南京师范大学
大学生创新创业训练计划项目申报表
（创新训练项目）

学院名称：	计算机与电子信息学院/ 人工智能学院
项目名称：	面向中小学科学教育大模型的 领域知识增强方法研究与实现
项目类型：	☑ 省级项目 ☐ 校级项目
所属一级学科名称：	计算机科学与技术
所属重点领域：	人工智能
项目负责人：	时子延
联系电话：	18351997562
指导教师：	周俊生
联系电话：	18951917729
申报日期：	2023 年 12 月 21 日

南京师范大学教务处 制

二〇二三年十一月

填表说明

一、申报表要按照要求逐项认真填写，填写内容必须实事求是表述准确严谨。空缺项要填“无”。

二、格式要求：表格中的字体采用小四号宋体，单倍行距；需签字部分由相关人员以黑色钢笔或签字笔签名。

三、项目类型为省级项目、校级一般项目。

四、项目来源：1. “A”为学生自主选题，来源于自己对课题的长期积累与兴趣；“B”为学生来源于教师科研项目选题；“C”为学生承担社会、企业委托项目选题。2. “来源项目名称”和“来源项目类别”栏限“B”和“C”的项目填写；“来源项目类别”栏填写“863 项目”、“973 项目”、“国家自然科学基金项目”、“省级自然科学基金项目”、“教师横向科研项目”、“企业委托项目”、“社会委托项目”以及其他项目标识。

五、所属重点领域：拟申报**省级项目选填**，如果属于重点领域的则填报。具体包括 10 类：泛终端芯片及操作系统应用开发、重大应用关键软件、云计算和大数据、人工智能、无人驾驶、新能源与储能技术、生物技术与生物育种、绿色环保与固废资源化、第五代通信技术和新一代 IP 网络通信技术、社会事业与文化传承。

六、表格栏高不够可增加。

七、填报者须注意页面的排版。

<table>
<tr><td>项目名称</td><td colspan="5">面向中小学科学教育大模型的领域知识增强方法研究与实现</td></tr>
<tr><td>项目所属
一级学科</td><td colspan="2">计算机科学与技术</td><td>项目所属
二级学科</td><td colspan="2">人工智能</td></tr>
<tr><td>项目类型</td><td colspan="5">（ √ ）省级项目　　（ ）校级项目</td></tr>
<tr><td>所属重点领域</td><td colspan="5">人工智能</td></tr>
<tr><td rowspan="2">项目来源</td><td>A</td><td>B</td><td>C</td><td>来源项目名称</td><td>来源项目类别</td></tr>
<tr><td></td><td>√</td><td></td><td>高原造峰专项学科群建设
——计算机科学与技术</td><td>南师大学科建设项目</td></tr>
<tr><td>项目实施时间</td><td colspan="5">起始时间：　2023 年 12 月　完成时间：2025 年 6 月</td></tr>
<tr><td>项
目
简
介
(限 200 字)</td><td colspan="5">本项目是一个基于知识增强的科学教育问答系统，通过整合中小学教育知识，我们建立了庞大的 NebulaGraph 知识库与知识图谱，致力于提供高效、精准的中小学教育知识服务。在实现中，本项目采用向量化索引技术，实现了对知识库的智能检索，使得系统能够在海量数据中快速准确地定位相关信息；同时通过知识图谱推荐算法，我们将相关的知识点相连接，实现了知识的智能推荐。最后，通过 neo4j 和 langchain 的协同作用，实现系统的部署与发布，保证了系统的高效运行，也为未来的更新和维护提供了可行性和可持续性。</td></tr>
</table>

申请人或申请团队		姓名	年级	学号	所在院系/专业	联系电话	QQ 邮箱
	主持人	时子延	22	19220422	计算机与电子信息学院/人工智能	18351997562	szy@nnu.edu.cn
	成员	董文杰	22	19220448	计算机与电子信息学院/人工智能	19805119351	2785794768@qq.com
		马艺轩	22	19220402	计算机与电子信息学院/人工智能	15996880830	1023006192@qq.com

指导教师					
	第一指导教师	姓名	周俊生	单位	计算机与电子信息学院
		年龄	51	专业技术职务	教授
	主要成果		周俊生教授先后主持 3 项国家自然科学基金面上项目和多个省部级项目与智慧教育横向合作项目的研究工作，在 IJCAI、ACL、COLING 等人工智能和自然语言处理领域的顶级国际学术会议国内外期刊上发表论文 60 余篇。目前担任中国人工智能学会语言智能专委会副主任、江苏省人工智能学会自然语言处理专委会副主任、中国计算机学会自然语言处理专委会委员、中文信息学会语言与知识计算专委会委员。		

	第二指导教师	姓名	张博	单位	计算机与电子信息学院
		年龄	28	专业技术职务	讲师
	主要成果		张博近年来已在国际顶级期刊 IEEE TKDE、IEEE TIP、IEEE TCYB 及国际顶级会议 ACL、EMNLP、ISWC 等刊物上发表论文十余篇，获得知识图谱顶级会议 ISWC2022 候选最佳论文（CCF-B 类会议，<2%）。获得授权发明专利 2 项。		

一、申请理由

项目主持人：时子延，南京师范大学人工智能专业 22 级本科生。在读期间活得“蓝桥杯”工信部计算机程序设计竞赛省赛二等奖、“高教社杯”全国大学生数学建模竞赛省赛二等奖、2023BMW 大学生联合创新挑战营——上海站：漾 AI·灵感闪耀奖等奖项；高中曾获全国中学生生物学竞赛国家一等奖, 江苏省一等奖, 江苏省第 13 名。在 The 10th INTERNATIONAL CONFERENCE ON BEHAVIOURAL AND SOCIAL COMPUTING 发表论文 MAgic: A Morphable Attention Based Algal Tiny Object Detection Model 一篇；熟悉 Python，Java，C++，C，HTML，CSS，JavaScript，C#，Rust，Swift 等编程语言，具备网络开发前后端全栈软件开发的能力和算法编写能力。熟练使用以 Ubuntu 和 Kali 为代表的 Linux 和 MacOS，Windows 等操作系统,熟悉 Git，bash，vim，tmux，gdb 等命令行实用工具，能够使用 makefile 管理项目，python，sh 编写脚本。擅长使用 LaTeX 进行论文撰写与排版，了解深度学习，自然语言处理，大语言模型、数据结构与算法、数学建模相关算法，网络安全相关计算机基础知识。具备渗透测试相关工具使用经验。对生命科学，脑科学，认知科学领域下的认知，智能，学习，推理，记忆，意识与自我等课题感兴趣。

项目成员：董文杰，南京师范大学人工智能学院人工智能专业 2022 级本科生。学习态度严谨，综测绩点均位于专业前 10%；对学科充满热情，搭建“南京云锦”数字化信息检索平台并持续运营与维护；曾获“电工杯”全国大学生数学建模竞赛三等奖，“高教社杯”全国大学生数学建模竞赛赛区二等奖，“蓝桥杯”工信部程序设计竞赛赛区二等奖。有良好的数学基础，精修高等数学、线性代数与概率论。对 AI 领域动态时刻关注，并热爱主动实践相关算法，有一定的 LLM 大模型学习基础。热爱编程，对于传统算法与科学计算有一定的熟练度，对于相关系统的开发也有一定的熟练度。

项目成员：马艺轩，计算机与电子信息/人工智能学院人工智能专业 2022 级本科生，在团队中主要负责数据处理，数据分析，程序构建等。该生在校期间，勤勉学习，积极参加各种比赛和项目，如校算法悬赏令，蓝桥杯，数学建模，icpc 等比赛，并取得优异的成绩。同时，除专业学习以外，大一时参与项目研究，帮助处理数据和搭建神经网络的数据集和模型。该生还积极学习算法设计，机器学习模型，数学建模等相关知识，熟悉并掌握 matlab，opencv，torch，pandas，C++等等。在校期间多次参与院的微课活动：通过个人博客，视频录制等为在校学生讲解计算机相关知识，累计获得志愿时长 57.5h。该生做事认真负责，注重工作效率和质量，在各种各样的实践活动中锻炼了自身的组织领导能力，问题分析处理能力，沟通协调能力等。

二、项目方案

（一）项目研究背景

1. 项目研究意义

科技进步的浪潮席卷全球，使得我们的生活、工作、学习方式发生了翻天覆地的变化。在这个信息爆炸的时代，人们对快速、准确、便捷地获取知识和信息的需求日益增长。问答系统能够理解用户提出的问题，从海量知识库中检索出最准确的答案，并以=个性化？的方式呈现给用户。问答系统作为人工智能领域的一项重要技术，正逐渐成为满足这一需求的关键途径。

基于大语言模型的科学教育问答系统，满足了我们对于高效、全面科学教育知识库的需求。这种系统具有强大的语义理解和知识推理能力，不仅能够精准深入地解答学生问题，还能为教师提供丰富的教学资源，使他们能够更好地在教学中应用科学知识。总的来说，构建这样的系统是响应智慧教育的重要举措，能为广大师生提供便捷、高效的服务，推动我国科学教育事业的发展。

然而，由于大语言模型尚存在不可解释性、不可溯源性和知识性幻觉等问题，导致其在垂直领域的应用中出现准确性缺失，问题答案与事实不符等。特别的，在科学教育领域，这种错误可能会对学生的理解和学科知识的正确传递产生负面影响。因此，在使用大语言模型时，本项目组类比人类的的思维运行方式，即根据已有的认知进行思考，将大语言模型视为人的思维中心，引入知识图谱作为认知中心，通过知识图谱增强的方式引导大语言模型，以提高其在特定垂直领域的准确性。

基于以上问题，本组通过构建特定学科的知识图谱，设计出一套基于知识增强的科学教育问答系统。该系统可以给出知识点的出处和来源，保证知识点的准确性，避免出现一些知识性错误。同时，引入图谱增强的技术，提取模型初始回答中的实体作为查询实体，并结合已构建好的科学教育领域的知识图谱，让大模型针对这些邻实体生成知识信息推荐或者提问，帮助学生构建相应的知识体系，从而实现启发式教育。

启发式教育也是近几年国家鼓励和支持的教育方法。本组设计的科学教育问答系统，注重激发学生的兴趣和主动性，鼓励学生独立思考和创造性思维，从而充分调动学生学习的积极性，使学生在与系统对话的过程中逐步构建相应的知识体系，主动地参与学习过程，从而更好地理解和掌握知识。

2. 国内外的研究现状

2.1 问答系统

问答系统（question answering，QA）能够自动回答用户提出的自然语言问题，是信息检索和自然语言处理的交叉研究方向。问答系统的实现途径主要包括基于知识图谱、基于语言模型、基于信息检索的等方式。

基于知识图谱（Knowledge Graph，KG）的问答系统通过构建知识图谱，将现实世界中的实体、概念和关系进行结构化表示，利用查询解析、信息检索和知识嵌入等技术实现对问题的精确解析和回答。（2）基于语言模型的问答系统利用深度学习技术训练语言模型，使其能够理解自然语言中的语法、语义和上下文信息。在问答系统中，语言模型可以用于生成回答。（3）基于信息检索的问答系统通过索引技术建立问题和答案之间的关联，跟据用户提出的问题，从大量文本数据中检索出相关的答案。这类系统通常采用关键词匹配、文本相似度计算等技术来找到最匹配的答案。（4）基于机器学习的问答

系统通过训练机器学习模型，让系统从数据中学习到问题和答案之间的关系，从而实现对用户问题的自动回答。这类系统通常采用文本分类、文本回归、序列标注等技术来实现。

本文主要采用大语言模型 (Large Language Model，LLM) 作为问答系统的基座模型，通过实现检索增强和图谱增强的两种技术来对完善问答系统的回答。

2.2 大语言模型

大语言模型在大量语料上进行预训练的大语言模型，在多种自然语言处理任务上取得了显著的成效。近年来，随着模型的参数量不断增加，使得 LLM 表现出涌现能力。像 InstructGPT[1]、ChatGPT[2]、GPT4[3] 这些自回归大语言模型[4] 通过预训练、微调（fine-tuning）等技术理解并遵循人类指令，在类似于教育[5] 、代码生成[6] 、推荐[7]。等复杂的实际任务中显示出巨大的潜力。

尽管这些模型在许多应用中展现出了强大的能力，但无法忽视其存在的局限性：（1）LLM 因为缺乏事实知识而饱受诟病。LLM 知识受到预训练语料和模型能力的限制：语料的覆盖范围有限，语料质量难以保证，语料存在时效性。模型难以学完并记住所有语料知识，无法获得关于最近事件的最新信息，模型的推理能力差，产生有害幻觉事实的回答[8]。。（2）LLM 作为黑盒模型，其也因缺乏可解释性而受到批评。LLM 在其参数中隐含地表示知识，通过 LLM 所获得的知识很难进行解释或验证。此外，LLM 通过概率模型进行推理，这是一个不确定的过程。用于进行预测或决策的 LLM 的特定模式和功能对人类无法直接访问或解释。尽管一些 LLM 能够通过应用思维链[9] 来解释他们的预测，但他们的推理解释也存在幻觉的问题[10] 。

由于这些局限性，将 LLM 直接应用于专业领域的问答仍然存在诸多问题。一方面，缺乏专业领域的知识或新的训练数据，在一般语料库上训练的 LLM 可能无法很好地推广到专业的领域。另一方面，很多工作通过采取数据微调的方式修改模型的参数，从而增强 LLM 应对专业领域问题的能力。然而，一些文献指出这些数据微调的方法会产生灾难性遗忘[11] ，致使模型丧失原始对话的能力，甚至在处理非微调的数据时会产生混乱的结果。

2.3 知识增强问答系统

（1）基于图谱增强的问答系统研究

KG 是由 Google 公司在 2012 年提出来的一个新的概念。从学术的角度，我们可以对 KG 给一个这样的定义：“知识图谱本质上是语义网络 (Semantic Network) 的知识库”。从实际应用的角度出发其实可以简单地把 KG 理解成多关系图 (Multi-relational Graph）。

知识图谱是一种大规模的、用来储存人类知识的数据集。将结构化的知识存储为三元组的形式$\mathrm{KG} = \{(\mathrm{h}, \mathrm{r}, \mathrm{t}) \subseteq E \times R \times E\}$，其中 E 和 R 分别表示实体和关系的集合，而 h 则表示头实体，t 表示尾实体，r 表示头实体和尾实体之间的关系。现有的知识图跟据存储的信息可以分为四组：1）百科全书式的知识图谱，2）常识性的知识图谱，3）领域特定的知识图谱，4）多模态的知识图谱。基于图谱增强大语言模型是指通过利用等结构化数据，来提高 LLM 的生成能力和准确性。

Rony[14]等人提出了 DialoKG，一种新颖的面向任务的对话系统，它有效地将知识整合到一个语言模型中。他们提出的系统将关系知识视为知识图，并引入了（1）结构感

知知识嵌入技术和（2）知识图加权注意掩蔽策略，对知识图谱中存在一些与当前问题无关的知识进行过滤，以促进系统在对话生成过程中选择相关信息。

Sun[15]等人认为现有的用于常识性问答的知识图谱增强模型主要集中在设计复杂的图神经网络（Graph Neural Network，GNN）来对知识图谱进行建模，却忽略了对问题上下文表征和知识图谱表征的有效融合和推理，以及在推理过程中自动从知识图谱中选择相关节点。对此，他们提出了一种新的模型 JointLK，通过 LM 和 GNN 的联合推理以及动态 KGs 剪枝机制解决了上述局限性。

Yu[16]。等人提出了一种新的联合预训练框架 JAKET，用于对知识图谱和语言进行建模。知识模块和语言模块提供相互协助的基本信息：知识模块为文本中的实体生成嵌入，而语言模块为图中的实体和关系生成上下文感知的初始嵌入。基于传统掩码[MASK]的预训练任务，预测知识中被掩盖的实体、实体类别与关系，同时预测知识三元组与释义文本中的被掩码内容。

Sun[17]。等人提出了一种新的 LLM-KG 集成范式——将 LLM 作为一个代理，交互式地探索 kg 上的相关实体和关系，并基于检索到的知识进行推理。他们通过引入一种称为 Think-on-Graph （ToG） 的新方法实现了这种范式，从用户问题中提取出实体后，在知识图谱中搜索相关实体与关系链，即找出 Top-K 个实体间关联的路径。用 LLM 判断路径是否合理。在知识图谱内得到足够证据后送回 LLM，综合所有结果生成合理的答案。

（2）基于检索增强的问答系统研究

基于检索增强是指在用户提出问题时，利用外部知识库或搜索引擎等检索手段，找到与问题相关的信息，并将这些信息提供给 LLM，以增强其生成准确和相关答案的能力。检索增强的方法包括向量相似性搜索和实体链接。

Izacard[18]。等人提出了一种精心设计和预训练的检索增强语言模型 Atlas，能够以很少的训练示例学习知识密集型任务。Atlas 拥有两个子模型，一个检索器与一个语言模型。当面对一个任务时，Atlas 依据输入的问题使用检索器从大量语料中生成出最相关的 top-k 个文档，之后将这些文档与问题 query 一同放入语言模型之中，进而产生出所需的输出。

Liu[19]。等人提出了 RETA-LLM 的策略，LLM 可以跟据知识检索系统从外部语料库中检索到的相关内容作为参考，生成更多的事实文本来响应用户输入。此外，通过整合外部知识，检索增强 LLM 可以回答那些依靠存储在参数中的世界知识来回答的领域内的问题。从而缓解了 LLM 倾向于产生幻觉并生成对用户请求的虚构响应的问题。

Luo[20]。等人提出了新颖的参数化知识引导 PKG 框架，该框架为 LLM 配备了知识指导模块，可以在不改变 LLM 参数的情况下访问相关知识。跟据用户问题从外部知识库中获取知识文本，知识文本被添加在 prompt 中，作为背景知识供 LLM 参考，并联合问题输入生成最终的答案。

3. 项目已有的基础

在最近的一段时间里，本课题收集和阅读大量关于知识增强大模型领域的文献。通过深入了解这些文献，本课题得以把握当前该领域的研究进展，并掌握了一些新兴的自

然语言处理技术。这些技术为知识增强大模型的发展提供了强大的支持，使得我们能够更好地理解和处理大量的文本数据。

与此同时，本课题与本组的老师、同学们一起对凤凰出版社提供的中、小学课本和教辅资料进行了标注。这些资料被用作检索知识库，为我们的项目提供了宝贵的数据基础。通过标注这些资料，本课题实现了基于检索增强问答系统的可追溯性，以及重点内容的突出能力。

在这个过程中，本课题采用了多种标注方法，包括实体标注、关系标注等。这些标注方法有助于我们更好地理解文本中的实体和关系，从而为知识增强大模型的训练提供了更加准确的数据。

此外，本课题还对检索知识库进行了优化，提高了其检索效率和准确性。这使得本课题的知识增强大模型能够更好地为用户提供准确、有用的答案。

通过收集和阅读相关领域的文献，以及与老师、同学们一起进行标注和优化检索知识库的工作，本课题深入了解了知识增强大模型的研究进展和应用前景。同时，本课题也掌握了一些新兴的自然语言处理技术，为未来的研究工作打下了坚实的基础。

4. **与本项目有关的研究积累和已取得的成绩**

在过去的暑假中，课题组致力于实现一个基础版的科学教育领域大模型，并在这一过程中积累了丰富的研究经验和技能。通过采用先进的 LangChain 和 Toolformer 技术，成功地实现了对外部工具的调用与针对性优化，以增强模型的输出效果，有效提高了模型的准确性和全面性。在本项目中，课题组深入研究了 LangChain 和 Toolformer 的原理与应用，将它们有机地融入到科学教育领域的大型模型中，为后续研究和应用奠定了坚实基础。

5. **已具备的条件**

数据方面具有江苏凤凰教育出版社提供的高质量课本教参PDF数据, COIG、pCLUE、alpaca_chinese_dataset 在内的开源指令数据集与教育领域的数据集。硬件方面团队具有 8 张 A40 显卡与 8 张 4090 显卡。

6. 尚缺少的条件及方法

无

（二）**项目研究目标及主要内容**

基于邻域知识增强的科学教育问答系统设计，是以大语言模型为核心构建的针对科学教育领域专业知识，服务于 K12 教育阶段师生的科学式问答系统。该系统旨在通过问答的形式，培养学生的抽象思维和科学思维，从微观知识点出发，进行推理和归纳，培养逻辑思维和建设知识体系。为了解决目前存在的数据质量低，数据来源不可靠，不可溯源；大语言模型编造不存在的知识点；混淆不同教育版本之间的知识内容与要求；问答不能体现学科思维，只能被动回答而缺乏系统的引导对话等痛点，本课题研究目标如下：

1. **科学教育知识库系统构建：**本课题拟针对如何构建基于大语言模型的问答系统知识库进行研究。本课题拟通过实体发现或命名实体识别 NER、实体链接 EL、关系提取 RE、事件提取 EE 等相关算法，借助大语言模型将自然语言处理为格式化 JSON 文本，通过标准化 Schema 存储于基于 NebulaGraph 构建的知识图谱，最终使用 Streamlit 和 Docker 将我们的对话系统打包为一个 Web 软件进行发布。
2. **科学教育知识增强问答算法设计：**该目标旨在通过设计相关算法，实现问答系统的知识点溯源能力和重点内容突出的能力，解决科学教育问答的准确性和可溯源性。同时结合目标一构建的知识图谱，实现相关知识点推荐功能，帮助用户建立相应的知识体系。
3. **科学教育知识增强问答系统搭建**: 该目标旨在搭建一个科学教育领域的问答系统，并用知识增强的技术增强完善问答系统的输出，解决了项目实例化落地的问题。系统部署基于 B/S 进行，实现了动态数据更新以及与用户实时交互。为了更好的将研究实例化，本课题以 LLaMa2 为基座模型，中、小学的物化生课本和教辅资料作为训练语料，训练微调出一个中、小学科学教育领域的模型作为问答系统的基础模型。以检索增强的技术，增强问答系统的输出的可解释性、准确性等。以图谱增强的技术，实现问答系统的知识信息推荐和问答的能力。最终基于开源微框架 Streamlit 与 Docker 容器技术将模型部署上云，实现与用户的实时交互。

（三）项目创新特色概述

本项目具有以下创新特色：

采用知识图谱+检索的方式，进而增强问答系统的可解释性：在我们的项目中，我们引入了知识图谱与检索相结合的方式，以显著提升问答系统的可解释性。传统的问答系统在回答用户提问时通常基于模型的预测，用户往往难以理解系统的决策过程。通过引入知识图谱，我们将丰富的实体关系信息与检索技术相结合，使得系统生成的答案更具可解释性。

针对当前科学教育领域的知识图谱的缺失，结合中、小学科学教育领域不同学科的特点，设计并构建一个有一定数量规模且质量较好的科学教育领域知识图谱。

基于图谱增强的技术，通过提取用户提问和问答系统中的实体，查询科学教育领域的知识图谱，获取具有体系化的专业知识注入给问答系统，通过的多轮对话的方式使问答系统具有知识信息推荐和启发式问答的能力，使问答系统的输出更具体系化。

（四）项目研究技术路线

本课题拟分为三个部分：数据层面的知识图谱构建，算法层面的知识增强问答算法设计，系统层面的知识问答系统开发建设。下面是每个部分的技术路线：

1．科学教育知识图谱构建

本课题拟使用 NebulaGraph 来构建知识图谱数据库。然后使用 TransE 模型来进行知

识图谱嵌入，以便对知识图谱进行表示学习。最后使用 Neo4J 图数据库来存储和查询知识图谱数据。图 1 是构建科学教育知识图谱的技术路线。

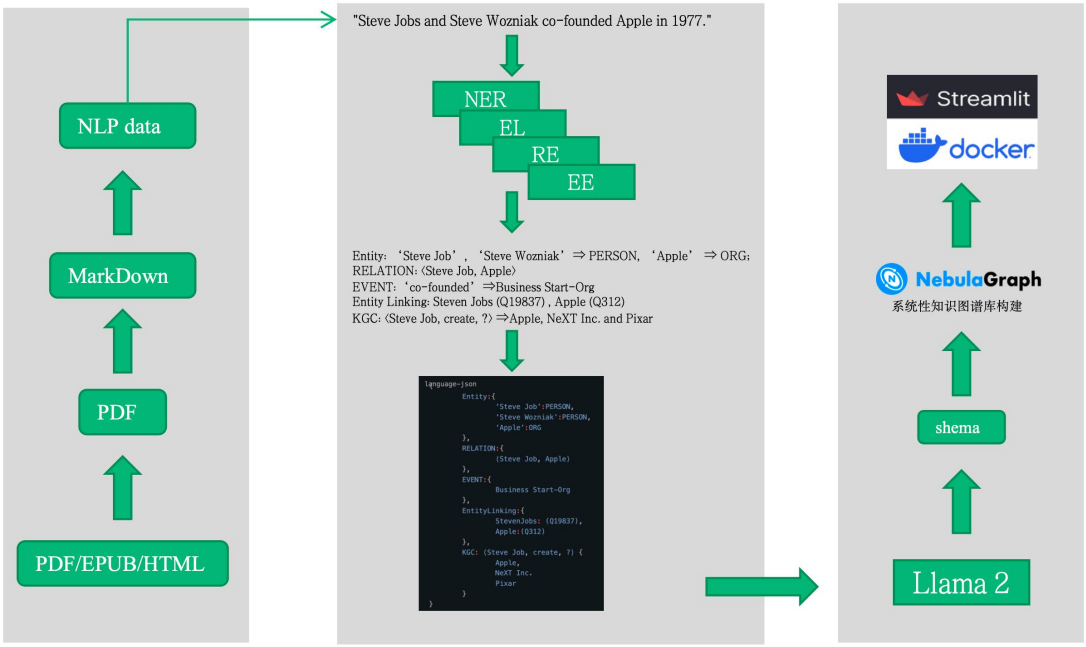


图 1：构建知识图谱技术路线

具体步骤细节如下：

1.1 数据的获取

本课题拟通过使用具有版权使用权的优质教材教辅的 PDF数据、网络开源的学科教育数据集、网络开源知识图谱、网络开源指令数据集等数据作为原始资料。通过 marker 等开源格式转换工具将 PDF、json、txt、markdown 等格式数据文件统一转换为 markdown 格式存储。

1.2 系统性知识图谱库构建

本课题拟通过 LLM 将自然语言格式的文本数据转换为格式化的图谱 I/O 语句，通过实体发现或命名实体识别 NER、实体链接 EL、关系提取 RE、事件提取 EE 等语义分析（Semantic Analysis）获得知识条例数据中的格式化成分，通过大语言模型将其转换成 JSON 格式的格式化文本，进一步转换成 NebulaGraph 的图查询语言 (Nebula Graph Query Language，NGQL）来执行数据导入、图查询和图分析操作，创建实体和关系的顶点和边，以及定义图模式和索引。以此开源分布式高性能图数据库 NebulaGraph 构建学科知识库。

1.3 知识图谱嵌入

如图 2 所示，本项目拟使用 Python 中的深度学习框架 PyTorch 来实现TransE 模型。使用知识图谱数据训练 TransE 模型，以学习实体和关系的向量表示。训练完成后，通过 TransE 将学习到的向量表示存储起来，以备后续在知识图谱数据库中使用。

1.4 教育知识图谱的构建

本课题拟根据不同学科之间的学科逻辑与学科素养要求，开发针对每个学科的模块化知识图谱子库。

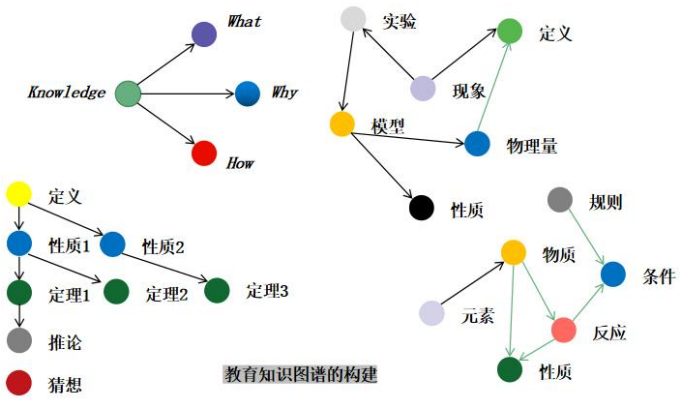


图 2：教育知识图谱的构建

1.5 知识图谱可追溯性化

本课题拟通过生成时溯源标签设置、节点索引化、相关性聚类、索引集合与溯源标签对应等步骤实现知识库中的知识来源溯源定位，实现知识的可追溯性（见图 3）。

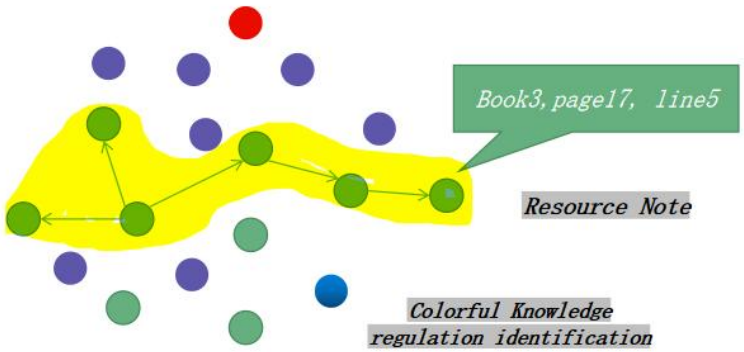


图 3：知识图谱可追溯性化

1.6 打包和发布

本课题拟将训练好的系统模型通过 Streamlit 开发成 Web 软件，并通过 Docker 容器挂载到服务器上发布。

1.7 数据检索与提示工程

本课题拟使用 Text2Cypher 来把自然语言的文本转换成 Cypher 查询语句，然后使用 schema 进行检索匹配，返回 JSON 格式化数据，最后通过知识条例提取（Knowledge Regulations Extraction）与相关性检索算法确定相关知识边界，将知识边界内的图谱结构化打包返回作为检索结果输出给大语言模型，最终将检索结果生成可视化前端交互图，

依照模板构造 Prompt，按照学科思维与逻辑展开与用户的引导性交互问答。

2. 科学教育知识增强问答算法设计

2.1 向量化索引检索知识库

本课题拟采用向量化索引检索知识库+构造 Prompt 的方法实现目标二中知识点溯源，重点知识标注两个子任务。这两个子任务都要求跟据用户输入或者模型生成的内容去检索知识库得到想要形式的内容，去增强、完善问答系统的输出，所以选择一个合适的检索算法是关键一步。现采用 LangChain 和 Toolformer 两种框架。

LangChain 和 Toolformer 两种框架，旨在帮助开发人员使用语言模型构建端到端的应用程序，可以帮助开发人员轻松建立知识库与 LLM 间的链接，将知识注入到 LLM 中。如下图 4 所示，使用 LangChain + LLM，生成更为完善的回答。通过微调或者 ICL 的方式，让模型学会从用户的输入或者模型的初始输出中提取出关键文本，基于 LangChain 在知识库中检索出与关键文本相关的知识，并与模型的初始输出跟据不同子任务构造特定的提示，让模型依据提示对模型初始的输出进行完善，最终得到回答文本。可以选用 LLaMa2、ChatGLM-6B 等作为大模型，本文采用 LLaMa2 作为基座模型。

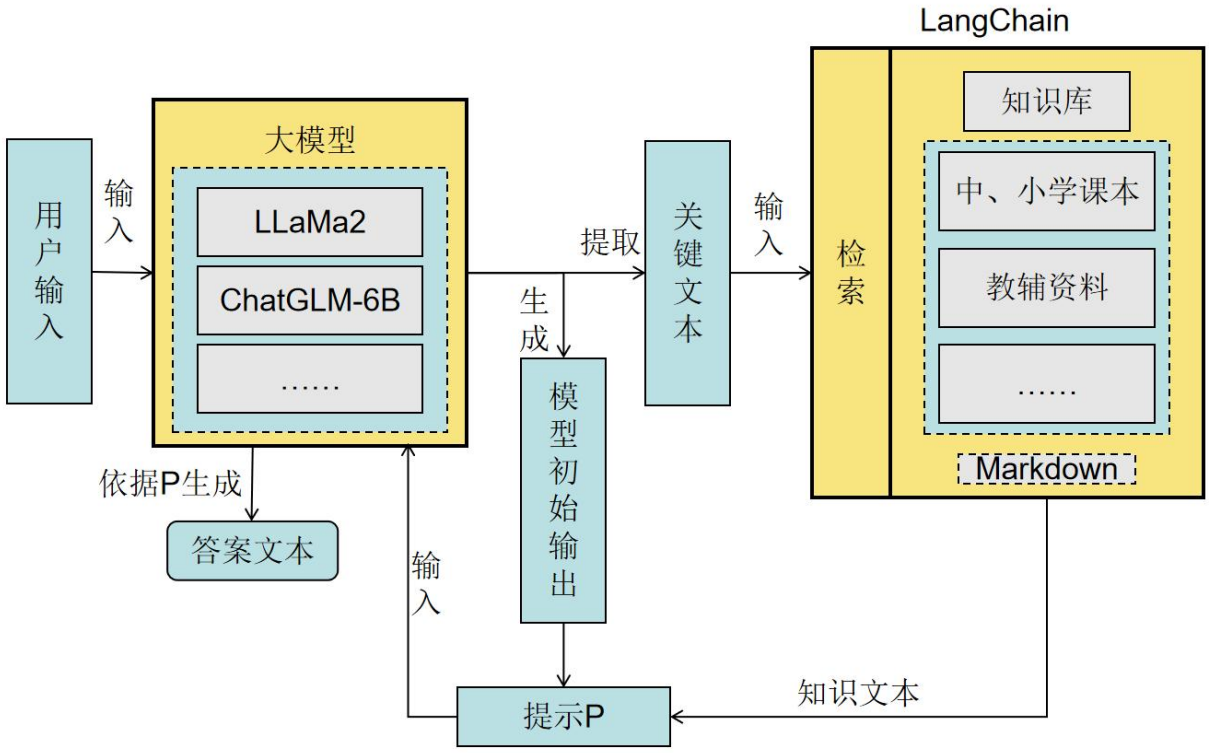


图 4 子任务 1,2 技术流程图

假设知识库中的第 i 个文件（1 个小 block 即为 1 个文件）为$F_i(i=1,2,\ldots,n)$。基于 LangChain 进行检索会将各个文件中的文本进行分块，$D_{ij}(i=1,2,\ldots,n;j=1,2,\ldots,m)$表示第个文件的第 j 个分块。然后对每一块文本建立向量索引$V_i(i=1,2,\ldots,n\times m)$，在检索时将提取出的关键词或者关键文本向量化，得到关键文本向量 Q，最后通过向量相似度计算出和 Q 最相似的 k 个向量索引，并返回对应的文本块（包括文本块对应的文件名以及整个文本内容），构成知识文本 D。将检索出的文本 D 和模型的初始输出以一定的形式拼接构成提示 P，输入给 LLM 对初始输出按照提示 P 进行修改，最终得到完善后的模型输出。

该过程的伪代码如下图 5 所示：

算法1 向量化索引检索问答
输入：通过微调或者少样本的ICL让模型学会从问题文本或者模型的初始输出中提取到的关键文本q，知识库文件f。
输出：LLM的回答文本result。
1. for $i: = 1\ to\ N\ do$
2. $d_i \leftarrow split(f_i)$ //第i个文件划分文本块
3. $end\ for$
4. $for\ i: = 1\ to\ N \times M\ do$
5. $V_i \leftarrow trans(d_i)$ //生成每个文本块的向量索引
6. $end\ for$
7. $Q \leftarrow trans(q)$ //将关键文本转化为关键文本向量
8. $V_k \leftarrow score(Q, V_{n\times m})$ //计算关键文本向量和索引向量的相似度，得到k个最相似的索引向量
9. $d_k \leftarrow de\_trans(V_k)$ //根据匹配到的k个文本向量转化为相应的知识文本
10. $result \leftarrow model(P(q, d_k))$ //将得到的相应的知识文本和模型的初始回答构造成提示P的形式输入给LLM以获取最终的回答

图 5 向量化索引检索问答伪代码

算法中 q 表示 LLM 从用户的提问或者模型的初始回答中提取出的关键文本，f 表示知识库文件，d 表示知识文本块，Q 表示关键文本向量，V 表示文本块的向量索引，split 表示划分文本块的过程，trans 表示从文本转化为向量,detrans 表示从向量转化为文本，score 将返回 k 个最相似的向量索引,需要进行一些文本处理, 为表示得到的相应的知识文本和模型的初始回答构造成提示 P 的形式输入给大模型 LLaMa2。

如下表 1 所示，跟据不同的子任务，构造不同的提示 P。

表 1 基于检索增强的提示示例

子任务	知识文本	提示文本
知识点溯源	知识文本所在的文件名称（即知识点在教辅中的具	{知识文本}是（模型初始输出）内容的来源，请复

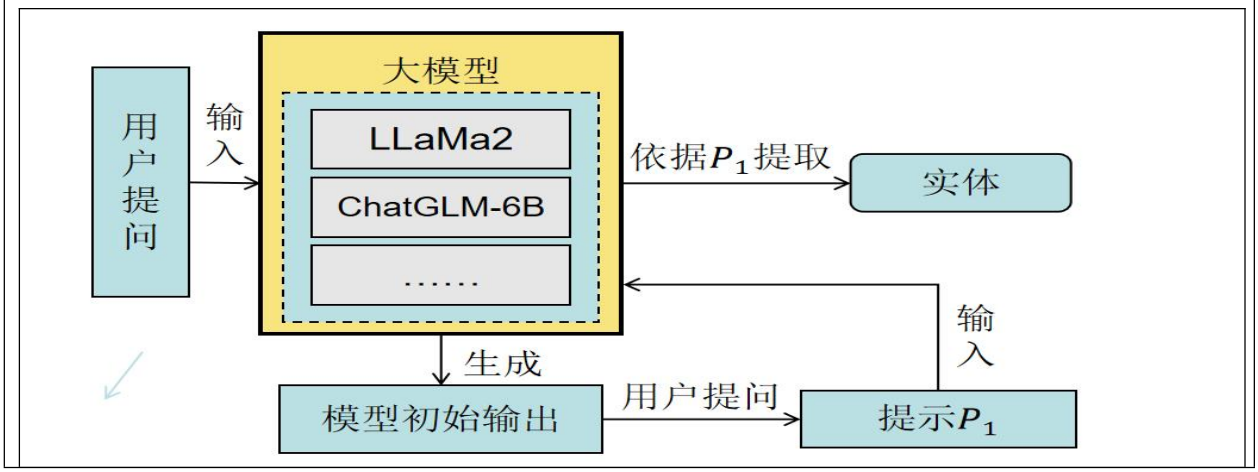
	体来源)	述(模型初始输出），并在恰当的地方插入{知识文本}。
重点知识点标记	带有标记的知识文本（“<u> </u>”表示带下划线的内容，“** **”表示加粗的内容)	请跟据文本{知识文本}中的标记内容（“** **”或“<u> </u>”），找出（模型初始输出）中与其表达意义相同的内容，并在这些内容前后加上“** **”进行标记，最后输出标记后的（模型初始输出）。

注：“{}”内的内容是检索得到并进行一定处理后的知识文本，“（）”内的内容是模型的初始输出。

2.2 知识图谱推荐算法

本课题组拟采用 KG 推荐+Prompt，实现目标二的相关知识点推荐的子任务。

如下图 6 所示, 构造提示 P1 让模型去提取用户的提问模型初始输出中涉及到的实体作为查询实体，并将查询实体与目标一构建的知识图谱进行匹配，按照直接或间接相连的邻居实体找到关联知识点。然后基于 Network 框架，根据关系权重，路径信息，与学科重点知识等设计相应的相关度评估方法。将相关度高的知识点结合模型的初始输出构造不同的提示 P2，输入给模型，让模型依据提示 P2 实现知识信息推荐。



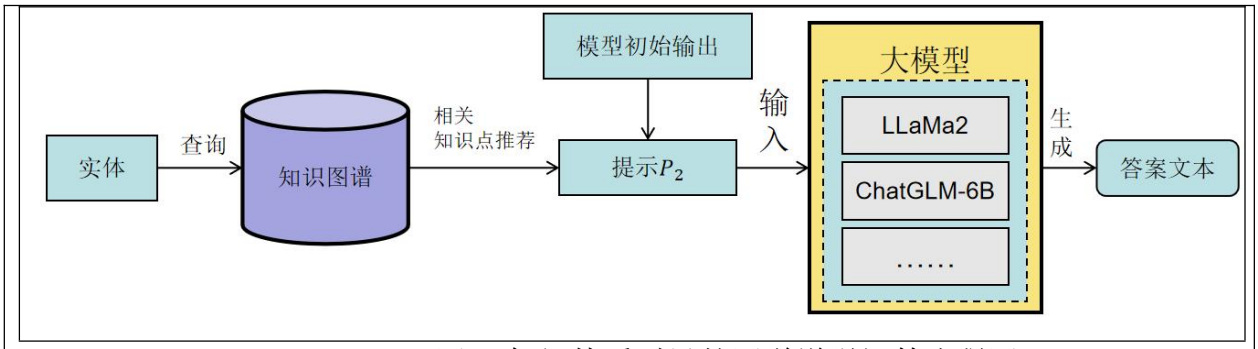


图 6 知识体系引导的图谱增强问答流程图

如下表 2 所示，为提示 P1 和提示 P2 的两个示例：

表 2 基于图谱增强的提示示例

提示	子任务	提示文本
P1	提取涉及到的实体	{模型初始输出}是对于{用户提问}的回答，请提取两者涉及到的实体，并以一个集合的形式将实体输出，例如{浓硫酸具有腐蚀性吗}和{具有，浓硫酸同时具有脱水性}分别为用户提问和模型初始输出，我希望提取的实体集合为{浓硫酸，腐蚀性，脱水性}。
P2	基于图谱的知识信息推荐	给定一个{相关度高的知识点}，以及一段文本（模型的初始输出）。复述文本的同时，询问用户是否想了解{相关知识点}信息。

3. 科学教育知识增强问答系统搭建

如图 7 所示，科学教育知识增强问答系统主要包括以下三个模块：提取模块、检索模块以及增强输出模块。

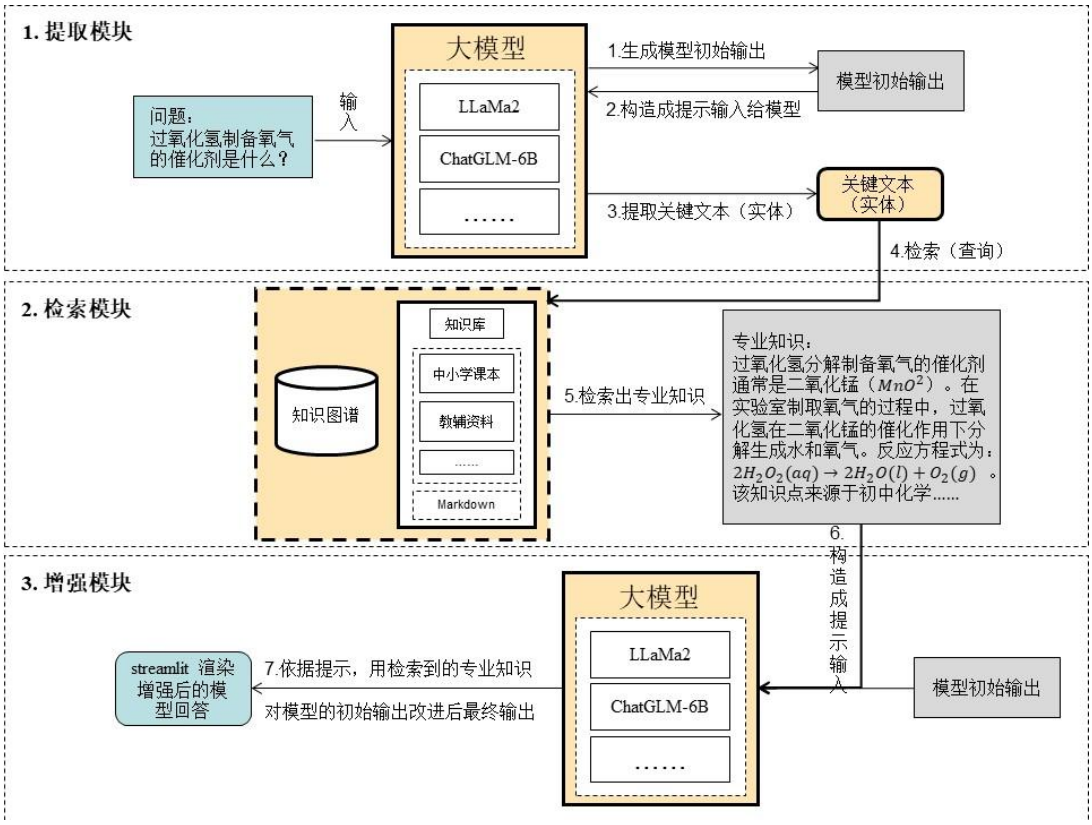


图 7 科学教育知识增强问答系统

3.1 提取模块

主要通过构造提示的方式，让问答系统对用户的提问和问答系统的初始输出中涉及到的关键文本（实体）进行抽取。

3.2 检索模块

根据抽取出的关键文本（实体），查询知识库与图谱，得到与之相关的专业知识形成体系。

3.3 增强模块

将检索到的专业知识和问答系统的初始输出构造成提示，输入给问答系统，让问答系统依据专业知识的内容，按照提示的要求，对问答系统的初始输出进行改进。从而增强问答系统最终的回答。

3.4 最终对于输出流

通过 streamlit 进行前端渲染，将知识点以 markdown 格式呈现，知识体系通过思维导图 graphviz_chart 的形式呈现，思维导图中的每一个结点可以作为新输入迭代读入。

4. 参考文献

[1] Yang J, Jin H, Tang R, et al. Harnessing the power of llms in practice: A survey on chatgpt and beyond[J]. arXiv preprint arXiv:2304.13712, 2023.

[2] Wei J, Tay Y, Bommasani R, et al. Emergent Abilities of Large Language Models[J]. Transactions on Machine Learning Research, 2022.

[3] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[4] OpenAI. ChatGPT. [EB/OL]. 2022. https://openai.com/blog/ChatGPT/.

[5] OPENAI. GPT-4 technical report[R]. arXiv:2303.08774, 2023.

[6] Wang Y, Kordi Y, Mishra S, et al. Self-instruct: Aligning language model with self generated instructions[J]. arXiv preprint arXiv:2212.10560, 2022.

[7] Malinka K, Peresíni M, Firc A, et al. On the educational impact of ChatGPT: Is Artificial Intelligence ready to obtain a university degree?[C]//Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1. 2023 .

[8] Li Z, Wang C, Liu Z, et al. Cctest: Testing and repairing code completion systems[C]//IEEE/ACM 45th International Conference on Software Engineering (ICSE). 2023.

[9] Liu J, Liu C, Lv R, et al. Is chatgpt a good recommender? a preliminary study[J]. arXiv preprint arXiv:2304.10149, 2023.

[10] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023.

[11] Wang X, Wei J, Schuurmans D, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models[J]. 2023.

[12] Golovneva O, Chen M P, Poff S, et al. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning[C]//The Eleventh International Conference on Learning Representations. 2022.

[13] Toneva M, Sordoni A, des Combes R T, et al. An Empirical Study of Example Forgetting during Deep Neural Network Learning[C]//International Conference on Learning Representations. 2018.

[14] Rony M R A H, Usbeck R, Lehmann J. DialoKG: Knowledge-Structure Aware Task-Oriented Dialogue Generation[C]//Findings of the Association for Computational Linguistics: NAACL. 2022.

[15] Sun Y, Shi Q, Qi L, et al. JointLK: Joint Reasoning with Language Models and Knowledge Graphs for Commonsense Question Answering[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022.

[16] Yu D, Zhu C, Yang Y, et al. Jaket: Joint pre-training of knowledge graph and language understanding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022.

[17] Sun J, Xu C, Tang L, et al. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph[J]. arXiv preprint arXiv:2307.07697, 2023.

[18] Izacard G, Lewis P, Lomeli M, et al. Few-shot learning with retrieval augmented language models[J]. arXiv preprint arXiv:2208.03299, 2022.

[19] Liu J, Jin J, Wang Z, et al. RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit[J]. arXiv preprint arXiv:2306.05212, 2023.

[20] Luo Z, Xu C, Zhao P, et al. Augmented Large Language Models with Parametric Knowledge Guiding[J]. arXiv preprint arXiv:2305.04757, 2023.

（五）研究进度安排

开始时间	结束时间	工作内容
2023-12	2024-01	课题调研，信息检索，论文综述
2024-01	2024-03	技术栈学习
2024-04	2024-08	按照技术路线实现最小可用产品
2024-09	2024-12	在最小可用产品的基础上拓展功能
2024-12	2025-02	前端美化，交互优化
2024-02	2024-04	调试，打包，发布

（六）项目组成员分工

1. 课题相关论文检索，阅读与综述：时子延，董文杰，马艺轩
2. 技术栈调研：时子延，董文杰，马艺轩
3. 技术栈学习分工
   (1) streamlit：董文杰
   (2) docker：董文杰，时子延
   (3) Python：时子延，董文杰，马艺轩
   (4) TransE：时子延
4. Github 仓库维护：时子延
5. 产品手册维护：董文杰
6. 开发日志维护：马艺轩

三、学院提供条件

1. 周俊生老师，张博，孔力，刘海峰，郑智超等老师在研究方向、理论分析、技术实现等方面给予了充分的指导。
2. 校方采购订阅的数据库学术资料丰富、网络信息渠道畅通。
3. 学院和组内具备相应的硬件条件。学院配有服务器资源，便于更好地进行实验。

四、预期成果

1. 构造一个针对 K12 教育的知识图谱数据库，实现知识图谱的前端可视化展示。
2. 在知识图谱知识库基础上使大模型系统具备知识点溯源，重点知识标注，关联知识点推荐功能，且将系统部署上云并实现模型与用户的交互逻辑。
3. 申请软件著作权 1 至 2 项。
4. 发表国际会议或期刊论文 1 篇。

五、经费预算

总经费（元）

注：总经费、财政拨款、学校拨款按照规定金额填写，校企合作项目企业资助金额不少于 5000 元。	**3000**	**财政拨款/企业资助（元）**	**0**	**学校拨款（元）**	**3000**

具体包括：

1、调研、差旅费；
2、用于项目研发的元器件、软硬件测试、小型硬件购置费等；
3、资料购置、打印、复印、印刷等费用；
4、学生撰写与项目有关的论文版面费、申请专利费等。

六、导师推荐意见

签名：

年　月　日

七、院系推荐意见

院系负责人签名:　　学院盖章:

年　月　日

八、学校推荐意见:

学校负责人签名:　　学校公章

年　月　日