# （一）基于大语言模型的科学假设生成系统

## 系统背景

随着知识指数级增长和跨学科研究的复杂性增加，机器学习研究人员面临信息过载和探索新想法的重大挑战。在此背景下，如何有效借助外部工具生成创新科研想法变得至关重要。最新的大语言模型GPT4、Qwen等，展示了在文本生成方面的巨大潜力。这些模型不仅能够理解和生成复杂的学术内容，还能对齐多模态信息，构建隐含的思维链，揭示隐藏的联系。通过充分利用大语言模型的这些能力，可以加速研究人员新的科研想法的产生，提升科研效率。

但是目前的通用大语言模型在科学假设生成领域存在以下问题：1. 产生能解决背景中问题的新颖且可行的想法，需要总结前人工作，目前通用大语言模型无法自主检索相关文献；2. 在线文献档案仅限于简单的关键字匹配，无法充分利用用户提供的信息或现有文献，导致检索结果不完整和不准确；3. 科学论文想法既需要新颖性，也需要可行性。然而，如何使大语言模型能够产生全新的想法同时确保其可行性，具有挑战性。

## 主要功能

基于大语言模型的科学假设生成系统的主要目的是结合大语言模型以及检索技术进行科学假设生成，主要模块包括文献数据库构建、文献数据库检索和科学假设生成三大模块。

（1）文献数据库构建

从机器学习领域收集大量文献，并使用实体提取、语义编码、摘要和引文分析等技术为每篇论文提取多个维度的信息，构建文献数据库。

（2）文献数据库检索器

提出基于语义、实体和引文共现的文献检索方法。在这个框架中，语义检索捕获论文的全局信息，实体检索关注局部细节，引文共现检索反映了之前研究人员发现的隐藏关系。通过在这三个不同的粒度级别上进行匹配，提供了更全面的文献检索。

（3）科学假设生成

根据给出的问题，通过文献检索器到的文献产生灵感，并使用头脑风暴方法来产生新的想法。根据基于文献检索生成和基于头脑风暴生成的想法进行结合，并经过进一步筛选和细化，从而提高其新颖性和可行性。

## 开发环境

操作系统及版本：Linux Ubuntu22.04

支撑软件及版本列表：Python 3.10 PyTorch 2.0.0

数据库环境：Neo4j 社区版 5.20.0

## 基本的架构体系

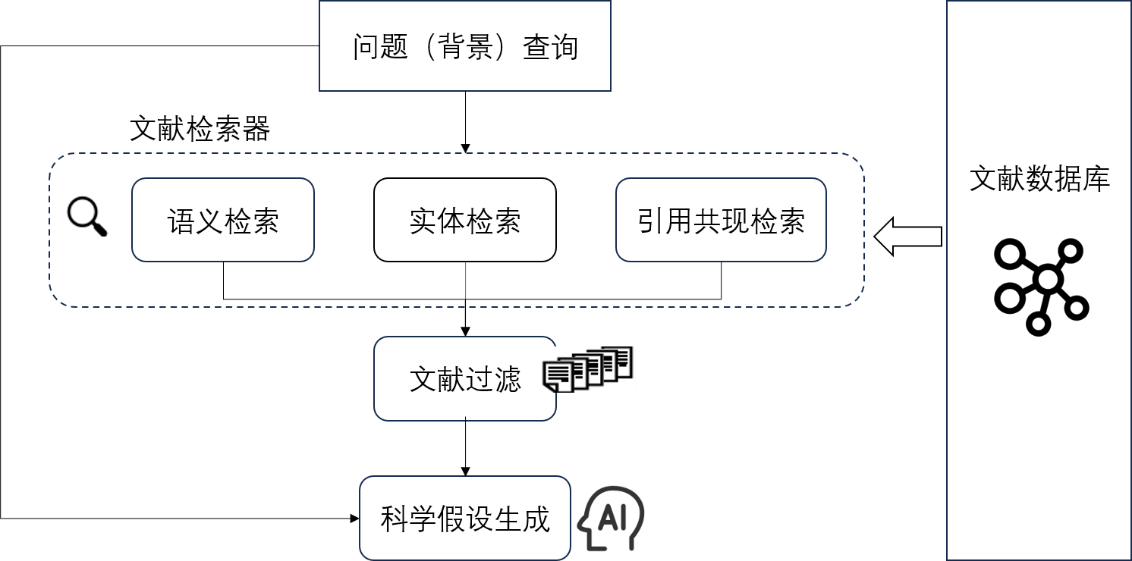


图 1 科学假设生成系统架构图

文献数据库：主要包括对机器学习领域大量文献的收集，使用实体提取、语义编码、内容总结和引文网络构建等技术，提取论文的多维度信息。具体实现包括：1. 通过识别文献摘要中的关键概念、研究方法、问题描述等重要信息，提取文本中的关键实体；2. 使用大语言模型自动生成精简的摘要，得到文献的核心贡献；3. 文献中的文本信息转化为高维度的语义表示，利用预训练语言模型例如（Bert）编码总结内容，从而形成语义向量；4. 分析文献间的引用关系，构建引用网络。

文献检索器：主要分为语义检索、实体检索和引文共现检索。1. 基于语义编码，通过深度学习模型理解文献的整体内容，捕获论文的全局信息，帮助用户从语义层次上搜索相关文献；2. 通过实体提取技术检索特定的关键词，构建实体网络；3. 通过分析引文的共现情况，识别文献间的隐性联系。

科学假设生成：系统根据用户提出的问题，结合文献检索器中的检索结果，使用大语言模型来启发性地产生假设。其具体实现包括：1. 基于检索到的相关文献，从文献的研究问题、方法和结果中产生启发；2. 大语言模型通过头脑风暴方法生成新的想法，将文献中已有的研究成果与用户当前问题进行创新性结合；3. 结合基于文献检索的生成和头脑风暴生成的想法，经过模型对假设的新颖性、可行性进行筛选和细化，以确保提出的假设具备科学性和创新性。