大学生创新创业训练计划项目任务书

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 | | | | | | | | | |
| 项目管理  学院 | | | 计算机与电子信息学院/人工智能学院 | | | 指导老师 | | 周俊生 | | 工作单位 | | 计算机与电子信息学院/人工智能学院 |
| 项目实施时间 | | | 起始时间：2023年 12 月 完成时间： 2025 年 6 月 | | | | | | | | | |
| 项目级别 | | | □国家级（省级重点）  ☑省级（含省级一般、省级指导）  □校级 | | | | | | | | | |
| 团队成员 |  | 姓名 | | 年级 | 学院 | | 专业 | | 联系电话 | | E-mail | |
| 主持人 | 时子延 | | 22 | 计算机与电子信息学院/人工智能学院 | | 人工智能 | | 18351997562 | | szy@nnu.edu.cn | |
| 成 员 | 董文杰 | | 22 | 计算机与电子信息学院/人工智能学院 | | 人工智能 | | 19805119351 | | 2785794768@qq.com | |
| 马艺轩 | | 22 | 计算机与电子信息学院/人工智能学院 | | 人工智能 | | 15996880830 | | 1023006192@qq.com | |
|  | |  |  | |  | |  | |  | |
| **一、项目简介**(限200字）  随着人工智能技术的快速发展，基于自然语言处理（NLP）的大模型在教育领域的应用展现了巨大的潜力。尤其是面向中小学科学教育的智能教学助手，能够帮助学生解决疑难问题并提升学习效率。然而，现有的大模型在面对中小学科学教育场景时，存在诸多挑战，例如缺乏领域知识的充分融入、生成答案的准确性与可解释性不高，以及对于复杂连续性问题的理解能力不足。  本项目旨在研究并实现面向中小学科学教育大模型的领域知识增强方法，通过构建科学教育领域知识图谱（TechEdu KG）和改进检索增强生成框架（Advanced RAG），提升大模型在科学教育领域的应答能力，为中小学科学教育提供更精准、高效的知识服务。 | | | | | | | | | | | | |
| **二、项目研究内容**（即研究任务，要求分项表述）  本研究的目标是通过领域知识图谱（TechEdu KG）的构建与应用，优化基于RAG框架的生成方法，并开发出适用于中小学科学教育场景的智能问答系统。具体来说，研究目标包括以下几点：   1. TechEdu KG 构建：通过开源PDF转换工具结合人工精细微调的方式构建科学教育领域知识图谱，实现从课本PDF数据到知识图谱三元组的转换。 2. Advanced RAG 优化：改进原始检索增强生成框架，增加用户请求模块、内容提取模块和事实验证模块，提升信息检索的准确性与完整性。 3. Faiss on KG Subgraphs：提出以知识图谱子图作为检索单位的检索方法，在保证检索准确性的基础上大幅提升检索速度，进而有效降低了模型应答的延迟，提升了用户体验。 4. Benchmark 构建：通过大模型生成结合人工精细筛选的方式构建科学教育领域的问答基准对。 5. Streamlit Web界面：利用专为数据可视化打造的Python框架Streamlit，基于Browser/Server服务架构，快速搭建一个端到端的研究落地平台。便于收集内测用户的反馈并及时调整，为今后的应用推广打好坚实基础。 | | | | | | | | | | | | |
| **三、项目实施方案**（包含具体的研究方案、研究方法、技术路线等）  1. **研究方法**：   * **深度学习**：利用深度学习模型（如 GLM-4、BERT、T5）进行文本处理、信息抽取、知识图谱三元组生成 * **知识图谱**：结合相关文献，构建科学教育领域知识图谱，并用COMET模型补全。 * **RAG**:基于fassis检索与RAG框架优化，改进原有知识检索增强。 * **实验和评估**：通过实验验证方法的有效性，并通过构建benchmark使用评估指标进行性能评估。  1. **技术路线**:   2.1 **文本处理**：   * 利用 GLM-4、BERT 等预训练模型进行文本表示和语义理解。 * 使用句法分析技术进行句子成分分析和依存关系识别。 * 应用命名实体识别技术识别文本中的实体信息。   2.2 **信息抽取**：   * 利用 T5，GLM-4 等预训练模型进行关系抽取，识别实体之间的关系。 * 使用知识图谱嵌入技术将实体和关系映射到低维向量空间。   2.3**知识图谱构建**：   * 基于信息抽取结果，构建科学教育领域知识图谱。 * 使用图神经网络（COMET）等技术进行知识图谱的推理和扩展。   2.4**问答生成：**   * 利用 RAG 框架，结合知识图谱和预训练模型进行问答生成。 * 通过重写用户请求模块、内容提取模块和事实验证模块，提升问答的准确性和可靠性。 * 服务界面：基于streamlit搭建web界面并部署。   2.5**研究方案:**  2.5.1第一阶段：知识图谱构建   * 开发 PDF 数据到知识图谱三元组的自动化转换脚本。 * 使用 GLM-4+Prompt 自动提取高质量三元组。 * 基于 BERT 和 IE 模型构建三元组生成脚本。 * 进行知识图谱的可溯化和路径分析。   2.5.2第二阶段：RAG 优化   * 分析 Naïve RAG 的局限性，并提出改进方案。 * 增加 用户请求模块、内容提取模块和事实验证模块。 * 进行实验验证和性能评估。   2.5.3第三阶段：检索方法研究   * 研究 Faiss on KG Subgraphs 检索方法。 * 基于 DBSCAN 算法进行知识图谱子图划分。 * 进行实验验证和性能评估。   2.5.4第四阶段：基准测试（benchmark）构建   * 构建科学教育领域基准测试。 * 选择合适的评估指标（如忠实度、答案正确性）。 * 进行实验验证和性能评估。   2.5.5第五阶段：网站搭建   * Streamlit基本框架 * 导入TechEduKG，并可视化显示 * 知识推理问答界面。 * 尝试引入邻域知识问答系统生成的工作流 | | | | | | | | | | | | |
| **四、项目进度安排（**包含项目实施步骤、阶段性目标等）   * 第一阶段（1-3 个月）：完成 TechEdu KG 的构建和可溯化分析。 * 第二阶段（4-6 个月）：完成 Advanced RAG 的优化和 Faiss on KG Subgraphs 的研究。 * 第三阶段（7-9 个月）：完成 Benchmark 构建，并进行实验和评估。 * 第四阶段（10-12 个月）：进行项目总结和成果展示。 | | | | | | | | | | | | |
| **五、项目预期成果（**如论文、专利、作品、产品、研究报告等，应量化、具有可考核性）   * 专利：申请 1-2 项发明专利。 * 作品：开发一个面向中小学科学教育的大模型问答系统原型。网站Web应用程序 * 研究报告：撰写项目研究报告。 | | | | | | | | | | | | |
| **六、项目创新点**   * TechEdu KG 的自动化构建：实现从课本 PDF 数据到知识图谱三元组的自动化转换，提高构建效率。 * Advanced RAG 的优化：改进原始检索增强生成框架，提升大模型对复杂提问的理解和回答能力。 * Faiss on KG Subgraphs：提出以知识图谱子图作为检索单位的检索方法，提高检索效率和准确率。 | | | | | | | | | | | | |
| **经费使用计划**（各项开支的百分比）   1. **打印 ￥50** 2. **阿里云服务器 ￥500** 3. **域名 ￥20** 4. **智谱AI API ￥100** 5. **专利申请费用 ￥2000** | | | | | | | | | | | | |
| **指导教师意见**  **{93484D66-D01E-4FEC-BFBC-92A18EB6EF0E}**  签名  年 月 日 | | | | | | | | | | | | |
| **所在学院意见**  盖 章：  年 月 日 | | | | | | | | | | | | |
| **主管部门意见**  盖 章：  年 月 日 | | | | | | | | | | | | |

注：表格栏高不够可增加。