinjian[1]
（1. School of Information Management，Nanjing University，Nanjing 210023，China；
2. School of Public Administration，Nanjing University of Finance and Economics，Nanjing 210023，China；
3. School of Social and Behavioral，Nanjing University，Nanjing 210023，China；
4. Jiangsu Academy of Social Sciences，Nanjing 210004，China）

Abstract：［Purpose/Significance］Scientific questions serve as the starting point of scientific inquiry，determining the depth，breadth，and impact of research endeavors. However，amidst the exponential growth of global scientific publications，identifying high-value research gaps from the vast volume of literature has become an overwhelming cognitive burden for researchers. Consequently，developing automated methodologies to generate research questions from large-scale literature is of critical importance.［Method/Process］To address this need，this paper proposed the Automatic Generation Method of Scientific Questions（AGMSQ），a novel framework leveraging Large Language Models（LLMs）. By tailoring the

收稿日期：2025-09-02
基金项目：江苏省社会科学基金重点课题“人工智能对科学研究范式的影响研究”（项目编号：24TQA002）；国家自然科学基金面上项目“交叉科学对科技突破的影响模式、作用机制与政策优化研究”（项目编号：72374099）；中央高校基本科研业务费揭榜挂帅应用先导项目“遥感与开源数据综合的信息获取与知识构建软件系统”（项目编号：14380009）。
作者简介：周凝（2001-），女，硕士研究生，研究方向：科技大数据挖掘、科技情报分析。范涛（1995-），男，讲师，博士，研究方向：自然语言处理。刘雨萱（2005-），女，本科生，研究方向：计算社会科学。张雯（1994-），女，助理研究员，博士后，研究方向：产业科技创新。袁勤俭（1969-），男，教授，博士，博士生导师，研究方向：信息行为与信息经济。
通信作者：闵超（1990-），男，副教授，博士生导师，研究方向：科技情报挖掘、科技创新政策。

generation process to specific question types，AGMSQ guided LLMs to produce high-quality research questions that were structurally rigorous and deeply grounded in the literature context. The method comprised three core modules：the Scientific Question Classification Module，the Generation Template Design Module，and the LLM Generation Module. First，the Classification Module categorized questions into five types：descriptive，explanatory，methodological，evaluative，and normative. This fine-grained taxonomy enabled the model to capture the distinct logical patterns and semantic requirements inherent to different modes of scientific inquiry，thereby enhancing the precision of generation. Second，the Template Design Module constructed element-generation templates based on the structural principles of each question type. It integrated key element triplets extracted from "Future Work Sentences"(FWS) with domain extension search topics，which were matched to the triplets via semantic distance. Finally，the LLM Generation Module utilized parameter-fine-tuned models—including ChatGPT-4，ChatGPT-3.5，Claude 3 Sonnet，and Gemini Pro—to synthesize research questions based on the combined input elements. Additionally，the study introduced two quantitative indicators—the Utilization Rate of Prompts(URP) and the Occupancy Rate of New Words(ORN)—to evaluate and optimize the generation performance of the LLMs.［Result/Conclusion］The experiments utilize an FWS dataset sourced from the natural language processing domain，specifically targeting the generation of methodological questions. Expert evaluations indicate that the research questions generated by AGMSQ demonstrate favorable performance in terms of clarity，originality，feasibility，and academic value. Notably，among the evaluated models，Claude 3 Sonnet exhibits the superior generation performance. Furthermore，quantitative analysis based on URP and ORN metrics corroborates the expert findings，confirming that the optimized prompts effectively reduce semantic redundancy and increase the efficient utilization of input information. These findings validate the capability of LLMs to generate methodological questions within the natural language processing domain，offering empirical evidence and valuable insights for future exploration across diverse disciplines and question types. Overall，this study offers new insights and tools for automating research topic selection，representing a concrete practice of the "AI for Science" paradigm.

Keywords：research question；automatic generation；large language models；AI for science；natural language processing

在科研领域，科学问题的质量直接影响着研究的方向和深度。科学问题的生成传统上依赖于研究人员通过阅读文献、参加学术交流等方式，基于个人研究经验进行主观凝练，但这一过程耗时长，且容易受到知识背景和认知范围的限制。尤其在当前科技文献数量急剧增长的背景下，探索一种能够从文献中有效生成科学问题的自动化方法，对于提高科研选题的质量和效率有重要意义。

随着大语言模型(Large Language Model，LLM)在科研领域的广泛应用，其已在生成科学假设、设计研究目标和提出潜在解决方案等方面展现出巨大潜力。在此背景下，科研问题自动生成的研究范式正逐渐从基于编码器—解码器架构的传统生成模型(如 UniLM[1])向大语言模型探索转变，但在过程中仍面临诸多挑战。首要问题在于，科学问题常被同质化处理，未能引导 LLM 根据其内在结构与类型特征实现针对性生成；其次，对科技文献中丰富的启发式提示词挖掘不足，未能将其深层语义与结构信息有效整合到 LLM 的生成策略设计中。

基于此，本文提出基于大语言模型的科学问题自动生成方法(Automatic Generation Method of Scientific Questions，AGMSQ)，旨在引导 LLM 针对问题的不同类型，生成结构清晰并紧密依托科技文献语境的高质量科学问题。与以往研究相比，其贡献主要体现在两个方面：一方面，在生成策略设计中充分考虑了科学问题的结构与类型特征，根据不同类型的科学问题有针对性地设计了不同的要素组合模板；另一方面，系统挖掘并利用了科技文献中丰富的语义信息与结构线索，以未来工作句子(Future Work Sentence，FWS)为原料，提取其语义和结构层面的特征作为关键生成要素，进而引导 LLM 进行科学问题的生成。经过专家评分，发现生成的科学问题在清晰度、原创性、可行性和价值方面均表现出色，表明该方法在自动生成科学问题方面的可行性和有效性。研究聚焦面向科研实践的科学问题的生成，为科研选题自动化提供了思

路和工具，也是AI for Science的具体研究实践。

1 相关研究

1.1 科学问题的定义与分类

科学问题最初的界定是现有理论与经验陈述间的逻辑矛盾[2]。这种定义方式抓住科学问题可证伪、逻辑与经验并重的特性，但忽略了科学研究是一个社会行为[3]，且具有情境性，即科学问题的产生落地需要利益相关者结合社会历史状况，协商判断是否需要科学介入，能否利用科学解决。综上，本文认为科学问题是被科学共同体在特定情境中识别为值得且可通过科学方法加以解决的认知缺口或矛盾，其同时嵌入学术范式、资源条件与价值取向，并可能随情境演化发生漂移与重构。

在科学问题的分类上，学界同样缺乏一致共识，但总的来说可以大致归为三类。逻辑元维度层面，Popper K[2]和Laudan L[4]基于认知论将科学问题划分为“经验性问题”和“概念性问题”。Cook T D等[5]从方法论出发，提出四类科学问题：描述、关联、因果、机制。情景维度层面，记忆性、理解性、应用性、分析性、评价性、创造性的分类标准适应了人的教育认知规律[6]；事实类、定义类、方法类、原因类、假设类的分类原则满足了自然语言处理的需求[7]。除此之外，科学各领域因其性质的不同也呈现各具学科特色的分类方法，如自然科学领域的“理论问题”“实验问题”分类法[8]、社会科学领域的“描述性问题”“解释性问题”“意释性问题”分类法[9]、工程技术领域的“设计问题”“优化问题”“控制问题”分类法[10]等。

1.2 科学问题自动生成研究

在科学问题自动生成方面，相关研究可以分为3个阶段。第一个阶段是利用编码器—解码器架构的传统生成模型，并结合人工构建的模板来生成科学问题，如宋若璇等[1]将FWS中提取的关键词和扩展主题词相结合组成模板，并采用UniLM生成了可供科学家参考的科学问题。该类系统较为依赖于人工模板，且受限于架构与数据规模。第二个阶段是趋势预测，即研究者试图提取文章关键内容(如摘要、关键词、未来工作等)还原各领域研究的演绎模式，推导研究空白。Wang Q Y等[11]提出研究助手PaperRobot，基于记忆—注意力网络增量生成标题、摘要和未来工作。Li L等[12]提出的创意链(CoI)将覆盖32个学科的多模态数据集结构化为一个链条，以确定有意义的方向。该方法有效解决了知识结构成熟领域的鲁棒性问题，但在NLP等新兴领域仍存在挑战。随着LLM的兴起和普及，其多模态融合和跨学科整合的优势极大改变了科学问题自动生成的范式。最直接的方法是通过提示词引导LLM一次性生成想法[13]。在此基础上，其他研究工作结合了更多元化的技术，以提升科学问题构思质量。Hu X等[14]提出了一种基于迭代规划与搜索的研究提案生成框架。Su H等[15]采用多智能体协作的专家团队来集体生成、评估和完善想法，并提出五阶段协作框架。Li R等[16]引入多维度奖励机制和动态控制器，提出基于可控强化学习的科学问题生成框架。然而，现有方法往往过于依赖数据规模，而忽略了文档中蕴含的本质性、多维度及结构化语义信息，这类信息在领域知识密集型任务中至关重要，能帮助高效、精准地定位高度相关的潜在方向。此外，大多数方法将科学问题视为同质对象，缺乏针对不同类型的差异化生成策略，限制了生成结果的有效性。

1.3 未来工作识别研究

“未来工作句子”中蕴含了科研线索与方向，具备作为生成科学问题的语料来源的潜力。近年来，其识别与利用方法经历了显著的演进过程。早期研究[17-18]主要以正则表达式和人工规则对FWS进行提取与分类，已有研究对文献中“未来工作”和“局限性”部分段落做FWS的提取和分类。随着技术进步，研究转向于利用机器学习与深度学习方法。Zhu Z H等[19]和Hao W K等[20]分别采用Bert预训练模型和构建标注语料库的方法以提升识别、分类与提取FWS的效果。Zhang C Z等[21]对比了BiLSTM、TextCNN、BERT和SciBERT在未来工作句子分类的准确性。谢林蕾等[22]则进一步结合SelectKBest特征选择，实现了融合出版领域论文中未来工作句子的自动识别。最新研究开始探索大语言模型的应用潜力，Azher I A等[23]提出基于检索增强生成(RAG)与大语言模型反馈的机制，以多轮迭代的方式生成高质量的未来工作描述。这一发展路径表明，该领域正从依赖人工规则向自动化、智能化的识别和分类发展。

2 研究方法

本研究围绕自动生成科学问题的核心问题，设计了基于大语言模型的科学问题自动生成的方法（AGMSQ）。该方法通过建立类型学框架与结构化生成模板，引导大语言模型生成兼具结构规范性和语义创新性的科学问题。如图 1 所示，该方法由科学问题分类层、生成模板设计层和 LLM 生成层 3 个核心的模块构成。

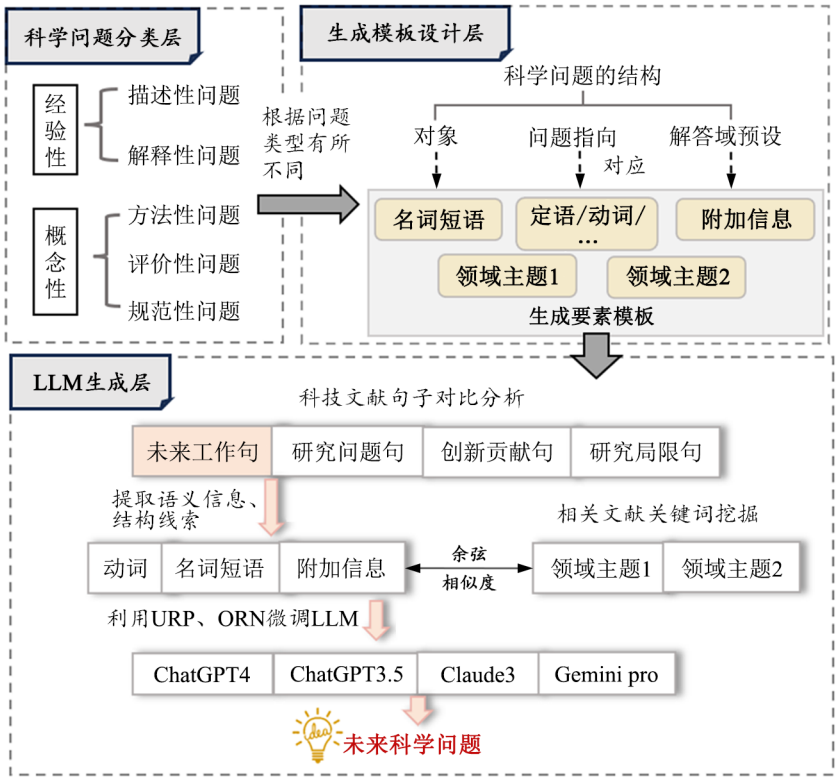

图 1 科学问题自动生成分析框架

Fig. 1 Analysis Framework of Automatic Generation of Scientific Questions

2.1 科学问题的分类

科学问题的分类是科学哲学与科学方法论的重要议题。根据 Popper K [2] 和 Laudan L [4] 的理论，科学问题有经验性问题与概念性问题的基本区分，强调科学理论的发展需要同时关注经验事实与理论结构的反思。本研究提出的科学问题分类框架以此为基础，结合社会科学与科学学中常见的问题类型，进一步将科学问题划分为描述性问题、解释性问题、方法性问题、评价性问题和规范性问题这五大类。

经验性问题关注科学对象的事实和规律本身，属于这一范畴的有描述性问题和解释性问题。描述性问题旨在回答“是什么”，即对事物的特征、状态或关系进行系统刻画 [24]，这类问题构成科学研究的经验基础。解释性问题则进一步探究事物内部或事物间的机制与规律，旨在回答“为什么会形成”或“背后的机制是什么”[24]，Fain H 等的“覆盖律模型”[25] 为此类问题提供了经典的理论范式。

概念性问题可划分为方法性、评价性与规范性问题，分别对应于科学活动的工具层面、批判层面和价值层面。方法性问题是关于如何构建或改进研究工具、方法及技术路径以获取知识的科学问题，其在科学哲学中的重要性已有深刻论述。Hacking I [26] 揭示了实验与干预相对于理论表征的独立地位。Chang H [27] 则通过“认知迭代”模型，论证了测量方法本身的演进构成科学进步的基本动力。评价性问题关注对现有理论、模型或方法的科学有效性进行系统分析或对比，靳玉乐 [28] 将评价性问题作为与知识型、理解型等并列的科学问题类型，认为评价性问题属于高级知识提问。规范性问题关注科学中“应当如何”的价值判断，涵盖研究伦理、技术责任及知识生产的社会规范等维度，其根植于“经验—规范”的哲学区分 [29]，将规范性问题独立划分，凸显了科学活动内在的伦理要求及实践的规范性导向。

总的来说，本文构建的分类框架旨在区分不同

性质的科学问题类型，从而为构建相应的科学问题生成策略提供理论依据。在实际科学研究中，科学问题往往具有相当的复杂性，可能同时涵盖多个问题类型，或难以被精确归类于某一特定范畴。因此，在应用本框架指导LLM生成科学问题或分析实际科研问题时，应结合领域背景与研究问题具体分析、灵活调整。

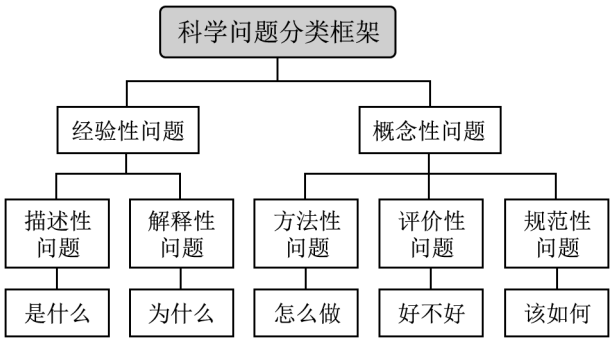

图 2 科学问题分类框架

Fig. 2 Classification Framework of Scientific Questions

2.2 不同问题类型的生成要素设计

生成高质量的科学问题需要规范、准确的内容要素作为输入，传统的生成科学问题或科学启发的提示词输入停留于主题或领域层面，如“基于你对[研究方向]最新进展的了解，请提出一个你认为当前最值得探索的科学问题。”这样的提示词模板信息密度低，缺少对不同问题类型的区分性，难以生成与文献研究直接相关且有价值的科学问题。

基于此，本文通过科学问题的结构原理，针对科学问题不同类型的特征，设计了生成要素模板，由“要素三元组”和“领域扩展搜索主题”结合构成，如图3所示。其中，要素三元组是从来源句子中识别得到的能够潜在表达科学问题对象、问题指向和解答域预设的句子成分；领域扩展搜索主题则来自领域相关的文献活跃主题，通过距离测算与要素三元组进行匹配。

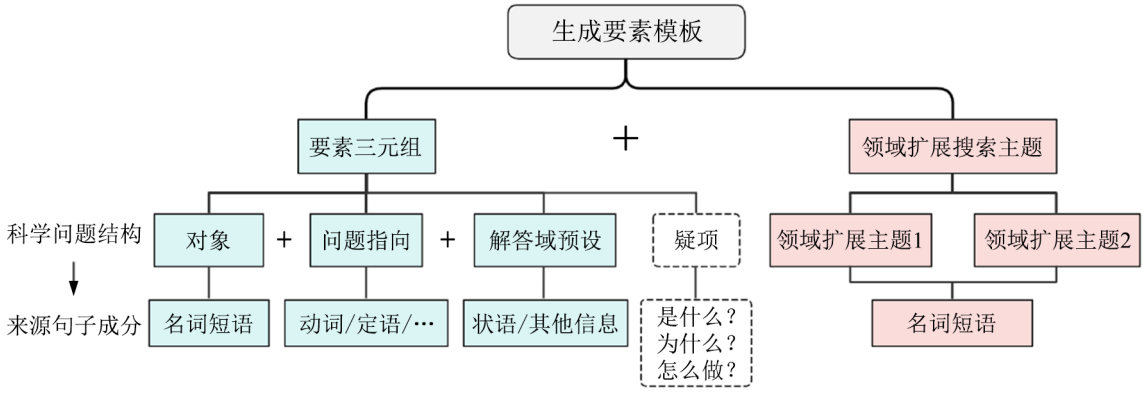

图 3 生成要素模板

Fig. 3 Template of Generation Elements

2.2.1 要素三元组设计

科学问题的结构一般可以分解为对象、问题指向、解答域预设和疑项[30]。研究对象是科学问题的基体；问题指向是研究对象的具体指向和方面，体现了问题求解的目标；解答域预设是科学问题自身所认定或假设的问题解答的范围界限；疑项则表示疑问，可用“疑问词+?”表示。为了强调生成问题的结构导向和简洁性，本文以生成的科学问题为陈述句，主要以对象、问题指向和解答域预设构成。

在语言学层面，这三类语义成分通常可以分别由句法成分承载[31]：名词短语表示论元与主题信息，适合作为潜在研究对象[32]；名词内部的定语或外部修饰用于定位对象的属性或方面，可以与问题指向相对应；状语或附加成分(时间、地点、条件、范围等)为命题设定语境与适用域，相当于解答域预设；动词或谓词编码事件类型、关系与操作性信息，可以指示因果、机制、方法或评价导向[33]。上述句法—语义映射是文本语篇分析中常用的要素抽取逻辑。将句子成分映射为科学问题要素，也为类型化生成策略提供了理论支撑与可操作的路径。

对于不同的问题类型，其研究对象和解答域预设所对应的句法成分是相似的。名词短语在句法上承担指称功能，在各个问题类型中都能潜在表示研究对象，附加信息/状语提供时空、条件等

信息，限定解答范围，对应解答域预设。

不同的是，各类问题在问题指向的侧重点上各有差异。描述性问题侧重于属性表征，由名词内部的定语或外部修饰来表示对象的属性或方面较为恰当。解释性问题以揭示因果或机制为目标，Levin B [33] 关于动词的研究表明，不同语义类动词在句法行为上具有可预测的分布特征，其中，因果、解释性动词可作为识别解释性句子的信号，将其与机制/因果关系结合起来可以潜在地表示问题指向，如“解释海洋酸化形成机制”中的“解释……机制”。

类似地，结合文献和对大量数据的分析发现，方法性问题中的操作或过程类动词(如设计、构建、测量、估计、实现、优化等)，在语料中常与直接宾语(名词短语)共现，能够用于识别方法性问题。评价性问题关注对理论、方法或系统的某一属性做批判或比较，对于批判意图的问题，需要识别出其批判的对象，也就是名词短语内部的定语或外部修饰；对于比较意图的问题，其一般具有较稳定的结构，如“比较A与B在某方面的差异”，采用比较词+多个名词短语可以潜在表示。规范性问题往往结合事实与价值，其生成需要在抽取事实性线索(前四类模板)之外，识别出情态或价值表达的词作为问题的目标与约束，因此没有设计统一的模板。总结这五类科学问题的生成策略如表1所示。

表1 不同问题类型及生成策略

Tab. 1 Question Types and Generation Strategies

问题类型	对象	问题指向	解答域预设
描述性问题	名词短语	描述性定语	状语(附加信息)
解释性问题	名词短语	动词+机制/因果关系	状语(附加信息)
方法性问题	名词短语	动词	状语(附加信息)
评价性问题	名词短语	定语或比较词	状语(附加信息)
规范性问题		没有统一的模板	

需要明确的是，以上针对五类科学问题所归纳的生成要素组合，其目的并非囊括识别该类科学问题的所有句法组合形式，事实上，关键句中对于科学问题的实际表达会因其研究领域、具体语境和学者写作风格而呈现出显著的多样性。在本文提取潜在句法生成要素并进行结构化生成的场景中，这些要素组合具有较高的典型性和可操作性，能够为科学问题的自动生成提供有效的指导。

2.2.2 领域扩展搜索主题

通过初步的实验发现，仅基于关键要素三元组生成科学问题，虽能确保语义的完整性和基本准确性，但其多样性和创新性往往受限，这主要是因为三元组结构高度聚焦于核心要素，难以引入更多的相关领域知识、新兴概念或边缘关联主题，而这些正是驱动科学创新突破的潜在因素。因此，为了提升生成问题的信息丰富度，并激发其创新潜力，对研究对象(名词短语)进行领域关联性扩展搜索。

扩展搜索的具体方法如下：获取相关领域论文的标题和摘要，并采用快速自动关键词提取算法(Rapid Automatic Keyword Extraction，RAKE)从摘要文本中提取领域活跃主题。采用余弦相似度的方法计算关键名词短语和领域活跃主题之间的距离并排序。为选择最合适的两个领域扩展主题，采用数据驱动阈值选择的方法，基于对实验数据多轮的相似度计算和对结果的观察发现，余弦相似度高于0.6的主题与关键名词短语重合度较高，普遍难以提供新的语义信息；低于0.4的主题则偏离核心语义太多，相关性有所缺失。余弦相似度介于0.5~0.55的主题在有区分度的同时，与关键名词短语还保持着较高的关联，是合适的相关性主题集合；介于0.4~0.45的主题在合理控制语义距离的前提下，能够显著引入创新性内容，是合适的创新性主题集合。在具体实验中，应结合问题类型和数据情况，从相关性主题集合和创新性主题集合中选择领域扩展主题。

结合要素三元组和领域扩展搜索的主题，形成最终的大模型的输入要素模板，以方法性问题为例，为[动词]+[关键名词短语]+[附加信息]+[领域主题1]+[领域主题2]。

2.3 基于LLM的问题生成

采用LLM生成科学问题的优化方法：在2.3.1节分析选取了适合用于科学问题自动生成的原料，从而提升生成要素的语义质量；在2.3.2节介绍了大模型基于要素模板生成科学问题时的所采用参数优化指标。

2.3.1 问题生成的来源语料

在科技文献的常见语料中，研究问题句、创新贡献句、研究局限句和未来工作句均具备作为科学问题生成来源的潜力。研究问题句常出现在摘要或引言，虽能提供规范的问题表述范式，但多聚焦于已有的问题；创新贡献句一般用于总结研究中的理论或方法突破，在关联到未解问题时可能引出新的研究方向；研究局限句用于指出当前研究的不足，提示未来可探索的空间，但有一定的语义转化成本；未来工作句则直接指出后续研究方向，集中体现了待探索的科学问题。

从时间指向看，研究局限句与未来工作句更直接关联未来的研究空白，且未来工作句相较于研究局限句的语义转化成本更低，已有较成熟的识别方法。因此，本文将其作为生成科学问题的主要语料来源。

2.3.2 参数优化指标

为提升大模型在该任务上的生成质量，设计提示词的利用率(Utilization Rate of Prompts，URP)和新词引入率(Occupancy Rate of New Words，ORN)作为参数调优的评价指标，其选择基于生成任务目标的契合性和理论上的支持。具体而言：

URP用于衡量生成文本对输入提示模板的利用程度。科学问题自动生成任务的核心目标是基于结构化的输入模板生成高语义相关性的科学问题，因此URP直接反映了模型能否有效地利用输入信息。URP的计算方法如式（1）所示：

$$URP\ of\ sentence = \frac{G \cap P}{P} \tag{1}$$

其中，G表示组成科学问题的单词集合，$G = \{g_1, g_2, \cdots, g_n\}$；P表示组成输入要素的单词集合，$P = \{w_1, w_2, \cdots, w_m\}$。

ORN用于衡量生成文本中新词的占比，其计算方法基于式（2）的提示词占有率(ORP，Occupancy Rate of Prompts)。ORN定义为$1 - ORP$，即生成文本中提示词以外的单词所占的比例。科学问题生成任务不仅要求生成的问题与输入提示高度相关，还需要具备一定的探索性和创新性。因此，ORN的引入能够衡量模型是否能生成具有扩展性和领域探索意义的新问题。

$$ORP\ \text{of}\ sentence = \frac{G \cap P}{G} \tag{2}$$

调用LLM的API，通过设置4个参数Temperature、Top_k、Presence_penalty和Frequency_penalty不同值的组合，分别计算生成的科学问题的平均URP和平均ORN，从而确定最优的LLM参数组合。

3 实验和结果分析

科学问题的生成涉及众多学科领域，为验证本文提出的科学问题结构化识别与生成方法的有效性，设计并实施了一项案例研究。将实验聚焦于计算机科学的一个子领域——自然语言处理(Natural Language Processing，NLP)，该领域具有明显的学科交叉性和应用性，在问题类型方面，以方法性问题居多，因此以方法性问题为生成的目标。

实验的具体流程如图4所示。首先，设计并评估了从该领域中提取未来工作句子的几种方法；其次，从未来工作句子中提取要素三元组［动词]+[关键名词短语]+[附加信息]，并进行领域主题搜索，形成完整的输入要素模板；最后，基于大语言模型来生成具体的科学问题。本实验旨在证明该框架在特定学科及问题类型下从科技文献文本中生成有效科学问题的可行性与应用价值。

3.1 数据来源

实验选取的数据是Zhang C Z等[21]研究团队发布的未来工作句子数据集，其来源于自然语言处理领域的3个代表性会议ACL、EMNLP和NAACL的文献。该研究团队收集了2000—2019年这3个会议共计10 032篇论文，提取所有论文中可能会出现FWS的段落，包括结论段落(Conclusion)、讨论段落(Discussion)和未来工作段落(Future work)。从这些段落中人工标注64 896个句子，其中9 009条为FWS，55 887条为非FWS。

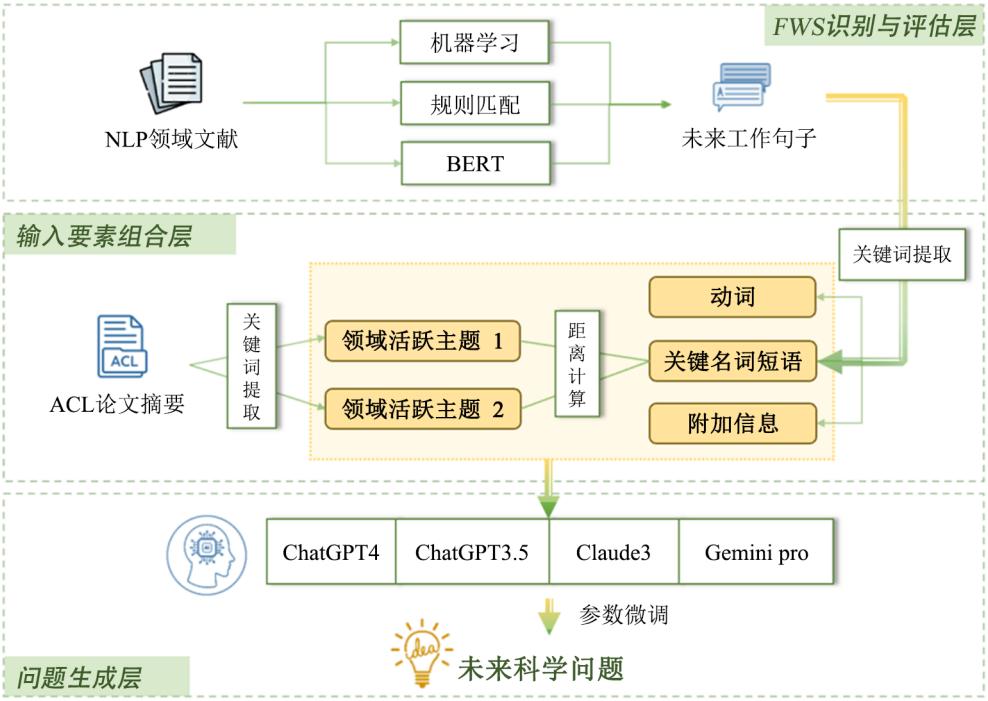

图 4　NLP 领域科学问题自动生成实验流程

Fig. 4　Workflow for Automatic Scientific Question Generation in NLP Domain

3.2　识别 FWS 结果与对比

该部分通过实验比较了 3 种识别 FWS 的方法，分别是基于机器学习的方法、规则匹配的方法和 BERT 预训练的方法。

1）基于机器学习模型的二分类预测法：分别采用朴素贝叶斯算法、随机森林算法、逻辑回归算法和支持向量机算法做 10 折交叉验证的训练，识别效果如表 2 所示。对比没有经过下采样的训练表现，可以发现下采样对于提升模型的召回率有显著的效果。综合比较 4 个模型，本文发现朴素贝叶斯在识别 FWS 上表现最好，其次是 SVM 模型。

表 2　机器学习模型在识别 FWS 上的表现

Tab. 2　Machine Learning Model Performance on FWS Identification

模型	训练集			测试集		
	精确度	召回率	F1 值	精确度	召回率	F1 值
随机森林	0. 851	0. 847	0. 849	0. 857	0. 845	0. 851
逻辑回归	0. 867	0. 862	0. 864	0. 870	0. 859	0. 865
SVM	0. 878	0. 875	0. 876	0. 875	0. 884	0. 879
朴素贝叶斯	**0. 919**	**0. 918**	**0. 918**	**0. 951**	**0. 951**	**0. 951**

注：加粗表示该指标最大值，下划线表示该指标次大值。

2）规则匹配方法：通过观察近 500 条句子（包括 300 条 FWS 和 200 条非 FWS）的结构和语言组成部分，归纳出三类正则表达式，分别是能够显著提取出 FWS 的规则（HSR，Highly Specific to FWS Rules）、能够显著提取出非 FWS 的规则（HNR，Highly Specific to Non-FWS Rules）、在 FWS 和非 FWS 中出现次数都较为频繁的规则（CR，Common Rules）。其中，HSR 类规则有 20 条，HNR 规则有 12 条，SR 类规则有 16 条。

规则匹配的识别效果如表 3 所示，通过引入 HSR 和 CR 类规则，去除 HNR 规则，并对数据做下采样，得到了较好的识别效果。

3）BERT 预训练分类方法：设置训练轮数为 5 轮，计算每次参数调整后的分类效果，得到训练结果如表 4 所示。该方法中，数据未经过下采样，发现当批量大小为 32、学习率为 1e-6、最大输入长

表 3 规则匹配方法的匹配效果

Tab. 3 Performance of Rule-Based Matching Methods

规则引入方式	训练集			测试集		
	精确度	召回率	F1 值	精确度	召回率	F1 值
A	**0.918**	0.519	0.663	**0.922**	0.532	0.675
A+C-F	0.512	**0.729**	0.602	0.526	0.734	0.613
A+C-F（下采样）	0.853	0.726	**0.784**	0.857	**0.745**	**0.797**

注：加粗表示该指标最大值，下划线表示该指标次大值。

度为256时，模型达到最优效果，此时模型在测试集上的精确度为0.883，召回率为0.774，F1值为0.825。

综合比较3种识别FWS的方法，在不经过下采样的情况下，FWS的识别难度较大，SVM模型、朴素贝叶斯模型和BERT模型都表现较好，其中，SVM模型在测试集的表现最好（精确度0.933，召回率0.768，F1值0.842）。在经过下采样后，朴素贝叶斯方法成为识别效果最好的模型（精确度0.951，召回率0.951，F1值0.951）。同时，规则匹配方法也能取得良好的识别效果（精确度0.857，召回率0.745，F1值0.797）。

表 4 不同参数下BERT模型的识别效果

Tab. 4 Performance of BERT Model

批量大小	学习率	最大输入长度	训练集			测试集		
			精确度	召回率	F1 值	精确度	召回率	F1 值
8	1e-6	256	0.914	0.849	0.880	0.777	0.671	0.720
32	1e-6	128	**0.949**	0.932	**0.940**	0.869	0.723	0.789
32	1e-6	256	0.947	0.932	0.939	**0.883**	**0.774**	**0.825**
64	1e-6	128	0.943	**0.937**	**0.940**	0.861	0.692	0.767

注：加粗表示该指标最大值，下划线表示该指标次大值。

3.3 生成要素组合输入模板

3.3.1 FWS中的要素三元组提取

对数据集中的FWS句子进行分析发现，绝大多数句子都与方法讨论相关，因此确定生成目标为方法类的科学问题，识别的要素三元组为[动词]+[关键名词短语]+[附加信息]。去除不相关的句子后，首先利用RAKE算法抽取FWS中的高频名词短语，采用的停用词包含①NLTK（Natural Language Toolkit）库的英文停用词表；②非名词；③与工作或研究有关的频繁词：study、research、work。以提取出的高频名词短语作为关键名词短语。保留RAKE算法中排名前10 000的长度大于等于2的名词短语，如表5所示。对无法表明研究内容的高频短语，如“nlp tasks”“future works”等做进一步的删除。可以发现机器翻译（machine translation）、关系抽取（relation extraction）、信息抽取（information extraction）、联合模型（joint model）等方向是目前NLP领域在未来工作中备受关注的话题。

表 5 出现次数最多的关键名词短语

Tab. 5 Most Frequent Key Noun Phrases

关键名词短语	出现次数
nlp tasks	60
machine translation	53
training data	41
language pairs	38
relation extraction	28
future works	23
wider range	18
information extraction	16
joint model	16

利用pos_tag词性标注的方法，标注出句子中的每一个动词，并将句子以动词为分隔，划分为若干动词+名词短语+状语的组合，匹配名词短语部分，最终形成[动词]+[关键名词短语]+[附加信息]的关键要素三元组。进一步删除不符合要求的

数据后，得到 1 828 个表示未来工作句子的要素三元组。

3.3.2 扩展搜索领域活跃主题的结果

选用NLP领域2021—2023年ACL会议论文的摘要信息作为拓展原料，生成的活跃主题如图 5 所示。分析发现，较为活跃的主题中有相当一部分是与论文的实验分析相关的，“f-score”“training data”等在每一年都频繁出现。值得关注的是，“language model”在 2021 年还未多次提及，但是在 2022 年和 2023 年出现频次逐渐升高，在 2023 年成为高频主题，反映出该主题在近年来越发受到关注，这也符合大语言模型等在 2022 年和 2023 年崛起并成为研究新热点的事实。

(a) 2021年 (b) 2022年 (c) 2023年

图 5 2021—2023 年的 ACL 摘要活跃主题

Fig. 5 Active Topics in ACL Abstracts From 2021 to 2023

计算[关键名词短语]和领域活跃主题之间的距离并排序，表 6 列举了 3 个[关键名词短语]和对应余弦相似度前十的领域活跃主题。计算排名前 20 的领域活跃主题的平均余弦相似度，与 2.2.2 节中的发现一致，前 2~3 个主题(相似度在 0.55 以上)与[关键名词短语]的结构和组成都非常相似，几乎很难提供新的信息。排名第 4~6 的主题平均相似度介于 0.5~0.55，是相关性主题集合，排名第 9~14 的主题平均相似度介于 0.4~0.45，是创新性主题集合。对数据做进一步的分析后，分别选取第 5 名与第 10 名的主题作为领域相关主题。此外，在实际操作中应根据数据分布与表现从集合中选取最合适的领域相关主题。

表 6 部分关键短语与余弦相似度前十的活跃主题

Tab. 6 Key Phrases and Top Ten Active Topics Based on Cosine Similarity

information retrieval	state space	robust model
information retrieval	state space	robust model prompt
information retrieval explainability	representation space	robust models
retrieval model	feature space	models robust
retrieval task	problem space	robust system
human computer information retrieval approach	search space	robust framework
text retrieval	online space	robust representations
retrieval results	answer space	robust encoder
retrieval performance	emotion space	robust ie
propose retrieval	test state	accurate robust
multi stage information retrieval pipeline	benchmark state	robust instances

在匹配领域相关活跃主题后，一共生成了 1 828 个[动词]+[关键名词短语]+[附加信息]+[领域主题 1]+[领域主题 2]的输入要素组合。采用 Word2Vec 方法表示组合向量，去除相似度超过 80% 的要素组合，最终剩下 1 124 个问题生成模型的要素组合。

3.4 大语言模型生成未来科学问题

3.4.1 大语言模型参数微调结果

选取前 100 条输入要素组合作为模型参数微调

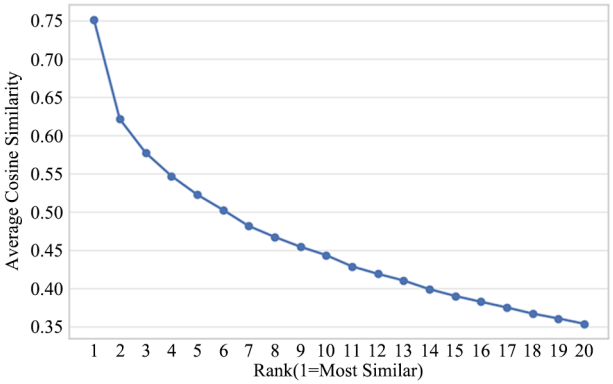

图 6 前 20 名的平均余弦相似度

Fig. 6 Top 20 Average Cosine Similarities

的原料，采用 ChatGPT-4 生成未来科学问题。对于 4 个参数 Temperature、Top_k、Presence_penalty 和 Frequency_penalty 不同值的组合，计算生成的科学问题的平均 URP 和平均 ORN 结果，如表 7 所示。在计算的时候对生成的科学问题和输入的提示词做了词性还原，以确保计算的一致性。Temperature 控制生成内容的随机性，值较高时生成的文本更加多样化。考虑到问题的生成需要一定的创新性和多样性，结合多次实验的结果，将 temperature 的基准设为 0. 6。

针对任务的特性，在更看重 URP 的情况下，发现第二个参数组合表现最好，该参数组合能够在充分利用提示词的基础上，引入大约 34% 的新词，兼顾了质量与多样性。第三个和第四个参数组合虽然获得了更好的多样性表现，但是未能引入足够多的提示词，无法很好地达到依据提示词生成科学问题的要求。

表 7 不同参数组合的生成效果

Tab. 7 Generation Performance of Different Parameter Combinations

参数组合	Temperature	Top_p	Presence_penalty	Frequency_penalty	URP	ORN
1	0. 6	1	0	0	0. 633	0. 30
2	0. 6	1	0. 5	0. 5	**0. 688**	0. 342
3	0. 8	1	1	1	0. 517	0. 43
4	1	1	2	2	0. 41	**0. 547**

注：加粗表示该指标最大值，下划线表示该指标次大值。

3. 4. 2 输入要素组合筛选结果

利用参数微调的 ChatGPT-4 对 1 124 个输入要素组合生成科学问题。生成的全体科学问题的平均 URP 为 0. 695，ORN 为 0. 372。为了筛选质量更高的要素组合，令 ChatGPT-4 每生成 50 个科学问题，就选出 10 个精选的问题，对于“精选”的提示是“符合自然语言处理前沿且具有研究价值”，再进行人工增删，得到共 224 个精选的要素组合和相应生成的科学问题，表 8 列举了其中的 3 个例子。

表 8 ChatGPT-4 生成的科学问题

Tab. 8 Scientific Questions Generated by ChatGPT-4

序号	输入要素组合	科学问题
1	［Verb］：include inferring the ［Keyword］：structure representations ［Additional information］：amr external smt knowledge ［Field keyword1］：structure level ［Field keyword2］：dialogue structure	Inferring Structure Representations in Dialogue Systems with AMR and External SMT Knowledge
2	［Verb］： integrate further nlp-based features ［Keyword］：coreference resolution ［Additional information］：question answering well citation classification graphical navigation ［Field keyword1］：event coreference resolution ［Field keyword2］：entity resolution	Enhancing NLP Features for Coreference Resolution in Question Answering and Citation Classification

表 8（续）

序号	输入要素组合	科学问题
3	［Verb］：include topics by ［Keyword］：latent dirichlet allocation ［Additional information］：additional view ［Field keyword1］：latent relations ［Field keyword2］：latent variables	Employing Latent Dirichlet Allocation to Explore Latent Relations and Variables in Text Data

3.4.3 对生成科学问题的评估结果

为评估基于要素模板的科学问题生成效果，本研究选取ChatGPT-4、ChatGPT-3.5、Claude3 Sonnet和Gemini Pro 4个代表性大语言模型作为主要评估对象，并进一步要求模型生成对应的创新摘要。模型参数统一设定为Temperature：0.6，Top_p：1，Presence_penalty：0.5，Frequency_penalty：0.5。计算大语言模型生成平均的URP和ORN指标，结果如表9所示。分析表明，模型在URP和ORN上各有侧重，表现出较高提示词利用率的模型有ChatGPT-3.5和Claude3 Sonnet，但这两个模型的新词引入率较低。ChatGPT-4和Gemini Pro在提示词利用率上略有降低，但引入了更多的新词，有更好的多样性。

为检验生成方法对模型迭代的适应性，本研究补充引入了新近发布的ChatGPT-5和DeepSeek R1进行科学问题的生成，并计算URP和ORN指标。结果显示，ChatGPT-5生成的URP为0.672，ORN为0.222；DeepSeek R1生成的URP为0.923，ORN为0.091。由此可见，ChatGPT-5与其他4个大模型表现相近，DeepSeek R1则利用了更多的提示词，引入了更少量的新词。可见，不同模型在具体任务中的表现有所差异，但均能基于要素模板稳定地生成结构规范、语义合理的科学问题，表明本文所构建的生成方法具备良好的通用性和可迁移性，其有效性不依赖于某一特定的模型架构或版本。

表9 大语言模型问题生成的URP和ORN

Tab. 9 URP and ORN Generated by LLMs

指标	ChatGPT-4	ChatGPT-3.5	Claude3 Sonnet	Gemini Pro	ChatGPT-5	DeepSeek R1
URP	0.694	<u>0.875</u>	0.848	0.707	0.672	**0.923**
ORN	<u>0.372</u>	0.327	0.16	**0.477**	0.222	0.091

注：加粗表示该指标最大值，下划线表示该指标次大值。

在完成自动指标的评估后，本研究进一步邀请了7位自然科学领域的专家对科学问题的生成内容评分。专家均来自南京大学，其中有教授3位，副教授2位，博士生2位。由于补充实验的指标评估结果已验证了生成方法的通用性，且ChatGPT-5的指标表现在前述4个模型范围之内，而DeepSeek R1虽URP值突出，但ORN值最低，问题新颖性相对有限。因此，专家评分环节将重点置于前述4个主要模型生成问题的质量比较上。考虑到评分过程的时间占用，从4个模型生成的问题中随机抽取30个科学问题予以专家打分。鉴于ChatGPT-3.5和Gemini Pro生成的创新摘要较为简短，且相较于另外两个模型在问题生成上风格更为统一，代表性高，因此这两个模型抽取更少量的问题（5个，其余两个模型10个）。

评分系统采用4个角度为科学问题打分（0~5分），分别是问题的清晰度（Clarity）、原创性（Originality）、可行性（Feasibility）、价值与意义（Significance）[34]。模型在4个维度上综合所有问题的平均得分，如表10所示。图7是根据四维平均得分绘制的雷达图，其中，最外圈表示得分最高的问题，最内圈表示得分最低的问题，介于两者之间的是得分均值。雷达图能够更清楚地体现各个模型在4个维度的优劣。

整体而言，经过参数微调的大语言模型，依据要素组合输出的科学问题，能够取得较好的清晰度和可行性，且问题大多是具有研究价值的，能够在本领域和相关领域产生影响。但在问题的原创性上多存在不足。其中，Claude3 Sonnet生成的问题

表 10 不同模型在 4 个维度上的平均得分

Tab. 10 Mean Scores of Models Evaluated on Four Dimensions

模型	清晰度	原创性	可行性	价值/意义	四维均分
ChatGPT-4	3.08	2.75	3.17	3.22	3.06
Claude3 Sonnet	**3.36**	**3.04**	3.04	**3.30**	**3.18**
Gemini Pro	3.14	2.64	**3.23**	3.14	3.04
ChatGPT-3.5	2.89	2.43	2.97	2.74	2.76

注：加粗表示该指标最大值，下划线表示该指标次大值。

在 4 个模型中表现最好，有着最高的原创性、清晰度、价值与意义的评分，且明显好于其他 3 个模型，在可行性上也有良好的表现。其次是 ChatGPT-4 和 Gemini Pro 模型，ChatGPT-4 在不同问题上的得分差异较大，得分最高的问题表现很好，但也不乏四维得分都极低的问题，Gemini Pro 则分差更小，表现更稳定。ChatGPT-3.5 在各个维度的评分都是最低的，相较而言生成效果最差。

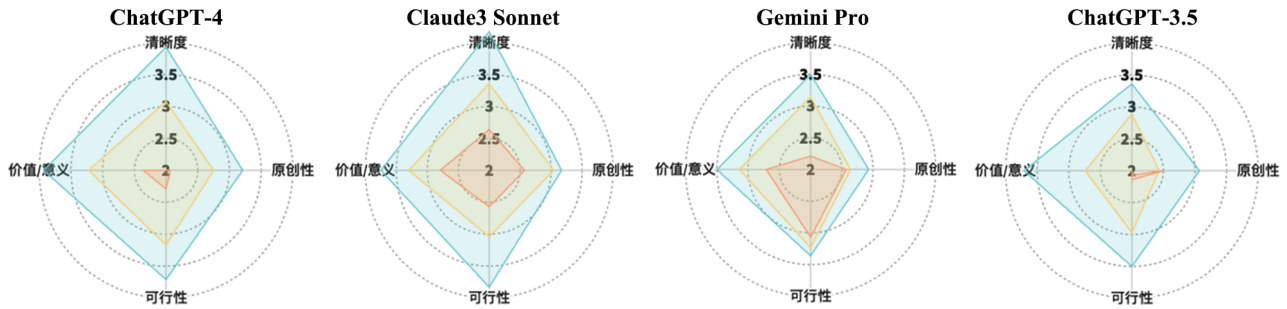

图 7 4 个模型在不同维度平均得分的雷达图

Fig. 7 Radar Chart Depicting Average Scores of Four Models

总体来看，各模型均能生成高质量的方法论问题，如表 11 所示。这些问题在 4 个维度上的平均得分基本在 3 分以上，话题包括通过人类情感需求改进聊天机器人算法、未见词生成的方法、语音与声调在模型中的集成和稀疏注意力矩阵等，是现今自然语言处理领域较为前沿和创新的方向。其中，未见词生成的问题在各方面的得分都是最高的，可以发现其输入要素组合的构成是最丰富的。总的来说，这些问题的可行性和价值/意义都非常高，证明问题具有很大的研究价值，较高的清晰度也表明问题表述足够明确和充分，可以为领域内的学者提供启发性的参考。

4 总结和展望

面对从海量文献中寻找新颖选题的挑战，本文提出了一种基于大语言模型自动生成未来科学问题的新方法(AGMSQ)。主要贡献在于：第一，针对当前普遍地将科学问题同质化处理的问题，提出了科学问题的分类框架，并根据不同类型的科学问题针对性地设计了不同的要素组合模板。第二，要素组合的设计基于科学问题本身的结构特征，并充分利用了生成原料的语义信息与结构化线索。第三，采用了 ChatGPT、Claude3 Sonnet、Gemini Pro 等大语言模型，验证了该框架在生成 NLP 领域方法性问题上的应用潜力，为进一步探索其他学科和问题类型提供了经验和数据支持。此外，本文还提出了 URP 和 ORN 两个评价指标为大模型的生成优化提供方向。

本研究仍具有一些局限性。第一，本研究实验部分仅测试了自然语言处理领域生成方法性问题的效果，并未验证其他学科和其余问题类型的生成模板的有效性。因此，当前研究结论的适用范围存在一定的边界，其在不同领域和其他问题类型的泛化能力有待进一步研究。未来可以在多个代表性学科中构建验证语料，针对不同类型的问题测试生成模板的有效性与稳定性。第二，当前对生成科学问题的评估依赖于专家打分，可能面临成本高、主观性强的问题。未来可以考虑将高质量科学问题的结构性特征与认知价值纳入自动化或半自动化的评估框架，可以借鉴如 Science 提出的“125 个未解科学问题”等权威问题集的结构

表 11 大语言模型生成的问题示例

Tab. 11 Examples of Scientific Questions Generated by LLMs

模型	科学问题及创新摘要
ChatGPT-4	标题：Analyzing Human Needs Through Affective Events，Narrative Stories， and Conversations 输入要素：[Verb]：analyzing the；[Keyword]：human needs；[Additional information]：affective events narrative stories conversations；[Field keyword1]：annotator needs；[Field keyword2]：nmt system needs 创新摘要：人类的情感需求和反应对于设计更为人性化的交互系统非常关键。这项研究通过分析情感事件、叙事故事和对话内容来理解人类的基本需求和情绪反应。这对于改进聊天机器人、推荐系统和其他基于用户交互的应用非常有价值，可以使这些系统更加贴近用户的实际需求和情感状态。 平均得分：清晰度3.92；原创性3.08；可行性3.67；价值/意义3.92
Claude3 Sonnet	标题：Incorporating Open Vocabulary and Copy Mechanism for Explicit Unseen Words Generation with Pointer-Copy Network and Novel Attention Mechanism 输入要素：[Verb]：incorporating open vocabulary and；[Keyword]：copy mechanism；[Additional information]：explicit unseen words generation；[Field keyword1]：pointer copy network；[Field keyword2]：novel attention mechanism 创新摘要：未见词生成一直是自然语言生成任务中的一大难题。该标题提出结合开放词汇和复制机制，利用指针-复制网络和新颖注意力机制进行显式未见词生成。具体来说，该方法将维护一个动态词汇表，在生成时同时预测来自已知词汇和未见词汇的token。对于后者，模型将利用指针网络从源序列复制相应token。同时，新颖的注意力机制将进一步增强未见词生成的能力。通过显式对待未见词，该方法能够有效缓解未见词带来的问题，从而提高生成质量。此外，开放词汇和复制机制也有助于提升模型的泛化能力，为构建通用生成系统奠定基础。 平均得分：清晰度4.17；原创性3.17；可行性3.83；价值/意义3.75
Gemini Pro	标题：Integrating a Richer Contextual Representation of Tone and Intonation Consistent for Improved Consistent Style and Way Consistent 输入要素：[Verb]：integrate a richer contextual representation of tone and；[Keyword]：intonation consistent；[Additional information]：phonetic theory within unsupervised；[Field keyword1]：consistent style；[Field keyword2]：way consistent 创新摘要：语调和声调在自然语言处理中扮演着关键角色。本研究旨在将更细腻的语调和声调表示集成到模型中，以提高一致风格和方式的处理能力。这一创新方法将推动涉及口语理解和情感分析等领域的发展，为构建更人性化的智能系统带来新的机遇。 平均得分：清晰度3.42；原创性2.92；可行性3.25；价值/意义3.42
ChatGPT-3.5	标题：Dig Deeper on the Benefits of Sparse Attention： Sparse Attention Matrix Estimation Module for Sparse Latent Representations 输入要素：[Verb]：dig deeper on the benefits of；[Keyword]：sparse attention；[Additional information]：terms better interpretability context aware nmt models；[Field keyword1]：sparse attention matrix estimation module；[Field keyword2]：sparse latent representations 创新摘要：这个研究深入探讨了稀疏注意力机制的优势，并重点开发了用于稀疏潜在表示的稀疏注意力矩阵估计模块。通过研究稀疏注意力在自然语言处理模型中的应用优势，旨在提高注意力机制的解释性、效率和性能。这项研究对于稀疏注意力在机器翻译、情感分析和文本分类等各种自然语言处理任务中的理解和应用具有重要意义。 平均得分：清晰度3.25；原创性3.08；可行性3.42；价值/意义3.67

与范式。在评估维度上，未来也可以纳入长期影响力、突破性等深度衡量科学潜力的指标。

参 考 文 献

[1] 宋若璇，钱力，杜宇. 基于科技论文中未来工作句集的学术创新构想话题自动生成方法研究［J］. 数据分析与知识发现，2021，5（5）：10-20.

[2] Popper K. The Logic of Scientific Discovery［M］. London：Routledge，1959.

[3] 默顿. 科学社会学：理论与经验研究［M］. 鲁旭东，林聚任，译. 北京：商务印书馆，2003.

[4] Laudan L. Progress and Its Problems：Toward a Theory of Scientific Growth［M］. Berkeley：University of California Press，1977.

[5] Cook T D，Campbell D T. Quasi-Experimentation：Design & Analysis Issues for Field Settings［M］. Boston：Houghton Mifflin，1979.

[6] Anderson L W，Krathwohl D R. A Taxonomy for Learning，Teaching，and Assessing：A Revision of Bloom's Taxonomy of Educational Objectives：Complete Edition［M］. Addison Wesley Longman，Inc.，2001.

[7] Mohasseb A，Bader-El-Den M，Cocea M. Question Categorization and Classification Using Grammar Based Approach［J］. Information Processing & Management，2018，54（6）：1228-1243.

[8] 申维玺，孙燕. 理论研究在中医药学关键科学问题研究中的作用［J］. 中医研究，2006，19（1）：1-3.

[9] 彭玉生. "洋八股"与社会科学规范［J］. 社会学研究，2010，25（2）：180-210.

[10] Bryson A E，Ho Y C. Applied Optimal Control：Optimization，Estimation，and Control［M］. New York：Taylor & Francis Group，1975.

[11] Wang Q Y，Huang L F，Jiang Z Y，et al. PaperRobot：Incremental Draft Generation of Scientific Ideas［EB/OL］.［2025-11-20］. https://arxiv.org/abs/1905.07870.

[12] Li L，Wang Y，Xu R，et al. Multimodal Arxiv：A Dataset for Improving Scientific Comprehension of Large Vision-language Models［EB/OL］.［2025-11-18］. https://arxiv.org/abs/2403.00231.

[13] Si C，Yang D，Hashimoto T. Can LLMs Generate Novel Research Ideas? A Large-scale Human Study With 100+ NLP Researchers［EB/OL］.［2025-11-24］. https://arxiv.org/abs/2409.04109.

[14] Hu X，Fu H，Wang J，et al. Nova：An Iterative Planning and Search Approach to Enhance Novelty and Diversity of LLM Generated Ideas［EB/OL］.［2025-11-24］. https://arxiv.org/abs/2410.14255.

[15] Su H，Chen R，Tang S，et al. Many Heads Are Better Than One：A Multi-agent System Has the Potential to Improve Scientific Idea Generation［EB/OL］.［2025-11-24］. https://arxiv.org/abs/2410.09403.

[16] Li R，Jing L，Du X. Learning to Generate Research Idea With Dynamic Control［C］//2nd AI4Research Workshop：Towards a Knowledge-grounded Scientific Research Lifecycle. 2024.

[17] Hu Y，Wan X. Mining and Analyzing the Future Works in Scientific Articles［EB/OL］.［2025-11-30］. https://arxiv.org/abs/1507.02140.

[18] Li K，Yan E. Using a Keyword Extraction Pipeline to Understand Concepts in Future Work Sections of Research Papers［J］. Computer Science，2019：87-98.

[19] Zhu Z H，Wang D B，Shen S. Recognizing Sentences Concerning Future Research From the Full Text of JASIST［J］. Proceedings of the Association for Information Science and Technology，2019，56（1）：858-859.

[20] Hao W K，Li Z C，Qian Y C，et al. The ACL FWS-RC：A Dataset for Recognition and Classification of Sentence About Future Works［C］//Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020，2020：261-269.

[21] Zhang C Z，Xiang Y，Hao W K，et al. Automatic Recognition and Classification of Future Work Sentences From Academic Articles in a Specific Domain［J］. Journal of Informetrics，2023，17（1）：101373.

[22] 谢林蕾，向熠，章成志. 面向融合出版前沿主题发现的学术论文未来工作句挖掘研究［J］. 情报工程，2023，9（5）：123-138.

[23] Azher I A，Mokarrama M J，Guo Z，et al. FutureGen：A RAG-based Approach to Generate the Future Work of Scientific Article［C］//2025 IEEE International Conference on eScience（eScience）. IEEE，2025：427-438.

[24] 康永征，辛申伟. 跨学科视阈下的社会科学研究方法［M］. 北京：中国社会科学出版社，2012.

[25] Fain H，Hempel C G. Aspects of Scientific Explanation and Other Essays in the Philosophy of Science［J］. American Sociological Review，1966，31（1）：130.

[26] Hacking I. Representing and Intervening：Introductory Topics in the Philosophy of Natural Science［M］. Cambridge University Press，1983.

[27] Chang H. Inventing Temperature：Measurement and Scientific Progress［M］. Oxford；New York；Tokyo：Oxford University Press，2004.

[28] 靳玉乐. 探究教学论［M］. 重庆：西南师范大学出版社，2001.

[29] Clarke B. Empirical-Normative Distinction［J］. British Journal of Political Science，1983，13（4）：511-512.

[30] 索传军，牌艳欣. 科学问题谱系构建研究［J］. 中国图书馆学报，2025，51（1）：61-81.

[31] Van Valin R D，Lapolla R J. Syntax：Structure，Meaning，and Function［M］. New York：Cambridge University Press，1997.

[32] Morley G D. Determining Objects，Adjuncts and Complements in English［J］. Word，1991，42（3）：295-302.

[33] Levin B. English Verb Classes and Alternations：A Preliminary Investigation［M］. Chicago：University of Chicago Press，1993.

[34] Baek J，Jauhar S K，Cucerzan S，et al. ResearchAgent：Iterative Research Idea Generation Over Scientific Literature With Large Language Models［EB/OL］.［2025-12-01］. https://arxiv.org/abs/2404.07738.

（责任编辑：李汇森）