**南 京 师 范 大 学**

**毕 业 设 计（论 文）**

**（2024届）**



|  |  |  |  |
| --- | --- | --- | --- |
| **题 目：** | 融合知识图谱的科学教育大模型  推理方法研究与实现 | | |
| **学 院：** | 计算机与电子信息学院/人工智能学院 | | |
| **专 业：** | 计算机科学与技术 | | |
| **姓 名：** | 陈昱衡 | | |
| **学 号：** |  | 19200322 |  |
| **指导教师：** | 周俊生 | | |

**南京师范大学教务处 制**

# 摘 要

大模型通过深度学习和自然语言处理技术来满足人们对快速、准确、便捷获取知识和信息的需求。这些模型具备语言理解和生成的能力，能够处理用户输入并生成有意义的输出，包括回答问题、解释概念、提供建议等。然而，现有的大模型技术一般存在知识性幻觉现象，在特定垂直领域的应用中，这可能导致准确性缺失，使得生成的答案与实际事实不相符。此外，模型自身的知识完全源于它的训练数据，现有主流的大模型的训练集基本都是构建于网络公开的数据，而由于其参数多，训练成本较大，使其无法获取一些具有时效性的数据。

因此，本文提出了一种融合知识图谱的科学教育大模型推理方法，通过利用知识图谱中的结构化数据，对于用户向大模型提出的问题，在知识图谱中检索相关的信息作为参考，并将检索到的信息和问题注入到大模型的提示中，是大模型生成对应的答案。我们结合清华大学公开的知识图谱EDUKG和开源中文知识图谱ownthink来构建自己的科学教育知识图谱作为检索的数据源。具体来讲，本文的主要工作包括以下几个方面：

（1）本文对大模型表现的不足之处以及对应的优化技术进行了梳理，对于构造科学教育大模型任务要求进行分析，确定以知识图谱为外部知识库的检索增强生成的技术框架。

（2）本文对现有的检索技术进行调研，针对知识图谱这种结构化数据的特点，提出了基于DBSCAN算法的层次聚类知识图谱子图划分算法，和以子图作为检索单位的Faiss相似度检索方式。同时，本文分别在预检索处理，后检索处理方面对于基础的检索增强生成框架进行改进，以提高其检索质量和生成质量。

（3）本文针对检索模块和检索增强生成框架结构这两个方面，构造了合理的对照组，使用RAGAs框架对提出的基于知识图谱子图的检索增强生成框架进行评价。实验结果显示该框架在忠实度和答案正确性两个指标下均优于其他框架。

**关键词：**大模型，检索增强生成，知识图谱

# Abstract

Large-scale models utilize deep learning and natural language processing techniques to meet people's demands for rapid, accurate, and convenient access to knowledge and information. These models possess the ability for language comprehension and generation, enabling them to process user inputs and generate meaningful outputs, including answering questions, explaining concepts, and providing recommendations. However, existing large-scale model technologies often suffer from the phenomenon of knowledge hallucination. In specific vertical applications, this may lead to inaccuracies, resulting in generated answers that do not align with actual facts. Additionally, the knowledge within the model is entirely derived from its training data, which is typically constructed from publicly available data on the web. Due to their numerous parameters and high training costs, mainstream large-scale models cannot access some time-sensitive data.

Therefore, this paper proposes a scientific education large-scale model reasoning approach that integrates knowledge graphs. By leveraging structured data from knowledge graphs, when users pose questions to the large-scale model, relevant information is retrieved from the knowledge graph as a reference, and this retrieved information along with the question is injected into the prompts of the large-scale model to generate corresponding answers. We construct our own scientific education knowledge graph as the data source for retrieval, combining the publicly available EDUKG from Tsinghua University and the open-source Chinese knowledge graph, OwnThink.

Specifically, the main contributions of this paper include the following aspects:

(1) We review the shortcomings of large-scale models and corresponding optimization techniques, analyze the requirements for constructing scientific education large-scale model tasks, and determine a technical framework for retrieval-augmented generation with knowledge graphs as external knowledge bases.

(2) This paper surveys existing retrieval technologies and proposes a hierarchical clustering algorithm for knowledge graph subgraphs based on the DBSCAN algorithm, tailored to the characteristics of structured data in knowledge graphs. The Faiss similarity retrieval method, which uses knowledge graph subgraphs as the retrieval units, is applied.

(3) To enhance the basic retrieval augmentation framework, this paper makes improvements in both pre-retrieval processing and post-retrieval processing to enhance the quality of retrieval and generation.

**Key words:** large-scale model, retrieval-augmented generation, knowledge graph

**目 录**

[摘 要 I](#_Toc165795838)

[Abstract II](#_Toc165795839)

[第1章 绪论 1](#_Toc165795840)

[1.1 研究目的及研究意义 1](#_Toc165795841)

[1.2 国内外的研究现状 2](#_Toc165795842)

[1.2.1 知识图谱概述 2](#_Toc165795843)

[1.2.2 大语言模型概述 3](#_Toc165795844)

[1.2.3 基于知识图谱的大语言模型推理方法研究 3](#_Toc165795845)

[1.3 研究内容 5](#_Toc165795846)

[1.3.1 科学类课程知识图谱的构建 6](#_Toc165795847)

[1.3.2 知识图谱辅助大模型推理方法 7](#_Toc165795848)

[1.3.3 科学类课程知识问答系统 7](#_Toc165795849)

[第2章 相关理论与技术 8](#_Toc165795850)

[2.1 聚类方法 8](#_Toc165795851)

[2.1.1 DBSCAN算法 8](#_Toc165795852)

[2.2 检索方法 9](#_Toc165795853)

[2.1.1 Faiss检索 9](#_Toc165795854)

[2.2.2 Lucene检索 10](#_Toc165795855)

[2.2 大模型检索增强生成方法 11](#_Toc165795856)

[2.2.1 概览 11](#_Toc165795857)

[2.2.2 发展过程 12](#_Toc165795858)

[第3章 融合知识图谱的科学教育大模型推理方法 14](#_Toc165795859)

[3.1 引言 14](#_Toc165795860)

[3.2 重写用户请求 15](#_Toc165795861)

[3.3 检索知识图谱 16](#_Toc165795862)

[3.4 背景内容提取 18](#_Toc165795863)

[3.5 答案生成与事实验证 19](#_Toc165795864)

[3.6 系统实现 20](#_Toc165795865)

[第4章 实验设计和结果分析 22](#_Toc165795866)

[4.1 实验数据和评价方法 22](#_Toc165795867)

[4.1.1 知识图谱构造 22](#_Toc165795868)

[4.1.2数据集 22](#_Toc165795869)

[4.1.3 评价方法 22](#_Toc165795870)

[4.2 实验和结果分析 24](#_Toc165795873)

[4.2.1 实验设置 24](#_Toc165795874)

[4.2.2 实验结果分析 25](#_Toc165795875)

[4.2.3 案例学习 27](#_Toc165795878)

[4.3 实验结论 30](#_Toc165795881)

[第五章 总结与展望 32](#_Toc165795882)

[5.1 总结 32](#_Toc165795883)

[5.2 展望 32](#_Toc165795884)

[参考文献 33](#_Toc165795885)

[致 谢 36](#_Toc165795886)

[本科期间主要研究成果 37](#_Toc165795887)

# 第1章 绪论

## 1.1 研究目的及研究意义

研究目的：

科技不断发展，改变了我们的生活、工作和学习方式。在信息爆炸的时代，人们对获取知识和信息的需求越来越迫切。大模型通过深度学习和自然语言处理技术来满足人们对快速、准确、便捷获取知识和信息的需求。这些模型具备语言理解和生成的能力，能够处理用户输入并生成有意义的输出，包括回答问题、解释概念、提供建议等。通过不断学习新数据，大模型能够保持更新，以适应不断变化的信息和知识需求，提供用户一种更加智能、灵活且高效的方式来获取知识和信息。科学教育大语言模型利用强大的语义理解和知识推理能力，不仅能解答学生问题，还为教师提供教学资源，推动智慧教育的发展，为师生提供便捷、高效的服务，促进科学教育事业的进步。

本课题的研究目的是研究和实现一种融合知识图谱的科学教育大模型推理方法。具体而言，课题首先要实现相关知识图谱的构建，重点针对科学类课程进行知识库的建设。随后，课题将探讨如何利用和融合教育知识图谱知识库，以改进大模型的推理过程，从而实现更准确的教育大模型推理效果。

本课题旨在为科学教育领域提供先进的大模型推理方法，为中小学科学教育的教学与学习提供更智能、个性化的支持。在实现这一目标的过程中，将深入研究知识图谱的构建和应用，以及如何将教育知识图谱知识库有机融入大模型推理过程。这将有助于提高科学教育的质量，推动教育技术在课堂上的创新应用，为培养具有科学素养的学生做出积极贡献。

研究意义：

目前，大语言模型在自然语言处理、编程辅助、医学、教育、创意生成、金融等领域取得了广泛的应用。这些模型能够执行多样化的任务，如文本生成、问答系统、医学诊断等。这些大语言模型的成功得益于深度学习和神经网络技术的进步，以及更强大的计算能力。它们通过在大量文本数据上进行训练，学习了语言的结构、语法和语义。这使得它们能够生成高质量的文本，模拟人类的语言能力。

然而，鉴于目前存在的大型语言模型在知识性幻觉方面的问题，其局限性在于仅能够记忆训练语料中的事实和知识，甚至无法回忆起这些知识而产生与事实不符的内容。在特定垂直领域的应用中，这可能导致准确性缺失，使得生成的答案与实际事实不相符。尤其是在科学教育领域，这种错误可能对学生的理解和学科知识的正确传递造成不良影响。

这一问题的严重性在于，误导性的信息可能对学生的学术进展和对科学的信任产生负面影响。学生可能会在接受虚假信息的情况下形成错误的观点，从而影响其对科学事实和概念的理解。因此，在解决语言模型的知识性幻觉问题方面，需要采取更为创新和全面的方法，以确保在教育过程中提供准确、可靠且更新的信息。

本课题采用构建知识图谱的方法，旨在解决大模型的知识性幻觉问题。知识图谱是一种将实体、关系和属性组织成图状结构的知识表示方式。通过迭代更新知识图谱，定期反映实际世界的变化，确保知识图谱知识库中的知识保持最新。在大模型的模型推理阶段，本课题引入知识图谱知识库作为外部知识库，根据用户问题从外部知识库中获取知识文本，使大模型避免生成与实际事实不符的答案，从而解决知识性幻觉这一问题。

此外，在科学教育领域，对于模型输出答案的准确性往往有着较高的要求。而本课题根据教材和教辅资料构建的知识图谱知识库可以保证知识点的准确性，避免出现一些知识性错误。教育领域的知识往往会随着教学大纲的变更而改变，而大模型本身无法适应这种变化。例如，当学生查询某个知识点时，如果知识点内容的侧重点随着教学大纲发生变化，大模型可能会给出偏离教学大纲的回答。这时，构建知识图谱的技术就能使大模型结合检索最新知识库得到的知识点内容，生成更为准确的输出。

## 1.2 国内外的研究现状

### 1.2.1 知识图谱概述

知识图谱（Knowledge Graph）的概念最早由Google 公司在 2012 年提出，其将知识图谱定义为用于增强搜索引擎功能的辅助知识库。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KG={(h ,r ,t)⊆E×R×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。

### 1.2.2 大语言模型概述

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[1]。近年来，随着模型的参数量不断增加，使得大语言模型表现出涌现能力[2]。像InstructGPT[3]、ChatGPT[4]、GPT4[5]这些自回归大语言模型[6]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[7]、代码生成[8]、推荐[9]等复杂的实际任务中显示出巨大的潜力。

尽管这些模型有着广泛的应用，但它们仍存在局限性：一方面，大语言模型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[10]。（2）另一方面，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[11]来解释他们的预测，但他们的推理解释也存在幻觉的问题[12]。

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[13]，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。

### 1.2.3 基于知识图谱的大语言模型推理方法研究

推理是人工智能的一个重要方面，其应用涵盖各个领域，例如问题解决、定理证明、决策和机器人技术[14]。自然语言推理是整合多种知识以得出关于世界的一些新结论的过程[15]。大模型在推理任务中也发挥着关键作用，通过学习大量数据，它们能够理解语境、关联信息，并进行有逻辑的推理。知识图谱允许研究人员构建领域特定的图谱，从而使模型更好地适应特定领域的推理任务，也能进一步提高推理的准确性。

Lu[16]等人认为现有的方法大多数将外部知识集成模块与改进的预训练损失相结合，并在大规模语料库上重新实现预训练过程。重新实现预训练通常非常消耗资源，并且很难适用于其他领域的知识图谱。对此，他们提出了一种基于微调过程的新型知识感知语言模型框架，该框架为预训练模型提供了一个统一的知识增强文本图，该文本图包含从知识图谱中提取的文本和多关系子图。此外，他们设计了一种基于层次关系图的消息传递机制，它允许注入的知识图谱和文本的表示相互更新，并可以动态地选择共享相同文本的模糊提及实体。

Sun[17]等人提出了情境化语言和知识嵌入（Contextualized Language and Knowledge Embedding，CoLAKE），它将语言和知识的情境化表示与扩展的掩码模型（Masked Language Model，MLM）目标相结合。CoLAKE不是只注入实体嵌入，而是从大规模知识库中提取实体的知识上下文。为了处理知识语境和语言语境的异构性，Sun等人将它们整合到一个统一的数据结构中，即词知识图谱（WK graph）。CoLAKE 使用改进的 Transformer 编码器在大规模 WK 图上进行了预训练。在知识驱动任务、知识探索任务和语言理解任务进行实验，并取得了令人惊讶的高性能。

Rony[18]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[19]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。具体而言，JointLK通过一种新颖的密集双向注意力模块在LM和GNN之间进行联合推理，其中每个问题标记都出现在知识图谱节点上，每个知识图谱节点出现在问题令牌上，并且两个模态表示通过多步交互相互融合和更新。然后，动态剪枝模块利用联合推理产生的注意力权重，递归地剪枝无关的知识图谱节点。

Wang[20]等人提出了K-Adapte框架，它保留了预训练模型的原始参数，并支持开发通用的知识注入模型。以RoBERTa为骨干模型，K-Adapter为每种注入的知识都有一个神经适配器，就像连接到RoBERTa的插件一样。不同适配器之间没有信息流，因此可以以分布式方式有效地训练多个适配器。他们在这项工作中注入了两种知识，包括从维基百科和维基数据上自动对齐的文本三元组中获得的事实知识，以及通过依赖解析获得的语言知识。从而解决了多种知识注入时更新预训练模型的原始参数，历史上注入的知识就会被冲走的问题。

Yu[21]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。

Wang[22]等人提出了一种KEPLE模型，以预训练语言模型作为嵌入对文本实体描述进行编码，然后共同优化知识嵌入和语言建模目标。解决了传统的知识嵌入模型无法充分利用丰富的文本信息的问题。

Sun[23]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。

## 1.3 研究内容

在围绕融合知识图谱的科学教育大模型推理方法研究与实现的研究目标中，如下图3-1所示，本文聚焦于构建科学类课程知识图谱，检索知识图谱中的相关知识和利用知识图谱知识辅助大模型推理三个关键方面。

首先，通过构建科学类课程知识图谱，我们旨在深入挖掘和呈现科学教育领域的关键知识点和概念之间的联系。这有助于建立一个全面而精准的知识体系，为大模型推理提供丰富的外部知识库信息。

其次，检索知识图谱中的相关知识用于为大模型提供参考，允许其基于特定查询检索信息。这种方法特别适用于回答特定查询或处理特定信息检索任务的场景。对于当前构建科学教育大模型的任务，该方法避免了重新对大模型进行训练，提高了其响应的准确性。

利用知识图谱知识辅助大模型推理是本课题的另一主要内容。我们需要设计合理的架构，将检索知识图谱获取的相关知识融入大模型的提示，让大模型能够参考相应的知识从而给出合理回答。因此，可以将该环节理解为“检索+生成”，前者主要是利用向量数据库的高效存储和检索能力，召回目标知识；后者则是利用大模型和提示工程，将召回的知识合理利用，生成目标答案。

最终，本文的研究目标还包括设计支持知识增强的科学教育问答系统。通过将构建的科学类课程知识图谱和知识图谱辅助大模型推理的方法结合，我们旨在开发一个智能问答系统，能够以更高效、准确的方式回答科学类课程的问题，促使用户深入理解科学知识。

总体而言，这三个研究目标相互关联，共同构建了一个综合而有力的研究框架，旨在推动科学教育领域智能教育技术的发展和创新。

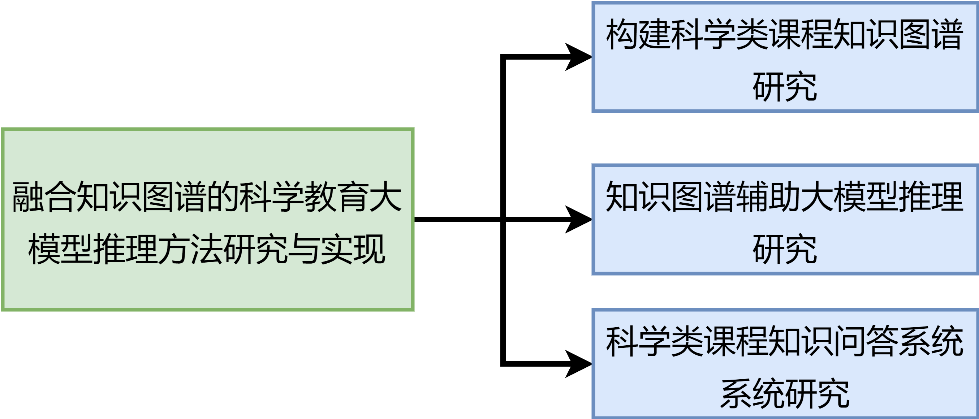


图1-1 融合知识图谱的科学教育大模型推理方法研究与实现研究内容图

### 1.3.1 科学类课程知识图谱的构建

目前，中小学科学教育领域尚未形成完善的知识图谱体系，并且在科学教育领域中，不同学科对于知识图谱构建的需求存在差异。为了弥补这一空白，本课题计划从设计并建立一个专注于科学类课程的知识图谱，旨在为未来相关研究提供一个坚实且全面的数据基础。这个专门针对科学教育的知识图谱将有助于整合和系统化科学知识，为学生和教育者提供更为清晰、有效的学科学习资源。

### 1.3.2 知识图谱辅助大模型推理方法

为了克服大模型可能出现的知识性幻觉问题，本课题提出在大模型的推理阶段引入外部知识库的策略。在这一方法中，大模型将利用事先构建的科学类课程知识图谱知识库，从中获取相关的知识文本。这些获取到的知识文本将被追加到模型的提示（prompt）中，以充当背景知识，供大模型在推理过程中参考和利用。通过将外部知识集成到模型的学习和推理过程中，有望有效提升模型的综合认知能力，减缓知识性幻觉现象的发生。这一策略为拓展大模型的认知范围、提高其知识获取的准确性提供了一种新颖而实用的途径。

### 1.3.3 科学类课程知识问答系统

本课题的核心关注点是开发一个具备用户界面的、专注于科学类课程知识的问答系统。该系统旨在借助大型模型和知识图谱技术，为用户提供高效且准确的科学类课程知识查询与问答服务。通过整合先进的自然语言处理模型和丰富的知识图谱，用户能够以直观友好的方式与系统进行互动，从而获得针对科学学科的详尽信息和精准答案。这一系统的设计旨在满足用户对科学知识的迅速获取和深入理解的需求，提升科学教育的学习体验。

# 第2章 相关理论与技术

## 2.1 聚类方法

### 2.1.1 DBSCAN算法

DBSCAN（Density-Based Spatial Clustering of Applications with Noise）是一种基于密度的空间聚类算法，旨在从包含噪声的数据中识别出簇（clusters）。这种算法在1996年由Martin Ester, Hans-Peter Kriegel, Jörg Sander和Xiaowei Xu提出[24]。DBSCAN的主要优势是它不需要事先指定簇的数量，而且能够有效地处理噪声和发现任意形状的簇。

考虑数据集合，首先引入以下概念与数学记号：

1. 邻域

设，则称为的邻域。

1. 密度

设，则称为的密度。

1. 核心点

设，为核心点阈值，若，则称为的核心点。记由中所有核心点构成的集合为，并记表示由中的所有非核心点构成的集合。

1. 边界点

若，且满足，则称为的边界点。记由中所有边界点构成的集合为。

1. 噪音点

记，若，则称为的噪音点。

DBSCAN算法的目标是将数据集合分成个cluster及噪音点组成，为此，引入cluster标记数组：

DBSCAN算法的流程如下：

|  |  |
| --- | --- |
| **算法1：DBSCAN聚类算法** | |
| 输入：样本集D，领域参数 | |
| 输出：簇划分 | |
| 1 | 初始化核心对象集合： |
| 2 | 对于每个三元组For |
| 3 | 确定样本的邻域 |
| 4 | If then |
| 5 | 将样本加入核心对象集合： |
| 6 | 初始化聚类簇数： |
| 7 | 初始化未访问样本集合： |
| 8 | While |
| 9 | 记录当前未访问样本集合： |
| 10 | 随机选取一个核心对象，初始化队列 |
| 11 |  |
| 12 | While Q |
| 13 | 取出队列Q中的首个样本q |
| 14 | If then |
| 15 | 令 |
| 16 | 将中的样本加入队列Q |
| 17 |  |
| 18 | ，生成聚类簇 |
| 19 |  |

## 2.2 检索方法

### 2.1.1 Faiss检索

Faiss的全称是Facebook AI Similarity Search，它是FaceBook的AI团队开发的面向稠密向量高效的相似性检索与聚类引擎，它能够快速处理大规模数据，并且支持在高维空间中进行相似性搜索。

给定一组维度为的向量集，Faiss在RAM中构建一个基于此的数据结构。构建完结构后，给定一个新的维度为的向量，它高效地执行如下操作：

其中是欧氏距离。

在Faiss中，这种数据结构就是一个索引，计算就是在索引上的搜索操作，寻找与最相似的向量数据。

Faiss不仅能返回最近邻的向量，还能返回第k个最近邻向量。同时，它能够实现批处理，一次搜索多个向量。Faiss在精确度和查询速度上做了权衡，通过模糊查询提升了其查询速度。

对于现有的向量数据或文本的嵌入数据，Faiss将根据用户选定的索引结构，将数据导入索引。这个过程也完成了对索引的训练。完成索引的建构之后，Faiss将用向量化的用户查询去检索上述索引结构。Faiss常用的索引结构包括暴力查询IndexFlatL2，查询优化IndexIVFFlat和存询优化IndexIVFPQ。

### 2.2.2 Lucene检索

Lucene是一套用于全文检索和搜索的开放源码程序库，由Apache软件基金会支持和提供。Lucene 由于其高性能和灵活性，被广泛应用于商业和开源项目中。它是许多流行搜索平台和服务的基础，如 Apache Solr 和 Elasticsearch，这些产品提供了更完整的搜索引擎功能，如分布式搜索、实时索引更新和易用的 API。

Lucene有着灵活的索引结构，允许将数据组织成多个文档和字段。每个文档可以包含不同类型的字段，如文本、日期等，适用于不同的数据存储和检索需求。通过算法优化和数据结构创新，Lucene 能够快速处理大规模数据的索引和查询，具有非常高的效率。

Lucene中包含了四种基本数据类型，分别是索引(Index)，文档(Document)，域(Field)和词(Term)。Lucene 的核心是索引的概念。在 Lucene 中，数据被组织成文档，每个文档是域的集合。域可以是文本字符串、URL、日期或任何需要进行搜索的数据类型。词是索引的最小单位，是经过词法分析和语言处理后的字符串。一个域通常包含一个或多个词。在索引过程中，Lucene 使用分词器(Analyzer)来处理每一段有意义的文本。

为了提高检索速度，Lucene采用了倒排索引(inverted index)的结构。通常的正排索引是从文档中检索是否包含关键单词。而倒排索引是指从单词检索文档，即需要为文档预建立关键字，然后在检索时通过匹配关键字来获取目标文档。

## 2.2 大模型检索增强生成方法

### 2.2.1 概览

大模型是指具有大规模参数和复杂计算结构的机器学习模型。这些模型通常由深度神经网络构建而成，拥有数十亿甚至数千亿个参数。大模型的设计目的是为了提高模型的表达能力和预测性能，能够处理更加复杂的任务和数据。大模型在各种领域都有广泛的应用，包括自然语言处理、计算机视觉、语音识别和推荐系统等。大模型通过训练海量数据来学习复杂的模式和特征，具有更强大的泛化能力，可以对未见过的数据做出准确的预测。

大模型的发展历程中一个重要的里程碑是Transformer模型的提出[25]。Transformer模型采用了自注意力机制（self-attention mechanism），极大地提高了模型处理长距离依赖关系的能力，成为了大型模型发展的基础。而后，GPT（Generative Pre-trained Transformer）系列模型[26]的问世更是推动了大型模型的发展。这些模型在大规模文本数据上进行预训练，然后在特定任务上进行微调，取得了令人瞩目的成绩，包括自然语言理解、生成、对话等多个领域。

大语言模型虽展现出强大能力，但在实际应用中，例如在准确性、知识更新速度和答案透明度方面，仍存在挑战。神经网络通常通过对模型进行微调来适应特定领域或专有信息，从而将知识参数化。尽管这种方法取得了显著成效，但它需要消耗大量计算资源，成本高昂，且需要专业技术知识，因此难以适应不断变化的信息环境。

检索增强生成 (Retrieval-Augmented Generation, RAG) [27]是指在利用大语言模型回答问题之前，先从外部知识库检索相关信息，将非参数化的语料库数据库与参数化模型相结合。RAG 被证明能显著提升答案的准确性，并特别是在知识密集型任务上减少模型的错误输出。通过引用信息来源，用户可以核实答案的准确性，从而增强对模型输出的信任。此外，RAG 有助于快速更新知识并引入特定领域的专业知识。RAG 有效结合了大语言模型的参数化知识和非参数化的外部知识库，成为实施大语言模型的关键方法之一。

### 2.2.2 发展过程

（1）原始RAG

原始RAG (Naive RAG)代表了早期研究方法，在ChatGPT广泛应用后迅速崭露头角。原始RAG的流程包括传统的索引、检索和生成步骤。原始RAG也被概括为一个“检索 - 阅读”框架[28]。

原始RAG包含三个重要组成部分，分别是索引，检索和生成。索引指的是在离线状态下，从数据来源处获取数据并建立索引的过程。而后根据用户的输入，原始RAG采用与第一阶段相同的编码模型将查询内容转换为向量。系统会计算问题向量与语料库中文档块向量之间的相似性，并根据相似度水平选出最相关的前 K 个文档块作为当前问题的补充背景信息。将给定的问题与相关文档合并为一个新的提示信息。随后，大语言模型被赋予根据提供的信息来回答问题的任务。根据不同任务的需求，可以选择让模型依赖自身的知识库或仅基于给定信息来回答问题。

原始RAG在检索质量、生成质量方面都面临着挑战。首先，检索集中的文档块并不都与查询内容相关，这可能导致信息错误或不连贯。其次是低召回率问题，即未能检索到所有相关的文档块，使得大语言模型无法获取足够的背景信息来合成答案。此外，过时信息也是一个挑战，因为数据冗余或过时可能导致检索结果不准确。而在生成方面，最突出的问题是制造错误信息，即模型在缺乏足够上下文的情况下虚构答案。另一个问题是回答不相关，即模型生成的答案未能针对查询问题。

（2）高级RAG

为了克服 Naive RAG 的局限性，高级 RAG 进行了针对性的改进。高级 RAG 在预检索处理和检索后处理方面提出了多种优化检索流程的方法。

在预检索处理方面，高级RAG主要在索引和嵌入方向进行了优化。在索引方面，高级RAG向原始文本中添加元数据信息如日期和用途等。通过元数据进行初步筛选可以提高检索的效率和准确性。同时，高级RAG将会采用混合检索的方式，其优势在于它结合了不同检索技术的长处。它智能地融合了关键词搜索、语义搜索和向量搜索等多种技术，适应不同类型的查询需求，确保能够一致地检索到最相关和内容丰富的信息。在嵌入方面，高级RAG将会采用微调嵌入的方法，让检索到的内容与查询之间的相关性更加紧密，优化检索内容对最终输出的影响。

在检索后处理方面，高级RAG希望减少检索得到的信息中的噪音，使得大语言模型聚焦关键信息。同时，也能保证检索内容不要超出大模型的限制token长度。因此也涌现了诸多技术来压缩检索结果的长度，提取检索内容的关键信息。

# 第3章 融合知识图谱的科学教育大模型推理方法

## 3.1 引言

大模型面临的知识性幻觉、知识更新以及数据相关的问题影响了大模型本身的可靠性以及其应用场景。在开放领域问题回答 [27]和常识推理[29]这些领域，大模型均面临着挑战。

检索增强生成(Retrieval-Augmented Generation, RAG)结合了一个预训练的检索器和一个预训练的生成器，通过端到端微调来以更可解释和模块化的方式捕获知识。在大型模型出现之前，RAG主要专注于直接优化端到端模型。而随着后续研究的发现，在大模型的上下文学习(In-Context Learning, ICL)中，检索增强生成可以有效解决上述问题。

此外，相比于大模型的其他优化技术，检索增强生成技术依靠着可更新的外部知识库，保证了知识的及时性。同时，用户可通过知识引用来源验证答案的准确性，提到了用户对大模型输出的信任。此外，检索增强生成技术在可为特定知识领域定制模型的同时，避免了参数的更新与模型的训练，使得其有着较强的可扩展性。

同时，相比于选用一些非结构化的文本数据作为外部知识库进行检索，选用知识图谱这种以三元组形式存储结构化知识的知识库，有效减少了预检索处理，不再需要进行例如剔除不相关内容，消除歧义和冗余等操作来提高数据的粒度。对于每一个知识图谱三元组，我们也能较为方便的创建标准索引和调整其块大小，提高检索的效率。

基于上述理由，本章选用检索增强技术来结合知识图谱的知识优化科学教育大模型的推理。相比于原始RAG(Naïve RAG)的“索引-检索-生成”的简单流程，本文设计在预检索处理，检索算法以及后检索处理以及三个方面进行改进，以提高其检索质量和生成质量。

本章设计的融合知识图谱的科学教育大模型推理方法流程如下图3-1所示。流程包括5个步骤：重写用户请求，检索知识图谱，内容提取，答案生成以及事实验证。在检索知识图谱环节中，本章选用DBSCAN聚类算法对知识图谱进行子图划分，以划分后的子图作为单位进行嵌入，使用嵌入后的向量进行相似度检索。

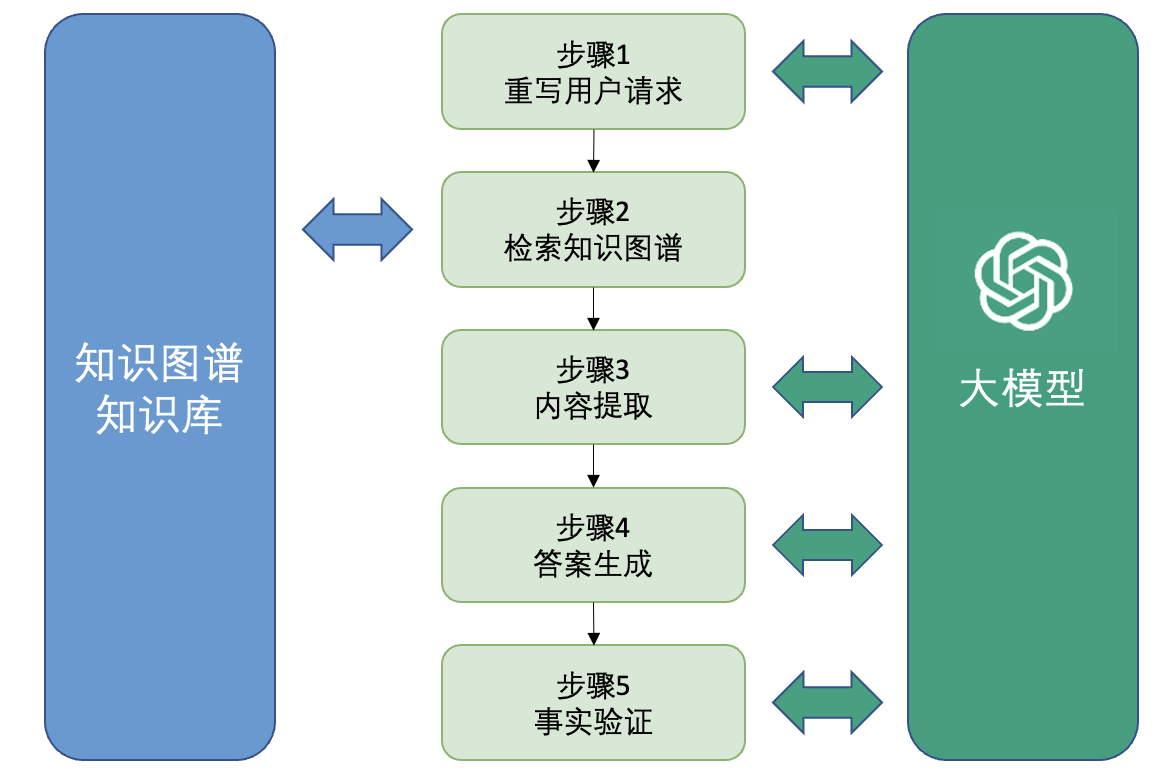


图3-1 融合知识图谱的科学教育大模型推理方法流程

## 3.2 重写用户请求

考虑到用户对本文构建的科学教育大模型的提问是连续的，因此用户的当前提问并不一定完整，而是包含着潜在的上文寓意。为了获取完整的用户提问来完成写下来对知识图谱的检索，我们构造了合理的提示用于与大模型交互，将先前保存的历史提问与当前提问共同输入进大模型中，得到重写后的用户请求。

我们构造的合理提示为：你现在作为资深心理学家，你的工作是理解用户的真实意图。在与你谈话的过程中，用户提出了一系列问题请求，这些问题请求是连贯的，但是这些问题请求本身可能并不完整，因为某个问题请求的主语或代词或其他部分可能是缺失的，出现在上一个问题中。因此，现在需要你通过理解上文用户提出的问题请求，把用户现在的问题请求补充完整。用户的已经提出的问题请求是:{history}，现在用户的问题请求是:{request}。请对用户的问题请求进行改写，使其含义明确。最后仅输出一个改写后的问题请求，不需要任何其他解释和中间过程。重写的输出结果如下表3-1所示。

表3-1重写输入输出结果

|  |  |
| --- | --- |
| **输入** | **输出** |
| history：介绍一下光的折射  request：那光的反射呢？ | 介绍一下光的反射 |

## 3.3 检索知识图谱

对于已经完成重写的用户请求，我们需要检索科学教育知识图谱中与请求最相关的信息作为参考背景。我们希望检索算法能够尽可能多的捕获用户提问中涉及的诸多实体对应的知识信息，同时将与用户提问中涉及的实体之间的关系最相关的知识信息排列在检索返回结果的最前列。为了实现上述目标，我们提出了基于DBSCAN算法的层次聚类知识图谱子图划分算法。

首先，我们进行对知识图谱子图的划分。如下图3-2所示，对于知识图谱中的一个三元组三元组，我们首先将其转换成一个向量用于表示。在本章中，我们选用了语义向量模型（Embedding Model）智源 BGE（BAAI General Embedding）模型作为嵌入模型。转换后的向量维度为1536。

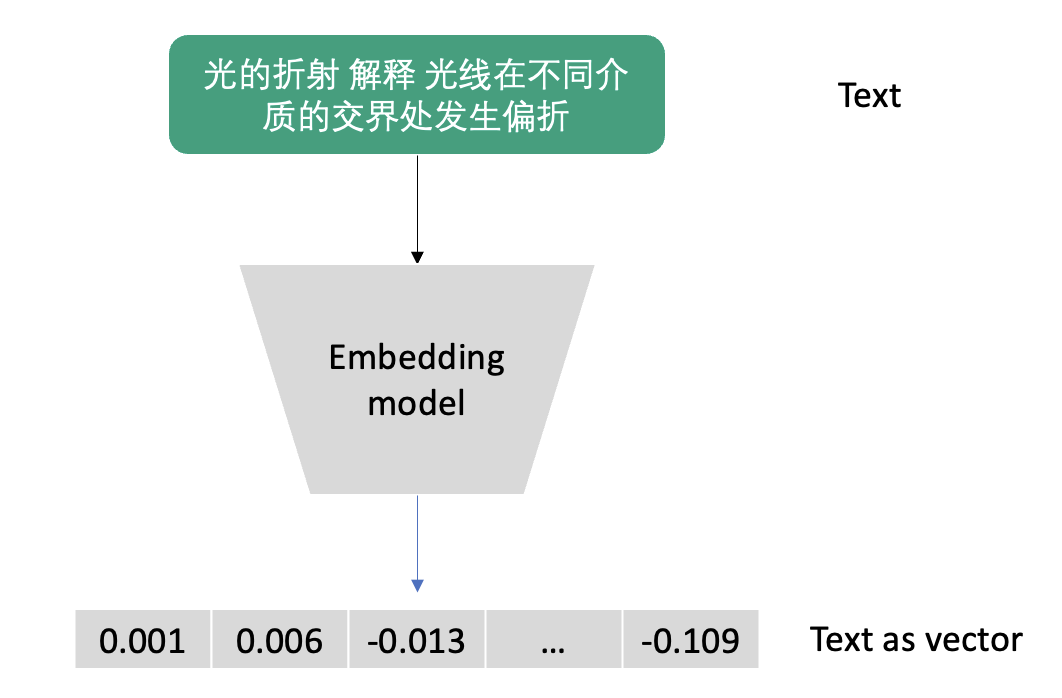


图3-2嵌入流程

因此，对于一个知识图谱三元组，我们可以将其转换为一个高维空间中的点，通过对高维数据空间的划分来实现点的聚类，从而实现知识图谱子图的划分。考虑到我们需要保证子图的规模大小以符合大模型的上下文长度限制，在本章中，我们选择了基于密度的聚类算法DBSCAN（Density-Based Spatial Clustering of Applications with Noise）。

在同DBSCAN聚类算法实现知识图谱子图划分的过程中，我们还采用了从上而下的层次聚类（Hierarchical Clustering）算法的思想。具体算法流程如下：

|  |  |
| --- | --- |
| **算法1：****基于DBSCAN算法的层次聚类知识图谱子图划分算法** | |
| 输入：知识图谱三元组 | |
| 输出：知识图谱子图划分 | |
| 1 | 对于每个三元组For |
| 2 | 通过BGE模型生成向量组 |
| 3 | 设DBSCAN算法的簇内最小样本数为， |
| 4 | 设DBSCAN算法的领域半径为，设初值 |
| 5 | 使用DBSCAN算法对进行聚类 |
| 6 | 对于每个簇划分For |
| 7 | If then |
| 8 |  |
| 9 |  |
| 10 | Else |
| 11 |  |
| 12 | 令 |
| 13 | If then |
| 14 | 再次进行聚类 Goto 5 |

在上述算法中，我们利用了DBSCAN算法的性质，通过不断减小邻域半径，实现了对聚类结果簇大小的控制。算法把簇看作数据空间中由低密度区域分割开的高密度对象区域，将足够高密度的区域划为簇，可以在有噪音的数据集中发现任意形状的聚类。然后对高密度簇进行进一步划分，完成了对整个知识图谱按密度进行层次划分，形成了合适的子图。最后共计划分成13647个子图。

完成了子图划分之后，我们将一个子图作为检索单位进行检索。我们进一步利用BGE模型对子图实现嵌入，并针对用户查询与子图使用Faiss算法进行检索。

Faiss检索是一种计算查询向量与所有存储向量的相似性得分的相似度检索。在知识图谱子图嵌入之前，我们先对子图进行预处理。考虑到嵌入模型的token长度限制以及后续构造成提示后可能超过的LLM的token长度限制，我们将所有以json格式保存的知识图谱子图截断为1000个tokens的固定长度，考虑到我们在子图划分阶段就已经控制了各个子图的规模，绝大多数知识图谱子图都并未超过目标长度，该步骤将不会损失很多语义信息。

完成子图嵌入之后，我们使用Faiss构建索引。考虑到科学教育知识图谱的规模，我们选用IndexFlatL2索引结构。这样在查询速度适当的同时能保证精确度和召回率。然后，将文本数据集中的每个文本对应的嵌入向量添加到Faiss索引中。这一步相当于在高维空间中建立了一个点的集合，以便于后续的快速搜索。完成索引的构建后，将索引进行保存。在这一步之后，索引就已经准备好接受查询并返回最相似的嵌入向量了。

对于之前预处理后得到的新的查询，将需要查询文本输入到BGE模型中，以获取对应的嵌入向量。然后将查询向量传递给Faiss索引进行搜索，Faiss会找到与查询向量最相似的知识图谱子图嵌入向量。最后检索到的嵌入向量可以被映射回原始知识图谱数据集中，从而得到与查询文本最相关的知识图谱子图。

最后，取搜索器返回的前3个知识图谱子图作为本次查询的参考文档。

## 3.4 背景内容提取

对于检索得到的原始文本数据，提取真正与查询相关的内容是有必要的。因为检索出的原始文本集合的长度结合构造的提示有可能超过大模型的输入token长度限制，同时检索出的结果可能仍包含与用户查询无关的内容，直接输入大模型有可能会造成干扰。

我们输入原始文本和用户查询构造提示如下：请你从用书名号《》分割的文档内容中，选出和引号“”分割的查询有关的文档内容片段。注意，必须是文档原文片段，不能够总结和生成新的内容。该文档为知识图谱中的一对实体关系三元组, 文档内容：《{content}》，查询：“{question}”。如果你认为没有相关内容，请只输出No。

背景内容提取的结果如下表3-2所示。

在进行提取时，我们构造了语句窗口提取器。由于大模型的输入token长度限制，无法直接将所有前3个相关知识图谱子图的内容连接起来作为它们生成答案的参考背景文本，且通过截断文档内容的简单方法可能会丢失其中的重要信息。我们重新利用大模型自身从检索到的文档中基于修改后的请求提取相关片段。设置滑动窗口的大小为256，降低信息内容提取的粒度，从而提高大模型对与用户查询相关内容提取的准确性，避免遗漏。

表3-2背景内容提取的输入输出结果

|  |  |
| --- | --- |
| **输入** | **输出** |
| content：['反丁烯二酸 中文名 反丁烯二酸', '反丁烯二酸 外文名 Fumaric acid', '反丁烯二酸 分子式 C4H4O4', '反丁烯二酸 分子量 116.07']  request：介绍一下反丁烯二酸 | ['反丁烯二酸 中文名 反丁烯二酸', '反丁烯二酸 外文名 Fumaric acid', '反丁烯二酸 分子式 C4H4O4', '反丁烯二酸 分子量 116.07'] |
| content：['反丁烯二酸 中文名 反丁烯二酸', '反丁烯二酸 外文名 Fumaric acid', '反丁烯二酸 分子式 C4H4O4', '反丁烯二酸 分子量 116.07']  request：介绍一下乙烯 | No |

此外，考虑到该环节将对大模型进行多次调用且构造的提示内容较长，我们设置了一个线程池，对于每一次提取都分配一个线程完成任务。

对于提取结果为No的原始文本数据，我们进行抛弃。然后整合剩下的所有文本数据作为用户查询的最终背景知识。

## 3.5 答案生成与事实验证

对于提取完成的背景，考虑到大模型的token长度限制，我们将其截断成长度为3000的文本块。将截断后的文本与预处理后的用户查询共同构造提示如下：现在，请完全的基于用书名号分割《》的参考信息，简洁和专业的来回答引号“”分割的用户的问题。不允许在答案中添加编造成分，答案请使用中文。参考信息:《{context}》 问题:“{question}”。 答案生成的结果如下表3-3所示。

表3-3背景内容提取的输入输出结果

|  |  |
| --- | --- |
| **输入** | **输出** |
| content：['五氧化二磷 描述 五氧化二磷（phosphorus pentoxide），化学式P2O5。']  ['五氧化二磷 熔点 340℃' '五氧化二磷 沸点 360°C (升华)' ]  request：五氧化二磷是什么？ | 五氧化二磷是一种无机化合物，化学式为P2O5。 |

最后，我们使用事实检查模块验证生成的答案是否含有事实错误，并为用户请求输出最终响应。尽管提供了额外的证据来生成，但是大模型可能也会产生幻觉。因此需要设计一个模块来进行进一步的事实验证。由于大模型具有强大的自然语言理解能力，我们将参考背景文本和生成的答案提供给它们进行判断。因此，大模型可以决定是否输出生成的答案，或者只是说“我无法回答这个问题”。

我们构造提示如下：现在，根据下面的参考，判断答案是否包括事实错误。参考用三括号[[[]]]分隔。答案由三括号((())分隔。请只输出是或否。参考：[[[{reference}]]]，答案：((({answer})))。事实验证的结果如下表3-4所示。

表3-4事实验证的输入输出结果

|  |  |
| --- | --- |
| **输入** | **输出** |
| reference：1.光敏电阻 描述 光敏电阻（photoresistor or light-dependent resistor，后者缩写为ldr）或光导管（photoconductor），常用的制作材料为硫化镉，另外还有硒、硫化铝、硫化铅和硫化铋等材料。  2.光敏电阻 中文名 光敏电阻  3.光敏电阻 外文名 photoresistor"  4.光敏电阻 又称 光敏电阻器或光导管  5.光敏电阻 制作材料 硫化镉  answer：光敏电阻（photoresistor）是一种光敏元件，也称为光敏电阻器或光导管，常用的制作材料包括硫化镉、硒、硫化铝、硫化铅和硫化铋等材料。 | 否 |
| reference：1.光敏电阻 描述 光敏电阻（photoresistor or light-dependent resistor，后者缩写为ldr）或光导管（photoconductor），常用的制作材料为硫化镉，另外还有硒、硫化铝、硫化铅和硫化铋等材料。  2.光敏电阻 中文名 光敏电阻  3.光敏电阻 外文名 photoresistor"  4.光敏电阻 又称 光敏电阻器或光导管  5.光敏电阻 制作材料 硫化镉  answer：光敏电阻（photoresistor）是一种将检测到的气体的成分和浓度转换为电信号的传感器，它是利用某些半导体吸收某种气体后发生氧化还原反应制成的，主要成分是金属氧化物。 | 是 |

## 3.6 系统实现

在科学教育知识问答系统方面，我们选用了Streamlit用于创建交互式网页应用。Streamlit 是一个开源的 Python 库，用于快速创建和共享数据应用。它被设计为简化数据科学和机器学习项目的展示过程，使开发者能够以最少的代码实现强大的交互式网页应用。

对于系统中每次必须被保存的用户历史提问，我们将其保存至Streamlit的会画状态中，以便于实时的添加和获取。

系统界面如下图3-3所示：



图3-3 科学教育知识问答系统界面

# 第4章 实验设计和结果分析

## 4.1 实验数据和评价方法

### 4.1.1 知识图谱构造

在本文中，我们需要需要构造科学教育知识图谱作为系统的外部知识库。为了选择合理的科学教育知识点，我们选用了清华大学构建的EDUKG，它是一个异构的、可持续的、跨学科的、细粒度的K-12教育知识图谱[30]。我们选择其中的一个可公开的收录了中学的物理、化学、生物、地理课程的知识点名称的知识图谱作为标准选取知识点。由于该知识图谱以RDF数据模型的形式存储，我们通过将其导入Neo4j图形数据库管理系统，通过编写Cypher 查询语句获取了所有知识点的标签名，共获取知识点名称5588个。

在知识点内容方面，我们选取开源中文知识图谱库ownthink。我们对于其中的所有知识图谱三元组的头实体进行分词，令分词结果为，EKUKG的知识点名词集合为，若满足：

则将这个知识图谱三元组加入我们的科学教育知识图谱中。

最终，我们构造了一个包含200000条知识图谱三元组的科学教育知识图谱。

### 4.1.2数据集

在本文中，为了更好的评价使用检索增强框架的基于知识图谱的科学教育大模型推理方法，我们从华东师范大学计算机科学与技术学院的EduNLP团队探索了针对教育垂直领域的对话大模型EduChat相关项目研发[31]中使用的测试数据集中抽取了50条与本文构建的科学教育知识图谱相关的问答对作为测试数据集。

### 4.1.3 评价方法

为了更好对本文中的使用检索增强框架的基于知识图谱的科学教育大模型推理方法进行评估，我们选用了RAGAs (Retrieval-Augmented Generation Assessment)框架[32]。在其中，我们选用了忠实度(Faithfulness)、答案正确性(Answer Correctness)这两个评价指标。

### 4.1.3.1 忠实度

忠实度衡量了生成的答案(answer)与给定上下文(context)的事实一致性。它是根据answer和检索到的context计算得出的。并将计算结果缩放到 (0,1) 范围且越高越好。

如果答案(answer)中提出的所有基本事实(claims)都可以从给定的上下文(context)中推断出来，则生成的答案被认为是忠实的。为了计算这一点，首先从生成的答案中识别一组claims。然后，将这些claims中的每一项与给定的context进行交叉检查，以确定是否可以从给定的context中推断出它。忠实度分数由以下公式得出：

忠实度指标可以很好的评价我们改进的检索增强框架是否能更好的帮助大模型从检索外部知识库得到的知识中捕获更多的信息，并且对其加以利用以生成包含更多信息的答案。

### 4.1.3.2 答案正确性

答案正确性涉及评估生成的答案与真实答案的准确性。这种评估依赖于真实答案和生成的答案，分数范围从0到1。更高的分数表示生成的答案与真实答案之间的一致性更高，表明答案的正确性更好。答案正确性包括两个关键方面：生成答案的事实正确性，生成答案与真实答案之间的语义相似性。这些方面通过加权方案组合，形成答案正确性得分。

生成答案的事实正确性量化了生成答案与真实答案之间的事实重叠，通过以下概念完成计算：

TP（真阳性）：在真实答案和生成答案中都存在的事实或陈述。

FP（假阳性）：存在于生成答案中但不在真实答案中的事实或陈述。

FN（假阴性）：存在于真实答案中但不在生成答案中的事实或陈述。

接下来，我们可以使用F1分数的公式来量化生成答案的事实正确性。

语义相似性评估基于真实答案和生成的答案，分值在0到1之间。分数越高，表示生成的答案与真实答案之间的一致性越好。测量答案之间的语义相似性可以提供关于生成回答质量的有价值见解。这种评估利用交叉编码器模型来计算语义相似性得分。在该指标中，分别使用指定的嵌入模型向量化真实答案和生成答案再计算两个向量之间的余弦相似度。设嵌入后的真实答案向量为，嵌入后的生成答案向量为。

在本章中，我们取事实正确性和语义相似性的加权参数分别为0.5，答案正确性分数由以下公式得出：

由于实验中的算法均基于我们构建的科学教育知识图谱开展检索与生成，答案正确性指标可以很好的评价我们改进的检索增强框架是否能更好的根据用户查询检索出尽量多的事实陈述，同时在整体上评估了整个流水线框架的性能。

## 4.2 实验和结果分析

### 4.2.1 实验设置

在本系统中，我们选用的大语言模型为OpenAI的gpt-3.5-turbo，我们通过在系统中使用OPENAI\_API\_KEY创建OpenAi的Client端对其进行调用。

在实验设置方面，我们提出了两组对比实验，用于分别评估本文对基于知识图谱检索模块的优化以及本文对检索增强生成框架的优化。

在第一组实验中，对于本文提出的基于DBSCAN算法的知识图谱子图划分作为嵌入检索单位的Faiss检索方法，我们选用常用于全文检索的Lucene检索方法和以单个三元组作为嵌入检索单位的Faiss检索方法进行对比。对于本文中提出的检索增强生成流水线框架，我们仅对其中的检索模块进行替换。

在第二组实验中，对于本文中提出的检索增强生成流水线框架，我们选用原始检索增强生成框架（Naïve RAG）进行对比，并保持其中的检索模块不变，均适用本文提出的知识图谱子图划分作为嵌入检索单位的Faiss检索方法。

### 4.2.2 实验结果分析

### 4.2.2.1 实验一 检索模块对比

表4-1 实验一对照结果

|  |  |  |
| --- | --- | --- |
| 模型 | 忠实度 | 答案正确性 |
| Advanced RAG + Lucene on KG Triples | 0.728 | 0.464 |
| Advanced RAG + Faiss on KG Triples | 0.754 | 0.449 |
| Advanced RAG + Faiss on KG Subgraphs | 0.850 | 0.472 |

**（1）实验对比阐述**

对于模型Advanced RAG + Lucene on KG Triples，我们将科学教育知识图谱中的一个三元组，将其以字典的格式保存在json文件中并预处理为的形式。Lucene检索将基于知识图谱三元组进行检索。整体的检索生成框架仍与本文中的框架一致。

对于模型Advanced RAG + Faiss on KG Triples，我们对科学教育知识图谱中的一个三元组进行嵌入，我们使用BGE模型作为嵌入模型，将其转化成一个长度为1536的向量。Faiss检索将基于知识图谱三元组的向量和查询向量进行相似度检索。整体的检索生成框架仍与本文中的框架一致。

**（2）实验分析**

从表4-1中可以看出本文中提出的Advanced RAG + Faiss on KG Subgraphs框架在忠实度和答案正确下两个方面都要优于模型Advanced RAG + Lucene on KG Triples和Advanced RAG + Faiss on KG Triples。

在忠实度方面，由于该模型是以算法划分的知识图谱子图为单位进行检索的，而在知识图谱子图划分方面，我们将语义相近的知识图谱三元组划分进同一子图。因此，对于一个子图来说，可从其中提取的事实陈述之间的语义信息必然是更为集中的。因此，对于科学教育领域出现次数较多的介绍性提问和涉及多个实体的复杂提问，以知识图谱子图这种包含了多个实体但语义信息又较为一致的这种单位进行查询，呈现出的综合忠实度和答案正确性结果较好。

而对于仅仅基于单个知识图谱三元组的检索来说，Lucene on KG Triples检索采用的是关键词匹配的方式，而忽视了用户查询的语义信息，导致检索结果的语义信息可能相差较大，同时真正符合用户查询语义的信息也更少。面对上下文背景信息的缺失，大模型也更难从其中捕获符合用户查询的信息，因此最后生成的回答正确性最低，而包含的相关事实陈述最少，综合的回答忠实度也最低。

Faiss on KG Triples检索由于是基于语义相似度的检索，整体的回答忠实度要略高于Lucene on KG Triples检索。但由于知识图谱三元组的“头实体-关系-尾实体”结构化形式，每一组三元组包含的实体之间的关系是较为简单的。而对于涉及三者即以上的实体或关系的复杂语义的用户查询，仅仅靠以三元组为单位的检索仍在捕获合适的三元组上存在困难。面对这样的复杂查询，很可能检索提供给大模型的上下文背景仍是不足的，从而导致答案正确性较低，而大模型最后生成的事实陈述数量也相应不足，这就导致综合的忠实度仍较低。

### 4.2.2.1 实验二 RAG框架对比

表4-2 实验二对照结果

|  |  |  |
| --- | --- | --- |
| 模型 | 忠实度 | 答案正确性 |
| Naïve RAG + Faiss on KG Subgraphs | 0.766 | 0.457 |
| Advanced RAG + Faiss on KG Subgraphs | 0.850 | 0.472 |

**（1）实验对比阐述**

对于模型Naïve RAG + Faiss on KG Subgraphs，我们删除了本文检索增强生成框架中的所有优化操作，仅仅保留了检索模块和生成模块。但对于检索模块仍使用了本文中的基于知识图谱子图进行检索的方法。

**（2）实验分析**

从表4-2中可以看出本文中提出的Advanced RAG + Faiss on KG Subgraphs框架在忠实度和答案正确下两个方面都要优于模型Naïve RAG + Faiss on KG Subgraphs。

原始检索增强框架（Naïve RAG）仅仅包含检索和生成两个步骤。很显然，在缺少了检索前处理和检索后处理等优化步骤之后，直接使用原始查询进行检索并且使用全部的检索内容作为上下文背景提供给大模型生成答案是效果较差的。原因主要有两个方面：第一是用户的查询很可能依赖于过往的上文，且包含隐含语义，直接将其作为检索依据将会影响检索的效果。第二是检索结果内容不一定都与用户查询相关，而这些内容之间通常又有着较高的语义相似度，直接将其连通用户查询提供给大模型可能造成大模型的混乱，从而无法完全捕获上下文背景中的事实陈述，或直接生成错误的答案。

整体上说，在缺少了检索增强框架优化的情况下，Naïve RAG + Faiss on KG Subgraphs仍呈现了相较于模型Advanced RAG + Lucene on KG Triples和Advanced RAG + Faiss on KG Triples更好的忠实性和答案正确性，也可见本文提出的基于DBSCAN算法的知识图谱子图划分作为嵌入检索单位的Faiss检索方法的有效性。

### 4.2.3 案例学习

### 4.2.3.1 实验一

在本案例中，我们选取数据集中的一组问答对，针对Lucene on KG Triples、Faiss on KG Triples和Faiss on KG Subgraphs三种方法获取的上下文背景和生成的答案，以及其对应的忠实度和答案正确性进行分析，问答对如下：

提问：蓝细菌和真菌有什么区别？

标准答案：蓝细菌是原核生物，属于细菌领域，而真菌是真核生物，属于真核生物界真菌门。在细胞结构、生活方式和生物学特征等方面有很大的差异。

如表4-3中所示，Advanced RAG + Lucene on KG Triples方法只捕获了提问中有关蓝细菌的信息，而考虑到Lucene检索使用了关键词匹配的方法，因此包含重复关键词“蓝细菌”的三元组则更容易被检索到。结果导致检索完全没有得到匹配提问的有效信息，因此生成答案与两个指标均不具备参考价值。

表4-3 Advanced RAG + Lucene on KG Triples结果

|  |  |
| --- | --- |
| 上下文背景 | 1. 蓝细菌 中文学名 蓝细菌 2. 蓝细菌 别称 蓝藻，蓝绿藻 3. 蓝细菌 标签 细菌 |
| 生成答案 | 蓝细菌是原核生物，真菌是真核生物。 |
| 忠实度 | 0.5 |
| 答案正确度 | 0.490 |

如表4-4中所示，Advanced RAG + Faiss on KG Triples方法并没有检索知识图谱中蓝细菌是细菌这个三元组，同时也没有检索出真菌有关的概念。但该方法成功检索处理出了蓝细菌能进行光合作用这一信息，因此答案也包含了相关内容。

表4-4 Advanced RAG + Faiss on KG Triples结果

|  |  |
| --- | --- |
| 上下文背景 | 1. 蓝细菌 别称 蓝藻，蓝绿藻 2. 蓝细菌 中文学名 蓝细菌 3. 蓝细菌 分类 产氧光合细菌' |
| 生成答案 | 蓝细菌和真菌的主要区别在于它们的分类。蓝细菌属于产氧光合细菌，能进行光合作用；而真菌不进行光合作用，属于真菌界。此外，蓝细菌在形态和生理功能上也与真菌有所不同。 |
| 忠实度 | 1.0 |
| 答案正确度 | 0.423 |

如表4-5中所示，Advanced RAG + Faiss on KG Subgraphs方法成功检索处理到了蓝细菌的分类、标签以及能进行光合作用这些信息。同时，考虑到实验的知识图谱并没有包含充足的有关真菌的信息，本次检索也实现了尽可能多的对真菌信息的提取。同时在第三个子图中，作为蓝细菌所属分类的细菌与真菌这两者信息都被包含进去了，因此也使得答案生成中包含了两者生物类型不同的这个重要结论。

表4-5 Advanced RAG + Faiss on KG Subgraphs结果

|  |  |
| --- | --- |
| 上下文背景 | 1. 蓝细菌 分类 产氧光合细菌   光合细菌 中文学名 光合细菌  光合细菌 分类 产氧光合细菌和不产氧光合细菌  光合细菌 英文名 Photosynthetic Bacteria   1. 蓝细菌 标签 细菌   蓝细菌 标签 微生物   1. 细菌 中文学名 细菌   细菌 拉丁学名 bacteria  细菌 英文名 germs  真菌 中文学名 真菌  真菌 拉丁学名 Fungus  细菌学 中文名 细菌学  细菌学 外文名 bacteriology |
| 生成答案 | 蓝细菌和真菌的区别在于生物类型不同，蓝细菌属于细菌 。蓝细菌是一种光合细菌，可以进行产氧光合作用，属于微生物的一种。相比之下，真菌不进行光合作用，属于不同的生物分类。 |
| 忠实度 | 1.0 |
| 答案正确度 | 0.536 |

### 4.2.3.2 实验二

在本案例中，我们选取数据集中的一组问答对，针对Naïve RAG + Faiss on KG Subgraphs和Advanced RAG + Faiss on KG Subgraphs两种方法生成的答案，以及其对应的忠实度和答案正确性进行分析，问答对如下：

提问：猪鼻龟是一种生活在水中的鱼类吗？

标准答案：不是，猪鼻龟是一种陆生龟类，属于爬行动物，常栖息于淡水环境中，而不是鱼类。

如表4-6中所示，Naïve RAG + Faiss on KG Subgraphs方法在完成检索后没有针对用户提问对检索结果进行提取，而原始检索结果中包含了许多与用户提问无关的信息，因此对大模型造成了背景信息利用的困难，最后没能充分利用其中的信息，答案正确性偏低。

表4-6 Naïve RAG + Faiss on KG Subgraphs结果

|  |  |
| --- | --- |
| 上下文背景 | 1. 猪鼻龟 中文学名 猪鼻龟   猪鼻龟 拉丁学名 Carettochelys insculpta  猪鼻龟 别称 大洋洲猪鼻龟，飞河龟，两爪鳖等  猪鼻龟 种 猪鼻龟  猪鼻龟 环境 淡水   1. 水金龟 标签 自然   水金龟 标签 生物物种  猪鼻龟 标签 科学百科生命科学分类  猪鼻龟 标签 自然  猪鼻龟 标签 生物物种  猪鼻龟 标签 动物  中国龟 标签 生物物种  中国龟 标签 古生物  中国龟 标签 动物   1. 猪鼻龟 科 两爪鳖科   猪鼻龟 属 两爪鳖属 |
| 生成答案 | 不是，猪鼻龟是一种生活在水中的爬行动物。 |
| 忠实度 | 1.0 |
| 答案正确度 | 0.239 |

如表4-7中所示，Advanced RAG + Faiss on KG Subgraphs方法成功过滤了与“猪鼻虫”无关的信息，捕捉到了“淡水”，“龟类”等重要信息，因此答案相对更完整，答案正确性也更高。

表4-7 Advanced RAG + Faiss on KG Subgraphs结果

|  |  |
| --- | --- |
| 上下文背景 | 1. 猪鼻龟 中文学名 猪鼻龟   猪鼻龟 拉丁学名 Carettochelys insculpta  猪鼻龟 别称 大洋洲猪鼻龟，飞河龟，两爪鳖等  猪鼻龟 种 猪鼻龟  猪鼻龟 环境 淡水   1. 猪鼻龟 标签 科学百科生命科学分类   猪鼻龟 标签 自然  猪鼻龟 标签 生物物种  猪鼻龟 标签 动物   1. 猪鼻龟 科 两爪鳖科   猪鼻龟 属 两爪鳖属 |
| 生成答案 | 不是，猪鼻龟是一种龟类，属于爬行动物，常栖息于淡水环境中，而不是鱼类。 |
| 忠实度 | 1.0 |
| 答案正确度 | 0.735 |

## 4.3 实验结论

在本章设置的实验中，我们使用RAGAs框架中的忠实度和答案正确性这两个指标，对于检索模块和检索增强生成框架结构这两个方面进行评价。

在检索模块方面，我们设置了方法Lucene on KG Triples和Faiss on KG Triples与本论文中提出的基于知识图谱子图进行Faiss检索的Faiss on KG Subgraphs方法进行对比。一方面，这种基于语义相似度检索的方法在诸多场景中均要优于以Lucene检索为代表的单纯进行关键词匹配的检索方法。另一方面，事先进行子图划分这一操作使得Faiss on KG Subgraphs方法在对于介绍性提问和多实体的复杂提问上的表现均要优于仅以单个三元组进行的Faiss检索。

在检索增强生成框架结构方面，我们设置了方法Naïve RAG + Faiss on KG Subgraphs与本论文中提出的Advanced RAG + Faiss on KG Subgraphs进行对比。实验结果表明Advanced RAG框架在针对用户提问语义进行检索和对检索结果进行后处理以便于大模型生成这两个方面的改进，均能够优化大模型最后答案的生成。

# 第五章 总结与展望

## 5.1 总结

论文中，为了缓解大模型的知识性幻觉问题，同时保证科学知识的实时性和准确的从而更好的构建科学教育大模型，我们设计了融合知识图谱的科学教育大模型推理方法，利用预先构建的科学教育知识图谱来实现检索增强生成技术，提出了基于DBSCAN算法的知识图谱子图划分作为嵌入检索单位的Faiss检索方法根据用户提问检索合适的知识。

同时，我们在预检索处理，后检索处理方面对原始检索增强生成技术框架进行改进，使用结构化的知识图谱数据作为外部数据源来优化索引结构，设计了合理的提示来对用户提问进行重写。此外，我们在检索完成后使用滑动窗口提取信息，对检索结果进行压缩和去除噪音，并且在大模型完成完成答案生成之后添加了事实验证模块，进一步降低幻觉现象。

最后我们在检索模块和检索增强生成框架两个方面设置了实验对照组，使用RAGAs框架针对忠实性和答案正确性两个方向进行评估，从而验证了论文中方法的有效性。

## 5.2 展望

在将来的工作中，会进一步优化高级检索增强生成框架，对于现在的流水线结构，考虑将其修改为模块化的框架，能够递归的调用大模型自身的能力，例如加入让大模型反思，评估，输出的功能。

同时，在针对知识图谱检索方面，考虑进一步对其进行探索。在知识图谱子图划分和嵌入的过程中，将知识图谱的结构信息和语义信息相结合加以利用。同时，在相似性检索的嵌入模块中加入微调嵌入的方法，使得检索到的内容与查询之间的相关性更加紧密。

# 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. Manning CD. Human language understanding & reasoning[J]. Daedalus 151(2):127–138
15. Yu F, Zhang H, Tiwari P, et al. Natural language reasoning, a survey[J]. arXiv preprint arXiv:230314725
16. Lu Y, Lu H, Fu G, et al. KELM: Knowledge Enhanced Pre-Trained Language Representations with Message Passing on Hierarchical Relational Graphs[C]//ICLR Workshop on Deep Learning on Graphs for Natural Language Processing. 2022.
17. Sun T, Shao Y, Qiu X, et al. CoLAKE: Contextualized Language and Knowledge Embedding[C]//Proceedings of the 28th International Conference on Computational Linguistics. 2020.
18. 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.
19. 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.
20. Wang R, Tang D, Duan N, et al. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP. 2021.
21. 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.
22. Wang X, Gao T, Zhu Z, et al. KEPLER: A unified model for knowledge embedding and pre-trained language representation[J]. Transactions of the Association for Computational Linguistics, 2021.
23. 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.
24. Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//kdd. 1996, 96(34): 226-231.
25. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
26. Radford A, Narasimhan K, Salimans T, et al. Improving language understanding with unsupervised learning[J]. 2018.
27. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rockt ¨ aschel ¨ et al., Retrievalaugmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.
28. Ma Y, Cao Y, Hong Y C, et al. Large language model is not a good few-shot information extractor, but a good reranker for hard samples![J]. arXiv preprint arXiv:2303.08559, 2023.
29. S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark et al., Improving language models by retrieving from trillions of tokens, in International conference on machine learning. PMLR, 2022, pp.2206–2240. Zhao B, Sun J, Xu B, et al.
30. EDUKG: a heterogeneous sustainable k-12 educational knowledge graph[J]. arXiv preprint arXiv:2210.12228, 2022. Dan Y, Lei Z, Gu Y, et al.
31. Educhat: A large-scale language model-based chatbot system for intelligent education[J]. arXiv preprint arXiv:2308.02773, 2023.
32. Es S, James J, Espinosa-Anke L, et al. Ragas: Automated evaluation of retrieval augmented generation[J]. arXiv preprint arXiv:2309.15217, 2023.

# 致 谢

四年的时间飞逝，……

# 本科期间主要研究成果

**发表论文：**

1. **……**

**完成项目：**

1. **高晨**，\*\*\*，\*\*\*. 大学生创新训练项目. 南京师范大学，2021~2022，0.3万.