

基于LLM Agent的教务查询系统 设计与实现

刘志忠

(南京传媒学院 传媒技术学院, 江苏 南京 210007)

摘要: 集成检索增强生成技术的大语言模型能较好地解决大语言模型的知识更新与“幻觉”问题,适用于教务查询领域。考虑到教务查询领域的通用问题、特定领域问题,根据大语言模型的不同需求和用户反馈持续优化系统。首先,利用基于LLM的Agent将文本分类模型与RAG组成处理流程;其次,根据问题分类结果融合Agent的记忆内容;再次,调用大语言模型生成系统应答;最后,在框架中引入基于用户反馈的强化学习机制,以优化Agent的记忆内容、查询结果及结果融合。实验表明,在Agent框架中引入文本分类模型可提升用户交互效率,降低大模型的调用开销;强化学习机制虽增加了系统复杂度,但用户满意率会随着用户交互不断提升。

关键词: Agent; 文本分类; 强化学习; 大语言模型; 检索增强生成; 教务查询系统

DOI: 10.11907/rjtk.251195

中图分类号: TP319

文献标识码: A

文章编号: 1672-7800(2025)010-0149-06

扫描二维码阅读全文:



Design and Implementation of Academic Affairs Inquiry System Based on LLM Agent

LIU Zhizhong

(School of Communication Technology, Communication University of Nanjing, Nanjing 210007, China)

Abstract: The integrated retrieval enhanced generation technology for large language models can effectively solve the problems of knowledge updating and "illusion" in large language models, and is suitable for the field of academic inquiry. Considering the general and specific issues in the field of academic inquiry, the system is continuously optimized based on the different needs of the big language model and user feedback. Firstly, use LLM based agents to combine the text classification model with RAG to form a processing flow; Secondly, based on the classification results of the problem, integrate the memory content of the agent; Again, call the big language model to generate system responses; Finally, a reinforcement learning mechanism based on user feedback is introduced into the framework to optimize the memory content, query results, and result fusion of the agent. Experiments have shown that introducing text classification models into the Agent framework can improve user interaction efficiency and reduce the overhead of calling large models; Although reinforcement learning mechanisms increase system complexity, user satisfaction rates will continue to improve with user interaction.

Key Words: Agent; text classification; reinforcement learning; large language model; retrieval augmented generation; academic affairs inquiry system

0 引言

大语言模型在自然语言处理领域的应用取得了突破性进展,正逐步渗透到教育、医疗、法律等行业^[1]。2023年11月,Nature杂志发表论文称大语言模型正在重塑教育行业^[2]。检索增强生成(Retrieval-Augmented Generation,

RAG)为大语言模型(Large Language Model, LLM)增加了额外的知识库,解决了困扰大语言模型的“幻觉”和知识更新问题。

目前,RAG与LLM的结合可有效利用通用知识与实时知识,但在教务查询系统中不同问题对大语言模型的要求并不一样。其中,通用知识要求大语言模型提供最终结果;实时知识则通常借助大语言模型对查询进行重写,以

收稿日期:2025-03-19

作者简介:刘志忠(1980-),男,博士,CCF会员,南京传媒学院传媒技术学院讲师,研究方向为Agent、大语言模型、检索增强生成、智能图像处理。本文通讯作者:刘志忠。

检索得到更多的外部知识。为此,本文利用该特性,对 RAG 与 LLM 的集成架构进行改进,以在一定程度上提升交互效率和用户体验。

1 相关研究

1.1 大语言模型

近年来,大语言模型研究取得了飞速发展。2023 年,Google DeepMind 推出的 Gemini 大模型,在传统大模型的基础上引入了对视频、语音、图片等的理解,实现了从 LLM 单模态到多模态的跨越^[3]。2024 年,OpenAI 继 ChatGPT、DaVinci 之后发布的 Sora 模型,在文本生成视频领域带来了重大突破,使多模态成为了大语言模型的研究热点^[4]。Claude 是美国人工智能初创公司 Anthropic 发布的大型语言模型家族,拥有高级推理、视觉分析、代码生成、多语言处理、多模态等能力^[5]。Claude 基于强大的保障措施和原则进行设计,提出“AI 宪法”的概念,减少偏见、尊重隐私,在提升安全性和透明度的同时减少了“过度拒绝”。

科大讯飞推出的新一代认知智能大模型星火大模型,拥有跨领域的知识和语言理解能力,能基于自然对话方式理解与执行任务,在专业知识方面表现较好,支持专业助手,有数百个专业助手涵盖各个领域,能为用户提供较为专业的帮助。文心一言是百度公司研发的知识增强大语言模型,由于百度拥有丰富的知识图谱和大量的行业知识,这些知识被整合到文心一言的训练过程中,使得模型能更好地理解、处理相关问题,提供更准确、更有深度的回答^[6]。同时,文心一言可与百度的搜索引擎技术相结合,利用搜索引擎的检索能力获取更多信息,从而提升回答的准确性和全面性。DeepSeek 是深度求索公司推出的大语言模型,通过优化混合专家模型(Mixture-of-Experts, MOE)架构,以超高的性价比获得了广大用户的青睐^[7]。

1.2 检索增强生成

RAG 是一种结合检索与生成技术的技术,引用外部知识库的信息生成答案或内容,具有较强的可解释性和定制能力,适用于问答系统、文档生成、智能助手等多个自然语言处理任务^[8]。RAG 模型的优势在于通用性强、可实现即时的知识更新,并通过端到端的评估方法提供更高效、精准的信息检索。

从技术上来看,RAG 与 LLM 之间具有很强的互补性,将 RAG 与 LLM 结合能更好地理解用户问题,从大量的文本数据中检索出相关信息,为用户提供更准确、更全面的答案^[9,10]。早期,RAG 技术仅在 LLM 的基础上扩展了外部知识库,对信息实施向量化并建立索引。具体为,用户在使用 LLM 前,先对该外部知识库的向量进行检索以获取更多信息,再将结果作为上下文提交给 LLM 进行处理,以实现更精准的查询。目前,知识库检索方式包括:token 检索、对数据块检索和对资源实体检索等^[11-13]。此后,研究

人员开始将 RAG 与 LLM 进行更紧密的结合,分别针对查询的前处理和后处理进行优化,通过查询扩展、查询重写、查询增强、查询后处理等方式优化 RAG+LLM 的输出^[14-17]。

1.3 基于大语言模型的 Agent

为了解决 RAG+LLM 无法处理复杂任务的困难,研究人员在 LLM 的基础上构建了大模型 Agent^[18,19]。其中,Agent 具备环境感知、自主理解、决策制定及执行行动的能力,通过感知、思考与行动的紧密结合完成复杂任务。HuggingGPT 是一个将 ChatGPT 作为任务规划器的框架,其规划器根据模型描述在框架中选择可用的模型完成任务,并依据执行结果对响应实施综合处理^[20]。此时,LLM 将作为一个中间协调者,协调多个模型的输出。Zhang 等^[21]提出基于 Agent 链(Chain-of-Agents, CoA)的多 Agent 系统框架,在该框架上通过自然语言实现超长上下文任务中的信息聚合和跨多个 LLM 的上下文推理。

在工业界,为了实现基于 LLM 的 Agent 开发。LangChain (<https://www.langchain.com/>)将 Agent 作为与现实世界沟通的主要模块,提供了各种类型的 Agent 及用于开发 Agent 的工具,通过 LLM 确定具体的操作与顺序。国内,Diffy (<http://diffyai.com/cn#llms-app-stack>)结合 LLM 的具体应用提供了 Agent 开发环境,基于特定流程调用相关的外部工具,以完成复杂的领域任务。

1.4 基于 LLM 的教务查询系统

EduChat 是由华东师范大学、复旦大学共同研发的,一个专门面向教育领域的聊天机器人系统,主要服务于教师、学生和家長群体,目标是支持个性化的智能教育^[22]。该系统提供了基于 Baichuan、Qwen 等大语言模型微调而得到的预训练模型,先在教育语料库上进行预训练以适应特定领域,并在设计的系统上微调提示词、指令来激发各种技能,在开放问答、作文评估、苏格拉底式教学和情感支持等教育特色功能方面表现出色。此外,该系统通过微调大模型来满足教务相关问题,所有问题均采用一致的方式使用 LLM 模型进行支持。

华中科技大学公开了一种基于 LLM 的知识图谱生成式辅助教学答疑方法及系统,将问题过滤后输入查询知识图谱,通过 LLM 生成回答以响应用户请求^[23]。该方法将外部知识存储为知识图谱,结合知识图谱查询与 LLM 来处理用户信息查询。分析表明,以 Agent 为基础,结合 LLM、RAG 技术解决信息查询问题,是 LLM 在教育领域应用的主要方向之一,但目前应用主要关注于扩展实时知识,较少关注查询效率。

现有应用中,所有查询交互都依赖于 LLM,系统的应用开销较高、查询效率较低。应用分析表明,不同种类的交互问题对 LLM 的使用要求不同,系统应在区分不同请求的基础上合理使用 LLM,降低应用成本、提升使用效率。本文以 LLM Agent 为基础,设计了一种新的增强大语言模

型架构,集成了一个简单的文本分类器来识别用户交互类别,针对不同交互类型按照不同的方式使用 LLM,以降低系统应用开销。此外,增加了相关的增强学习模块,利用用户对交互结果的反馈来改进模型交互效率和融合结果。

2 基于 LLM Agent 的教务查询系统设计

2.1 需求背景

教务查询系统在学校正常教学管理工作中承担着重要的角色,既包括各学科的教学内容,又涵盖了日常教学工作安排的内容。本文研究主要涉及教务内容查询,学生通过语音或文本的方式,从系统中获取所需内容。例如:本学期的数据结构课程是哪位教师授课?操作系统中的短作业优先调度算法如何理解?数据结构课程的期末考试成绩如何计算等?系统将根据知识库或大语言模型给出相关的答案。

从数据内容而言,系统包含的部分内容需要定期更新与维护;从使用角度而言,学生使用教学辅助系统存在以下特点:①在学生需求中,通用性查询与实时信息获取并重,既包含通用知识的获取(课程相关内容、教务系统使用的相关内容等),又包含实时信息的查询(学期课程安排,包括教室、任课老师、课程标准等);②在用户交互具有相关的场景性,交互问题需放在一定的上下文中进行理解,不同场景用户需按照不同的模式响应用户请求;③用户对交互响应的实时性较高,通常希望在提出请求后尽快给出相关答案。

2.2 解决方案

项目基本解决方案以检索增强的大语言模型为基础,利用 Agent 框架集成文本分类模型对交互文本进行分类,随后根据不同的分类结果执行不同的逻辑,以提升用户交互效率。

2.2.1 集成文本分类器的 LLM Agent

该方案以带记忆模块的 Agent 模型为基础,利用 Agent 的 pipeline 将记忆模块、检索模块、融合模块及 LLM 进行集成,基本结构如图 1 所示。其中,记忆模块(Memory Module)负责管理 Agent 的交互记录,保存在一定时间窗口内接触过的问题及其时间戳,所有问题按照访问时间先后排序,越新的问题排序越靠前,当问题一段时间未被访问时将从记忆中移除;检索模块以外部向量数据库存储的数据为基础,通过分类检索获取外部知识,并利用 Reranker 对查询结果进行重新排序。

因文本自身存在序列性特征,文本分类器通常基于双向长短期记忆网络(Bidirectional Long Short Term Memory Network, BiLSTM)所构建,数据主要包括:①通过前期线下调研收集的数据;②在引入分类系统前收集的线上数据;③通过 LLM 将线上、线下收集到的数据进行增强后的结果。数据经 LLM 分类后可得到初步标签,再进行人工审

核、调整后可得到训练数据集,数据集包含通用问题和实时问题。实验表明,在该训练集上实施训练的准确率在 94% 左右。

系统在接收到用户相关的查询后,调用文本分类器确定交互类型是否需要借助实时知识库辅助查询。如果需要辅助知识库支持,将产生检索增强的提示词,调用 LLM 生成扩展查询语句实时查询信息,基于语义相似度检索额外的向量知识库,并在实施重排序后将相关信息输出给融合模块。融合模块采用倒数融合排序的方式(Reciprocal Rank Fusion, RRF),将记忆模块中存储的排序信息与检索模块重排序后融合,输出 LLM 所需要的上下文信息。具体计算公式如式(1)所示。

$$Rank_{score}(d) = \sum_{i=1,2} \frac{1}{k_i + r_i(d)} \quad (1)$$

式中: k_i 为第 i 种排序方法对应的常数; $r_i(d)$ 表示第 d 个选项在第 i 种排序中的序列号; $Rank_{score}(d)$ 表示 d 的综合得分,得分越低表示重要性越差。

LLM 基于相关上下文产生多个输出供用户选择,在为 LLM 设计提示词时采用思维链方式,在有多个不一致的结果时,要求 LLM 必须输出不同结果的根源即推理规则,产生相关结果所对应的记忆或知识库条目与内容。

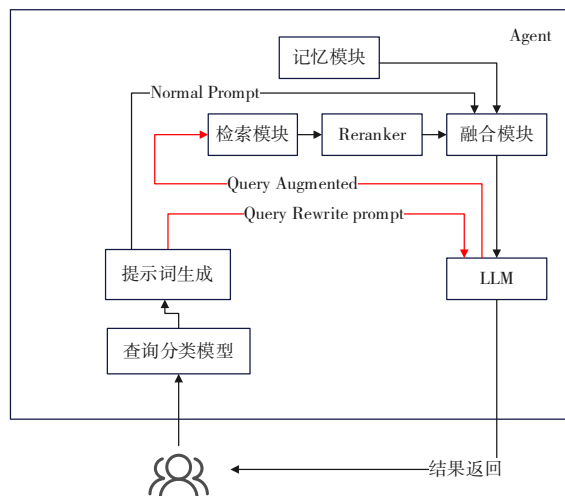


Fig. 1 Large language model with text classifier

图 1 带文本分类器的大语言模型

2.2.2 基于强化学习的 LLM Agent

目前,模型存在一个明显的不足,即系统无法基于用户反馈进行优化。为此,本文通过分析上述框架,基于两方面的反馈信息优化系统:①利用用户反馈信息调整信息融合模块,以确定检索模块的排序输出和记忆模块排序输出的优先级;②记忆模块可利用用户对相关交互结果的满意度,更新记忆内容、调整排序结果。

为此,本文在 LLM Agent 框架中引入强化学习模块,结合用户反馈信息调整记忆模块的更新策略与融合模块的融合策略,提出了基于增强学习的增强大语言模型,如图 2 所示。该结构与原模型的主要差别为图中灰色部分,即强

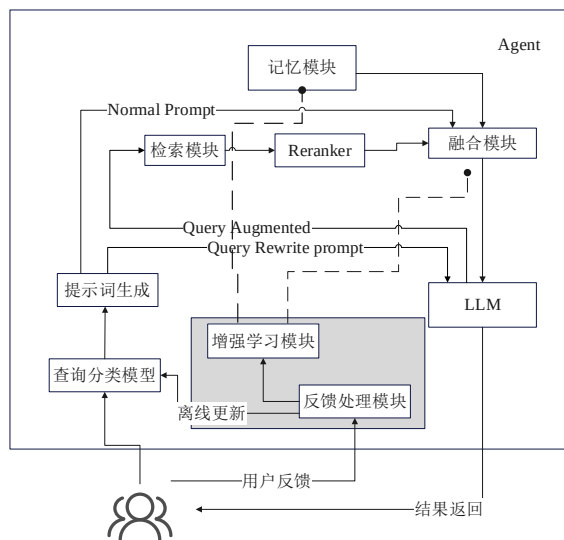


Fig. 2 Enhanced large language model based on reinforcement learning

图2 基于强化学习的增强大语言模型

化学习机制。

由图2可见,用户将Agent返回的结果进行满意程度判定,选项包括:{"非常满意" "满意" "基本满意" "不满意"},还能在模型返回的多个候选项中选择一个最佳选项,通过系统进行记录。

同时,为了优化记忆模块中的记忆内容,在记忆模块保存的内容中添加用户满意指标{"是否满足" "是否选中"}来记录模型反馈情况。此时,模块的问题将不单独依赖时间戳实施排序,而是基于用户满意指数(Satis)与时间戳实施综合排序,并将排序结果提交给融合模块。具体公式如式(2)所示。

$$Rank_{Mem} = \alpha * Satis + (1 - \alpha) * (Cur_Time - Stamptime) * K \quad (2)$$

式中: K 为常数; α 为权重系数。

权重系数根据用户反馈结果实施动态调整,对倒数排名融合(Reciprocal Rank Fusion, RRF)算法进行优化,主要对融合过程中的 k_i 值进行调整。如果最终用户的反馈结果多次来源于同一个选项,则应减小相应的 k_i 值,提升其在综合排序中的权重。

3 教务查询系统的实现与评估

3.1 方案实现

为了验证本文框架的有效性,以学校教务信息查询系统为应用背景实现以下3种方案。

方案①:传统的检索增强融合方法,即借助LLM实施分类,先查询重写专用知识,再查询实时知识,以生成结果并返回。

方案②:以LLM Agent框架为基础,集成文本分类器实施分类,根据不同分类结果进行相应的处理,如果涉及通

用知识将直接调用LLM返回结果,如果属于专用知识则查询重写,通过查询知识、融合结构返回结果。

方案③:在方案②的基础上,引入强化学习模块调整记忆排序和排序融合的权重。

图3为本文系统的具体框架,以星火大语言模型为后台大语言模型,前端模块提供了语音、手势等多种交互方式。实践证明,星火大语言模型的自然语言处理能力较强,提供了丰富的接口,可直接对接外部RAG的相关模块。

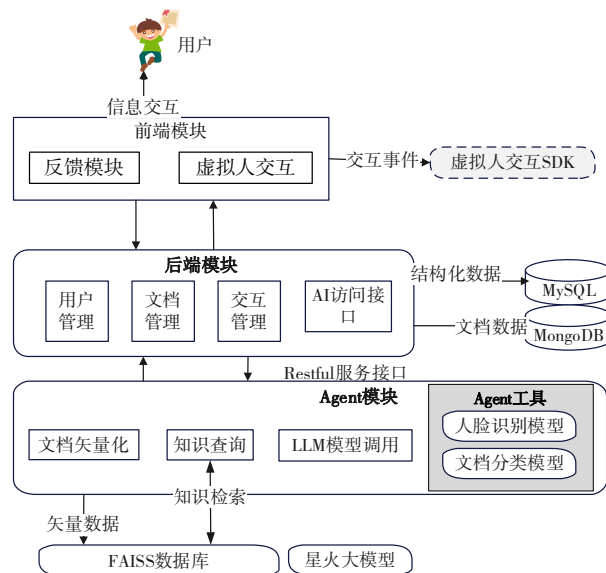


Fig. 3 Specific framework of the system

图3 系统具体框架

系统在星火大模型所提供的大量通用知识外,还收集了近两年某学院教学过程中涉及的教学资料,包括开课计划表、教学进度表、教学大纲、教学PPT等相关的教学资料及教务系统设计与使用文档等相关资料。将资料分块后,以BGE(<https://github.com/FlagOpen/FlagEmbedding>)的嵌入编码存储在FAISS数据库中,查询向量检索也以数据块为单位。Agent系统以Langchain为基础平台开发了相关的Agent架构,引入反馈模块收集用户对交互结果的反馈信息来优化系统。

3.2 模型评估

3.2.1 评估方案与数据概况

本文收集了系统实际运行过程中两个月内学生在各方案中的使用情况及反馈数据。反馈数据是用户使用系统后对查询结果的满意程度,从低到高可划分为:不满意、基本满意、满意、非常满意,用户在查询交互结束后给出。

由于用户查询的随意性,交互问题种类繁多,本文选择用户交互最频繁的30个问题作为测试问题,如表1所示。由此可知,前3类都与专有知识库信息相关,超过了总数的一半,在交互过程中各类问题交互的频率次数(见表2)、交互频次基本相近。

Table 1 Problem classification and quantity

表 1 问题分类与数量

类型编号	问题类别	数量
1	系统使用相关	5
2	学生日常事务相关	8
3	本学期教学相关	5
4	课程教学内容相关	12

Table 2 Distribution of various issues in each plan

表 2 各方案中的各问题的分布情况

类别	1	2	3	4	合计
方案①	290	421	300	797	1 808
方案②	322	430	278	870	1 900
方案③	315	450	305	725	1 795

3.2.2 评估结果

本文基于上述测试评估方案和测试数据集,对系统交互情况进行 3 个方面的分析,包括用户满意度、交互响应时间及满意率随时间的变化情况。

(1)用户满意度分析。在 3 种方案中,收集到的用户满意度反馈占比如表 3 所示。由此可知,在 3 种方案中用户满意的反馈均在 89% 以上;在方案②中,因采用文本分类机制替换 LLM 实施分类,在满意率上稍有下降;在方案③中,加入强化学习机制后长期的平均满意度相较于方案①、方案②更高。

Table 3 Proportion of satisfaction of each plan

表 3 各方案的满意度占比情况

分类	方案①	方案②	方案③
不满意	0.10	0.11	0.08
基本满意	0.35	0.35	0.36
满意	0.39	0.38	0.39
非常满意	0.16	0.16	0.17

(2)交互响应时间。本文分别统计了 3 种方案中所有有效查询中不同种类问题的平均响应时长(单位为 s),如表 4 所示。从时间角度而言,由于方案②通过简单文本分类模型对交互内容进行分类处理,因此降低了完成一次交互所需的时间;方案③在方案②的基础上引入了强化学习机制,使模型的整体交互效率有所下降。

Table 4 Average response time of various problems under different schemes

表 4 不同方案下各类问题的平均响应时间 (s)

方案	1类	2类	3类	4类	平均值
方案①	3.333	3.456	3.672	3.567	3.507
方案②	2.305	2.147	2.406	2.073	2.233
方案③	3.211	2.885	3.234	2.173	2.876

(3)满意率随时间的变化情况。为了验证强化学习算法的有效性,将 3 个系统部署应用了两个月左右,并以周为时间段记录学生使用系统的满意度,如图 4 所示。由此

可见,从满意度上来看,方案①与方案②的用户满意度基本保持不变;方案③引入了强化学习机制,通过动态改进记忆的排序方法和融合模块调整权重,使得查询结果能更好地满足用户查询需求,用户查询满意度随着运行时间增加缓慢提升。

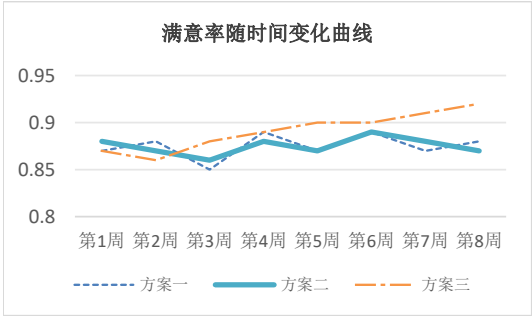


Fig. 4 The change of satisfaction rate over time
图 4 满意率随时间变化情况

4 结语

将 RAG 与大语言模型应用于教务查询系统,是实现教学智能化的一个重要发展方向。本文为了灵活支持不同类型的查询,减少大语言模型 token 消耗,以 LLM Agent 架构为基础,引入文本分类器分类交互内容,根据分类情况进行不同处理。同时,为了持续优化系统,在架构中引入强化学习机制,利用用户反馈调整 Agent 的记忆存储排序及融合模块调整权重,以更好地满足用户的需求。

未来,一方面将根据实际应用过程中存在的交互需求,细化相关的 Agent 处理流程,实现更复杂的查询应用逻辑;另一方面,随着交互增多,用户反馈系统将利用更多的用户反馈信息优化训练文本分类系统,以进一步提升文本分类的准确率和用户满意度。

参考文献:

[1] YIN S, FU C, ZHAO S, et al. A survey on multimodal large language models[DB/OL]. <https://arxiv.org/pdf/2306.13549v1>.

[2] EXTANCE A. ChatGPT has entered the classroom: how LLMs could transform education. [J]. Nature, 2023, 623(7987):474-477.

[3] REID M, SAVINOV N, TEPLYASHIN D, et al. Gemini 1. 5: unlocking multimodal understanding across millions of tokens of context [DB/OL]. <https://arxiv.org/abs/2403.05530>.

[4] ANONYMOUS. Video generation models as world simulators [EB/OL]. <https://openai.com/research/video-generation-models-as-world-simulators>.

[5] ANONYMOUS. Anthropic the claude 3 model family: opus, sonnet, haiku [EB/OL]. <https://paperswithcode.com/paper/the-claude-3-model-family-opus-sonnet-haiku>.

[6] SUN Y, WANG S, FENG S, et al. Ernie 3. 0: large-scale knowledge enhanced pre-training for language understanding and generation[DB/OL]. <https://arxiv.org/abs/2107.02137>.

- [7] LIU A, FENG B, XUE B, et al. Deepseek-v3 technical report[DB/OL]. <https://arxiv.org/abs/2503.23130>.
- [8] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 9459–9474.
- [9] GAO Y, XIONG Y, GAO X, et al. Retrieval-augmented generation for large language models: a survey [DB/OL]. <https://arxiv.org/pdf/2312.10997>.
- [10] FAN W, DING Y, NING L, et al. A survey on rag meeting LLMs: towards retrieval-augmented large language models [C]// *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024: 6491–6501.
- [11] MA X B, GONG Y Y, HE P C, et al. Query rewriting for retrieval-augmented large language models [DB/OL]. <https://arxiv.org/pdf/2305.14283>.
- [12] HE J X, NEUBIG G, TAYLOR B K. Efficient nearest neighbor language models [C]// *Association for Computational Linguistics*, 2021: 5703–5714.
- [13] KELVIN G, KENTON L, ZORA T, et al. Retrieval augmented language model pre-training[C]// *International Conference on Machine Learning*, 2020: 3929–3938.
- [14] WANG L, YANG N, WEI F R. Query2doc: query expansion with large language models [C]// *Association for Computational Linguistics*, 2023: 9414–9423.
- [15] MA X B, GONG Y Y, HE P C, et al. Query rewriting for retrieval-augmented large language models [DB/OL]. <https://arxiv.org/pdf/2305.14283>.
- [16] YU W H, ZHANG Z H, LIANG Z W, et al. Improving language models via plug-and-play retrieval feedback [DB/OL]. <https://arxiv.org/pdf/2305.14002>.
- [17] YANG H Y, LI Z T, ZHANG Y, et al. PRCA: fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter[C]// *Association for Computational Linguistics*, 2023: 5364–5375.
- [18] XI Z, CHEN W, GUO X, et al. The rise and potential of large language model based agents: a survey [DB/OL]. <https://arxiv.org/pdf/2309.07864>.
- [19] ANONYMOUS. Large language models revolutionized AI LLM agents are what's next [EB/OL]. <https://research.ibm.com/blog/what-are-ai-agents-llm>.
- [20] SHEN Y, SONG K, TAN X, et al. HuggingGPT: solving AI tasks with ChatGPT and its friends in hugging face [DB/OL]. <https://arxiv.org/abs/2303.17580>.
- [21] ZHANG Y, SUN R, CHEN Y, et al. Chain of Agents: large language models collaborating on long-context tasks [DB/OL]. <https://arxiv.org/pdf/2406.02818>.
- [22] DAN Y H, LEI Z K, GU Y Y, et al. EduChat: a large-scale language model-based chatbot system for intelligent education [DB/OL]. <https://arxiv.org/pdf/2308.02773>.
- [23] HUANG L Q, ZHANG J Q, WU F, et al. Generative AI-assisted teaching Q&A method and system using LLM-based knowledge graphs [P]. *China*, 202411809280.0, 2025-03-21.
- 黄立群, 张佳琦, 吴凡, 等. 基于 LLM 的知识图谱生成式辅助教学答疑方法及系统 [P]. *中国*, 202411809280.0, 2025-03-21.

(责任编辑:刘嘉文)