



计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331, CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目： 基于大语言模型的智能问答系统研究综述
作者： 任海玉，刘建平，王健，顾勋勋，陈曦，张越，赵昌瑛
网络首发日期： 2024-12-30
引用格式： 任海玉，刘建平，王健，顾勋勋，陈曦，张越，赵昌瑛. 基于大语言模型的智能问答系统研究综述[J/OL]. 计算机工程与应用.
<https://link.cnki.net/urlid/11.2127.TP.20241227.1952.011>



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

基于大语言模型的智能问答系统研究综述

任海玉¹, 刘建平^{1,2}, 王健³, 顾勋勋¹, 陈曦¹, 张越¹, 赵昌瑛¹

1.北方民族大学 计算机科学与工程学院, 银川 750021

2.北方民族大学 图像图形智能处理国家民委重点实验室, 银川 750021

3.中国农业科学院 农业信息研究所, 北京 100081

摘要: 智能问答是自然语言处理中的一个核心的子领域, 旨在理解并回答用户提出的自然语言问题的系统。传统的问答系统通常依赖于预定义的规则和有限的语料库, 无法处理复杂的多轮对话。大语言模型是一种基于深度学习技术的自然语言处理模型, 拥有数十亿甚至上千亿个参数, 不仅能够理解和生成自然语言, 还能显著提升问答系统的准确性和效率, 推动智能问答技术的发展。近年来, 基于大模型技术的智能问答逐渐成为研究热点, 但对该领域的系统性综述仍然较为欠缺。因此, 本文针对大模型的智能问答系统进行系统综述, 首先介绍了问答系统的基本概念和数据集及其评价指标; 其次介绍了基于大模型的问答系统, 其中包括基于提示学习的问答系统、基于知识图谱的问答系统、基于检索增强生成的问答系统和基于智能代理的问答系统以及微调在问答任务中的技术路线, 并对比了五种方法在问答系统中优缺点和应用场景; 最后对于当前基于大语言模型的问答系统面临的研究挑战和未来发展趋势进行了总结。

关键词: 大语言模型; 智能问答; 自然语言处理; 检索增强生成; 提示学习; 知识图谱

文献标志码:A 中图分类号:TP391 doi: 10.3778/j.issn.1002-8331.2409-0300

Research on Intelligent Question Answering System Based on Large Language Model: A Survey

REN Haiyu¹, LIU Jianping^{1,2}, WANG Jian³, GU Xunxun¹, CHEN Xi¹, ZHANG Yue¹, ZHAO Changxu¹

1. College of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China

2. The Key Laboratory of Images & Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, China

3. Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing, 100081, China

Abstract: Intelligent question answering is a core subfield in natural language processing, aiming at systems that understand and answer natural language questions posed by users. Traditional question answering systems usually rely on predefined rules and limited corpora and are unable to handle complex multi-round dialogues. Large language models are natural language processing models based on deep learning technology, with billions or even hundreds of billions of parameters. They can not only understand and generate natural language but also significantly improve the accuracy and efficiency of question answering systems, promoting the development of intelligent question answering technology. In recent years, intelligent question answering based on large model technology has gradually become a research hotspot, but a systematic review in this field is still relatively lacking. Therefore, this article conducts a systematic review of intelligent question answering systems based on large models. Firstly, it introduces the basic concepts of question answering systems, datasets, and their evaluation metrics. Secondly, it presents question answering systems based on large models, including those based on prompt learning, knowledge graphs, retrieval-augmented generation, and intelligent agents, as well as the technical route of fine-tuning in question answering tasks, and compares the advantages, disadvantages, and application scenarios of the five methods in question answering

基金项目: 北方民族大学中央高校基本科研业务费专项资金资助(2023ZRLG12); 国家自然科学基金项目 (32460444); 宁夏回族自治区重点研发计划 (2024YCZX0004)。

作者简介: 任海玉(2000—), 女, 硕士研究生, CCF 学生会员, 研究方向为自然语言处理、智能问答; 刘建平(1989—), 通信作者, 男, 博士, 讲师, CCF 会员, 研究方向为智能信息检索与推荐, E-mail: liujianping01@nmu.edu.cn; 王健(1971—), 男, 博士, 研究员, 研究方向为科学数据共享、信息行为、认知信息搜索; 顾勋勋(1998—), 男, 硕士研究生, 研究方向为自然语言处理、提示学习文本分类; 陈曦(2001—), 男, 硕士研究生, 研究方向为智能农业图像处理; 张越(2001—), 女, 硕士研究生, 研究方向为智能农业图像处理; 赵昌瑛(2002—), 男, 硕士研究生, 研究方向为智能信息检索与推荐。

systems. Finally, it summarizes the current research challenges and future development trends of question answering systems based on large language models.

Key words: Large Language Model; Intelligent Question-Answering; Natural Language Processing; Retrieval-Augmented Generation; Prompt Learning; Knowledge Graph

问答系统作为自然语言处理 (Natural Language Processing, NLP) 领域的重要研究方向之一, 因其能够让用户以自然语言提出信息需求并获得精准答案而备受关注^[1]。然而, 在处理自然语言问题时, 问答系统面临着语义复杂性和推理效率低的挑战。尽管这些系统能够分析问题并获取答案, 自然语言的灵活性和模糊性往往使其难以准确理解复杂问题的语义信息。

随着深度学习和自然语言处理技术的进步, 推动问答系统进入新时代, 能够更精准理解复杂问题, 支持多源知识表达和多轮交互, 实现智能化问答。传统的问答方法主要依赖于规则库、信息检索技术和浅层机器学习模型, 虽然在特定领域内表现较好且系统解释性强, 但在应对复杂语义和多轮对话时能力有限。近年来, 随着人工智能技术的迅速发展, 尤其是大模型 (如 GPT、BERT 等) 的出现, 智能问答系统的性能得到了显著提升。大语言模型 (Large Language Model, LLM) 的出现, 为理解和回答复杂语义问题提供了高效且准确的技术支持, 极大地推动了智能问答和文本生成的发展。Wu X 等人^[2]提出了一种针对开放域问答的段落提示调整方法, 通过微调参数有效的软提示, 结合问题-段落对的特定知识进行重新排序。Zong C 等人^[3]则提出了一个基于 LLM 的多角色代理框架, 专注于解决知识库问答 (Knowledge Base Question Answer, KBQA) 任务。Frisoni G 等人^[4]介绍了用于医学开放域问答的生成式框架 MEDGENIE, 通过提示生成医学语言模型的背景语境, 以此指导 LLM 进行问答。

目前, 在问答系统方面已经有一些综述性文章, Pan X 等人^[5]研究了如何利用外部知识来提升主题问答 (subject-area QA) 任务, 探索了开放域和封闭领域知识两种外部知识的利用方式。Mensio M 等人^[6]提出了一个基于 RNN 的上下文方法, 用于目标导向系统的意图分类, 验证了交互上下文在多轮问答中的重要性。王婷等人^[7]介绍了命名实体识别和知识问答技术的发展历程, 并指出基于大语言模型的方法是当前的主流研究方向。大语言模型的发展不仅提升了问答系

统的性能, 还为处理复杂和多轮对话开辟了新的途径。尽管大模型在问答领域展现了巨大的潜力, 仍有许多未解决的问题需要进一步研究, 目前缺乏系统性综述来全面总结大模型在问答系统中的应用及面临的挑战。

本文旨在填补这一空白, 系统地回顾和分析大模型在问答系统中的应用现状、挑战和未来研究方向, 通过综述和分析, 推动大模型在问答系统中的进一步发展。本文的主要贡献如下:

- 1、介绍了问答系统的重要性和研究背景及定义, 梳理问答系统的发展历程, 突出从早期基于规则的方法到大模型的转变。

- 2、对多个数据集及相关评价指标进行了系统的总结和分析, 通过对这些数据集和指标的深入研究, 发现它们在评估和比较不同问答系统方法方面的关键作用。

- 3、对基于大型语言模型的问答系统方法进行了系统的分类和分析。通过对现有的大模型技术在问答系统中的应用进行深入研究。按照提示学习、知识图谱、RAG、Agent 等方法对相关的文章进行总结, 系统性地分析了基于大语言模型的问答方法, 并详细讨论了它们优缺点及应用场景。

- 4、从优化提示模板设计、优化检索算法、动态知识更新和增强交互灵活性四个方面出发, 探讨了基于大模型的问答系统在未来的发展趋势和研究方向。

本文的其余部分组织如下: 第 2 节描述了本文所涵盖的相关文章的筛选程序; 第 3 节提供了关于问答系统的初步知识, 包括问答系统的基本框架及发展、以及问答系统的方法分类; 第 4 节总结了问答系统中常用的数据集及常用评价指标; 第 5 节对本文所涵盖文献中的基于大模型的问答系统方法进行了分类, 并对微调、提示学习、检索增强生成、大模型与知识图谱和 RAG 通用框架、Agent 等方法进行了深入讨论; 第 6 节讨论了当前基于大模型的问答系统的局限性和未来发展方向; 第 7 节对本文内容进行了总结。图 1 展示了本文的主要章节框架。

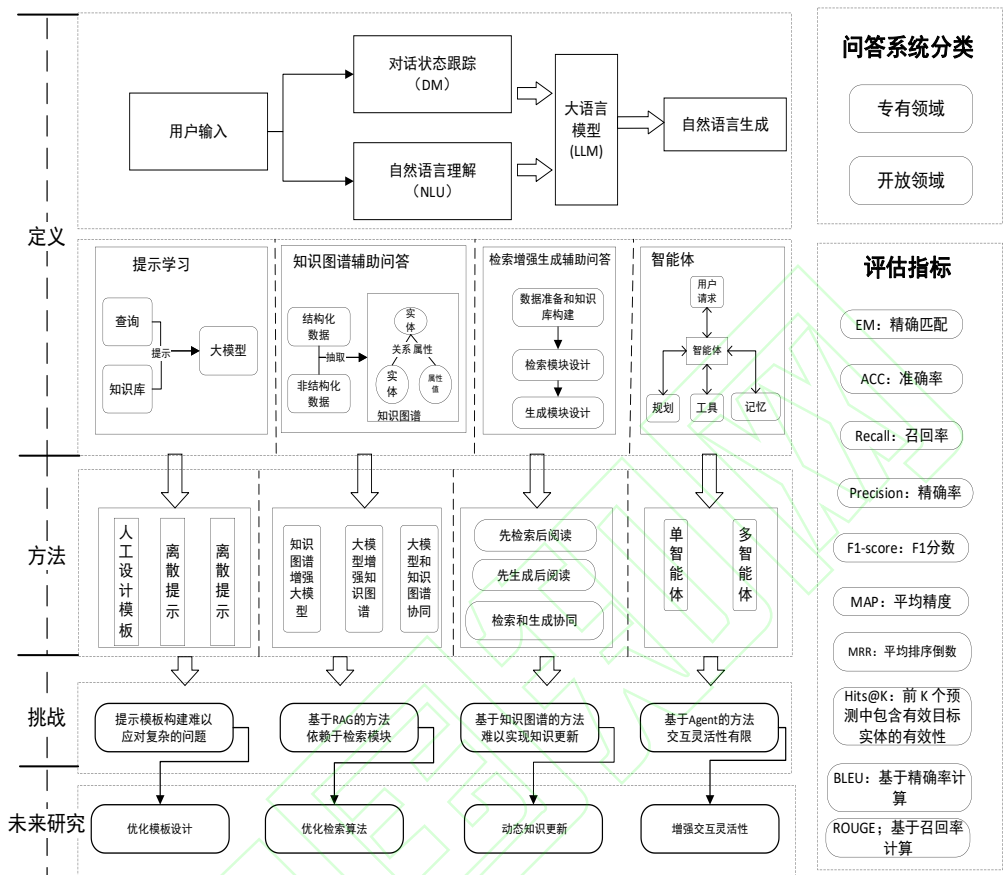


图 1 章节框架图

Fig.1 Chapter Framework Diagram

1 文献筛选过程

本文中主要包括 2012-2024 年的研究文章，其中包括会议、期刊等，这个时间段内，智能问答系统和大语言模型（如 GPT 系列）经历了快速的发展和应用，特别是在近几年内，大语言模型在问答系统中的应用取得了显著的进展。通过限制时间范围，确保了综述文献的相关性和前沿性。首先，“问答系统（Question Answering）”作为核心关键词，确保涵盖与问答系统相关的研究。随着大语言模型（LLM）的快速发展，其在自然语言理解与生成中的突破对问答系统影响巨大，因此 LLM 是检索的关键词之一。此外，为了提高问答系统的性能，“Prompt”作为引导大语言模型生成精确答案的重要工具也被纳入检索范围。随着信息检索技术的融合，“RAG（Retrieval-Augmented Generation）”方法，通过检索外部知识库增强生成答案的准确性，成为研究热点，故被选为关键词之一，智能问答系统常依赖于结构化数据如知识库（Knowledge Base）和知识图谱（Knowledge Graph），以增强推理能力和提供精准答案。因此，这两个关键词同样在文献检索中占有重要位置。智能问答系统与

智能代理（Agent）的关系日益密切，特别是在自动化决策和任务执行方面，因此将“Agent”作为重要关键词进行检索。

根据论文所要综述的内容及相关文献的整理，确定关键词：问答系统、Question Answering、Prompt、Knowledge Base、RAG、Knowledge Graph、Large Language Model、Agent，因此本文所涵盖的文献主要来自三个数据库：Web of Science、CNKI 和 IEEE Xplore，其中检索指令为：TS=(“Large Language Model”AND“Question Answering”)AND (“Prompt”OR“KnowledgeGraph”OR“RAG”OR“Agent”OR“Finetuning”)；TS=(“大语言模型”AND (“智能问答”OR“问答系统”))AND (“提示学习”OR“知识图谱”OR“检索增强生成”OR“智能体”OR“微调”)；上面的指令为一般检索指令但其中不同的关键词之间也会有组合在一起进行文献筛选，例如(“Large Language Model”AND“Question Answering”)AND (“RAG”AND“Knowledge Graph”)等情况，根据具体的需求进行关键词组合检索，以获得更加准确、更加相关的文献，使用 Google Scholar 和 arXiv 的搜索结果作为补充。如下图 2 所示，其中 CNKI 检索结果 69

条、Web of Science 检索结果 179 条、IEEE Xplore 检索结果 123 条。首先根据标题排除重复的、关键词缺失的论文，其次基于摘要和文章贡献进一步排除与本论文相关度不高的论文，最后得到人工筛选得到 71 篇

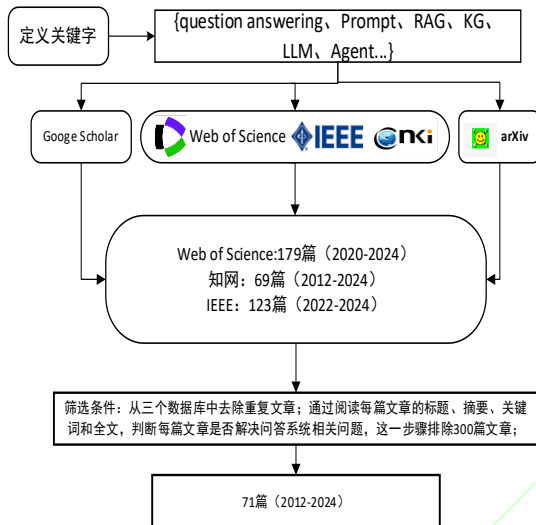


图 2 文献筛选过程

Fig.2 Literature screening process

2 问答系统基本概念

2.1 通用问答系统的基本框架

问答系统是一种人工智能应用，利用自然语言处理技术自动回答用户的问题，广泛应用于智能客服、虚拟助手、搜索引擎、在线教育和医疗咨询等领域。其核心目标是理解用户的自然语言输入，检索或生成相关信息，并以自然语言形式提供准确的回答。如图 3 所示，首先，将来自文档数据源（例如代码、各类文档、PPT 等）的内容会被拆分成一个个的文本块，并被输入到向量化模型之中。随后，这些文本块会通过检索增强生成（RAG）模块和知识图谱（Knowledge Graph, KG），与输入的问题进行精准匹配，从而找到最为相关的内容。

在此基础上，生成的语境信息与用户的查询会一同被传递至大语言模型（LLM）当中。通过 Prompt 以及增强后的上下文，最终生成高质量的答案。整个系统巧妙地将知识图谱、检索增强生成等技术结合起来，有效地提高了回答的准确性与相关性，为用户提供更加优质的问答服务。

2.2 问答系统的发展

问答系统研究起步于 20 世纪 50 年代的图灵测试，经历了早期基于规则和模板的问答系统、基于机器学习的问答系统、基于深度学习的问答系统以及最新的基于大模型的问答系统四个发展阶段。

20 世纪 60 至 70 年代，伴随着人工智能技术的崛起，基于规则和模板的早期问答系统应运而生，典型

相关文献。通过对文章进行分类统计，其中北大核心 1 篇，SCI（包括 1 区、2 区、3 区、4 区）9 篇，会议论文 40 篇，预印 21 篇。

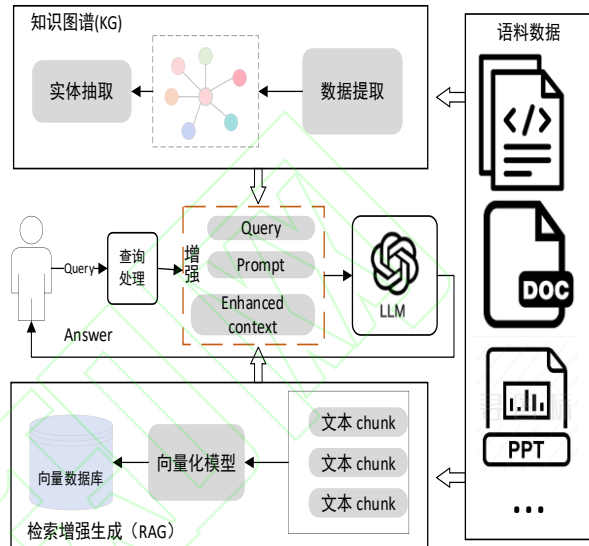


图 3 通用问答系统基本框架

Fig.3 Basic Framework of general Question-Answering System

的问答系统如 ELIZA 和 Lunar^[8]。ELIZA 模拟了心理治疗师的对话模式，而 Lunar 则用于回答关于月球岩石的科学问题。这一时期的问答系统虽然实现起来较为简单且响应迅速，但其问答范围较为狭窄，维护成本较高，且在扩展和更新方面存在一定难度。

20 世纪末，互联网的快速普及带来了海量的非结构化文本文档，如电子邮件、电子文档等，这些文档为信息检索、自然语言处理和数据挖掘等领域带来了新的挑战和机遇。问答系统的流程为从用户的自然语言问句中获取主题词，利用主题词在网络文档中搜索相关的文档，其中代表性系统为 ALICE^[9]聊天机器人。

21 世纪以来，随着深度学习的不断发展，问答系统的典型应用形式包括 FAQ 问答系统和社区问答系统。Sharma Y 等人^[10]探讨了基于深度学习的问答系统的发展与应用，文中详细分析了各种深度学习算法，包括长短期记忆网络（LSTM）和记忆网络等，通过将自然语言转化为语义表示，计算问题向量和答案向量之间的相似匹配得分确定最优答案。

2022 年底，OpenAI 公司推出了一款专注于对话问答生成的大语言模型 ChatGPT^[11]，标志着问答系统进入到一个全新时代。Hoffmann J 等人^[12]指出基于大语言模型的问答技术是一种利用大规模预训练语言模型来回答自然语言问题的方法，该方法的核心思想是先用海量文本数据预训练出一个大语言模型，然后在海量问答数据集上对该模型进行微调，使其适应下游问答任务，进而提供问答服务。

当前，随着大语言模型技术的不断发展，智能问答领域正经历着巨大的变革。大模型在问答系统等任

务中展现出了强大的潜能,使得计算机在理解和生成自然语言方面有了显著的提升,从而大大增强了问答系统处理用户提出的复杂问题的能力。当前,随着大语言模型技术的不断发展,智能问答领域正经历着巨大的变革。大模型在问答系统等任务中展现出了强大的潜能,使得计算机在理解和生成自然语言方面有了显著的提升,从而大大增强了问答系统处理用户提出的复杂问题的能力。目前,基于大型语言模型的智能问答系统已经取得了显著的进展。在金融、法律等专业领域,诸如 FinGPT^[13]、Chatlaw^[14] 等针对特定行业的大模型,推动了这些技术在行业内的实际应用,为专业知识的智能化处理提供了强有力的支撑。Frisoni G 等人^[4]探索在医疗开放域问答中,如何更有效地为模型提供上下文信息,以提高问答的准确性和性能。王婷等人^[7]针对果蔬农业技术领域的问答进行研究,旨在为果蔬种植户提供精准、有效的农技知识解答,助力农业生产。唐嘉等人^[31]利用大语言模型技术更好地服务于油气企业的制度管理和知识传播,推动企业数字化转型和智能化发展。Ossowski T 等人^[53]针对医学领域的视觉问答(Visual Question Answering, VQA)展开研究,旨在解决医学图像相关问题的回答任务,具有重要的研究意义和实际应用价值。王喆等人^[55]主要针对洪涝灾害应急管理领域的问答展开研究,旨在为洪涝灾害应急处置提供科学决策支持,提高应急指挥团队的决策效率。Lu P 等人^[60]科学领域的问答任务,旨在通过构建新的数据集和模型,提升人工智能系统在科学问题解答中的推理能力和可解释性。这些大模型不仅能够准确回答用户的问题,还能根据用户的需求提供个性化的建议和解决方案,提高了工作效率和质量。随着技术的不断进步,智能问答系统的性能和准确性将不断提高,能够更好地满足用户的需求。同时,智能问答系统也将与其他技术相结合,如人工智能、大数据、物联网等,为用户提供更加智能化、个性化的服务。

2.3 基于大模型的问答系统方法分类

在构建基于大语言模型(LLM)的问答系统时,本文综述的方法主要聚焦于提示学习(Prompt Learning)、知识图谱(Knowledge Graph, KG)和检索增强生成(Retrieval-Augmented Generation, RAG)。这些方法在构建高效、智能的问答系统中发挥着至关重要的作用。通过巧妙地将它们应用于 Agent(智能体),可以显著提升系统的性能和用户体验。如图4所示,微调方法通过主要包括全量微调、参数高效微调、少样本微调等技术。提示学习通过优化 Prompt 设计来提升 LLM 的响应精度。常见的 Prompt 模板构造方式包括以下三种:手动构造提示模板、连续提示模板构造和离散提示模板构造。知识图谱则通过整合实体和

关系,解决跨领域和复杂的问题。LLM 与 KG 的通用框架主要分为以下三类:KG 增强型 LLM、LLM 增强型 KG 和协同型 LLM+KG。检索增强生成通过结合外部知识库,提高 LLM 回答的准确性。主要范式包括三种:先检索后阅读(Retrieve-then-read)、先生成后阅读(Generate-then-read)、以及检索-生成协同(Retrieval-Generation Synergy)。

在结合 LLM、KG 和 RAG 技术时,通常采用以下策略:(1)以检索增强生成(RAG)为关键组件的混合框架,利用检索机制为生成模型提供相关的上下文信息,以增强生成过程;(2)通过 RAG 技术从向量数据库中检索与 Query 相关的信息,再从知识图谱中提取相关的上下文信息,以支持更准确的问答生成。

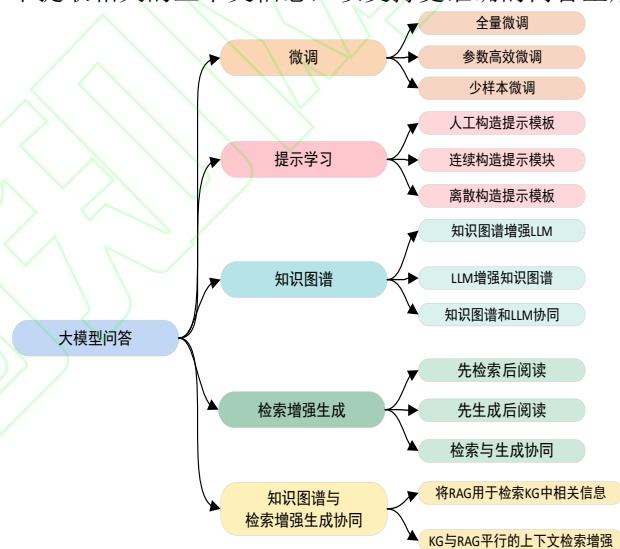


图4 方法分类

Fig.4 Method classification

3 数据集及其评价指标

随着一些大规模的 QA 数据集发布,推动了问答系统的发展,按照任务复杂度将数据分为简单问答数据集、复杂问答数据集,对话式问答数据集,其中划分数据集的主要依据是任务复杂度。简单问答数据集中主要包含单一事实型问题,通常通过一次查询就可以获得答案;复杂问答数据集包含需要多跳推理和复杂逻辑推理的问题,需要一定的推理过程才能得到答案;对话式问答数据集包含多轮对话和上下文信息的问答任务,涉及对话式问答,需要依据对话历史和多轮对话才能得到答案。

3.1 相关评价指标

关于问答系统的评价指标,常见的有 Accuracy、F1、Precision、Recall、EM、MAP、MRR 等指标。

精确匹配(EM)是衡量预测答案和实际事实之间的完美匹配,如式(1),其常用在问答任务数据 SQuAD 中,如(1)式:

$$EM = \frac{\text{Num}_{\text{right}}}{\text{Num}_{\text{total}}} \quad (1)$$

准确率 Acc(Accuracy)是指正确回答数占总问

题数的比, 如(2)式:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{N} \quad (2)$$

其中, TP 代表样本为真预测为真; FN 代表样本为真预测为假; FP 代表为样本为假预测为真; TN 代表样本为假预测为假; N 代表总问题数。

召回率(Recall)是指系统返回的正确答案数占所有正确答案数量, 如(3)式:

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{TN})} \quad (3)$$

精确率(Precision)是指问答系统所提供的答案中正确答案的比例, 如(4)式:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (4)$$

F1 Score 是指综合考虑精确率和召回率的指标, 如(5)式:

$$\text{F1} = 2 * \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

平均精度 MAP (Mean Average Precision) 是指多个问题的平均精度的平均值, 评估系统返回的答案排序质量, 如(6)式:

$$\text{MAP} = \frac{1}{Q} \sum_{i=1}^Q \text{Average Precision}(q_i) \quad (6)$$

平均排序倒数 MRR (Mean Reciprocal Rank) 是衡量检索结果排序质量, 如(7)式:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \quad (7)$$

式(7)中, MRR 经常被用作多结果问题中的一种度量方法, 其中 Q 表示查询总数, rank_i 是查询的序列。

Hits@K 是指正确答案出现在前 K 个返回结果的比例, 如(8)式:

$$\text{Hits@K} = \frac{\text{Num}_{\text{right in TOP K}}}{\text{Num}_{\text{total}}} \quad (8)$$

式(8)中, 通过调整 K 值, 可以评估问答系统中不同返回数量下的性能。

BLEU 适用于评估机器翻译质量的指标, 广泛应用于问答系统领域, BLEU 值越高, 表示生成的回答与参考回答越相似, 即回答的质量越高, 如(9)式:

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (9)$$

其中, BP 为长度惩罚因子, w_n 是针对不同的 N-gram 的权重, p_n 为修正精度, N-gram 是指一个语句里面连续 n 个单词组成的片段, N 为片段长度。

ROUGE 主要是基于召回率的计算, 是衡量生成的回答中包含了多少参考回答中的信息。

3.2 数据集介绍

本文针对现有的部分问答数据进行梳理, 将问答数据集根据任务复杂度的不同, 可以划分为多个类别。简单问答是指在给定的上下文中, 问题和答案之间具有明确且直接的对应关系, 通常是关于事实性和静态信息的查询; 复杂问答涉及多步骤推理和整合多个数据源来回答问题, 问题具有模糊性, 回答可能依赖于不同的上下文信息或外部知识库; 会话式问答是指基

于多轮对话的问答, 需要根据前文的对话历史进行推理, 维持话题的连贯性。事实上, 一些数据集可能同时包含简单问答、复杂问答和会话式问答。因此, 在数据集的实际使用中, 发现类型间的交叉是存在的, 尤其是在多轮对话的情境下, 简单问答与复杂问答的界限可能并不十分清晰。但在实际的数据集中, 由于问题表述的模糊性、数据标注的主观性以及问答系统处理方式的多样性等因素, 很可能存在不同问答模式类型交叉的情况。如下表 1 所示, 主要包括简单问答数据集、复杂问答数据集以及会话式问答数据集, 从数据集来源、特点、领域等方面总结数据集, 同时指出可能存在的交叉情况。SimpleQuestions^[15]是一个针对简单问题而构建的数据集, 它采用人工标注的方法根据知识库中的事实生成对应的问句, 并且以 Freebase 作为答案来源。WikiMovies^[16]是一个在电影领域包含原始文本和预处理知识库的 QA 数据集。WebQuestions^[17]数据集本身共包含 5810 条(问题, 答案)对, 其中简单问题占比在 84%, 复杂的多跳和推理问题相对较少。Natural Questions^[18]需要阅读整篇维基百科相关词条的文章, 更具挑战性。GrailQA^[19]需要对问题进行拆解和重新组合得到答案, 适合评估复杂逻辑推理能力。HotpotQA^[20]包含多样化、可解释的多跳问题, 要求模型能够理解和整合多个文档的信息。MetaQA^[21]是一个大规模的多跳 KGQA 数据集。ComplexQuestions^[22]处理需要复杂推理、多步推理以及跨文档信息整合的问答任务。WebQuestionsSP^[23]通过使用相应的 SPARQL 查询语句注释每个答案并删除模糊、不清楚或无法回答的问题来增强原始数据集。CWQ^[24]是一个用于测试模型回答复杂问题的数据集, 包含了大量复杂的问题, 这些问题需要在多个 web 片段上进行推理。QALD^[25]将自然语言问题转化成 SPARQL 查询以及知识推理。KQA Pro^[26]引入了一种组合式和可解释的编程语言来表示复杂问题的推理过程。CoQA^[27]基于新闻和文学作品, 每一轮问答包含一个问题和答案, 依赖于对话历史。QuAC^[28]基于对话文章, 要求模型理解对话的上下文, 并据此构成多轮对话。SQuAD^[29]是一个阅读理解数据集, 在文章中提出问题, 每个问题的答案都来自相应的阅读文章或一段文本。WikiQA^[30]由一组问题-答案对组成, 收集并注释以用于开放域 QA 研究。

表 1 常用数据集
Table 1 Common datasets

类型	数据集名称	来源	特点	领域	语言	评测指标
简单问答	SimpleQuestions	Freebase	每个问题都与相应的事实匹配	开放域	EN	Accuracy, F1
	WikiMovies	Wikipedia	问题通过使用知识库和相应的文档集来回答	非开放域	EN	hits@k
	SQuAD	Wikipedia	问题必须可由段落中一段连续文字来回答	开放域	EN	F1,EM
	WikiQA	Wikipedia	查询问题来自真实用户，候选句子直接从维基百科页面选择	开放域	EN	Precision,Recall,F1,MAP,MRR
	WebQuestions	Google	问题通常是事实性问题，可以直接获取	开放域	EN	F1,Precision,Recall
复杂问答	Natural Questions	Google 匿名查询和 Wikipedia 页面	问题是复杂多义的，需要在很长的文章中找到答案	开放域	EN	F1
	GrailQA	Freebase	需要对问题进行拆解和重新组合得到答案	开放域	EN	EM, F1
	HotpotQA	Wikipedia	多样化、可解释的多跳问题解答	开放域	EN	Hits@K
	MetaQA	WikiMovies Database	多步推理的复杂问题	非开放域	EN	F1
	ComplexQuestions	Freebase	处理那些需要复杂推理、多步推理以及跨文档信息整合的问答任务	开放域	EN	F1
	WebQuestionsSP	Freebase	用于研究语义解析和知识图谱问答	开放域	EN	F1
	CWQ	Freebase	多步推理和跨多个信息源的问题	开放域	EN	F1
	QALD	Dbpedia	利用关联数据来回答自然语言问题	开放域	EN	Precision,Recall,F1
	KQA Pro	Wikidata	对复杂的知识库问答进行深度分析推理能力	开放域	EN	Accuracy
	CoQA	新闻、文学作品	每一轮问答包含一个问题和答案，依赖于对话历史	对话	EN	F1
会话问答	QuAC	对话文章	理解对话的上下文，构成多轮对话	开放域	EN	F1
交叉	WebQuestions、Natural Questions、MetaQA、WebQuestionsSP 等数据集，包含了简单问答和复杂问答的交叉。					

4 基于大语言模型的问答系统

4.1 微调方法及范式

微调（Fine-tuning）是自然语言处理（NLP）任务中广泛使用的技术，其基本思想是在大规模通用数据集上预训练好的模型（如 BERT、GPT 系列等）基础上，进一步调整模型参数，以使其适应特定任务需求。如图 5 所示，全量微调(Full Fine-tuning)是对预训练模型的所有参数进行微调，预训练模型的所有层和参数均会被更新或优化，从而适应目标任务的需求。参数高效微调(Parameter-Efficient Fine-Tuning)旨在通过较少地调整预训练模型的参数，来实现对特定任务的良好适配。少样本微调主要适用于标注数据稀缺的场景。它基于已经在大规模数据集上训练过并具备通用语言知识的预训练模型，收集少量的任务特定标注数据，然后利用这些数据进行微调。唐嘉等人^[31]研究 LLM 在油气企业知识问答系统中的应用，通过将微

训练与检索增强生成技术进行优缺点对比。陈俊臻等人^[32]提出了一种融合大模型微调和图神经网络的知识图谱问答方法，通过微调大型预训练语言模型和引入模糊集理论，提升了知识图谱问答系统在处理自然语言问句时的语义解析精度和推理能力。文森等人^[33]梳理大语言模型发展现状及参数高效微调策略，其涵盖 Adapter、Prefix Tuning 及 LoRA 等主流类别，各有千秋。如 Adapter 类在模型网络层间嵌入新模块精准适配任务；Prefix Tuning 类巧设可训练前缀优化模型输入；LoRA 类以低秩矩阵近似权重更新高效微调。张钦彤等人^[34]综述了 LLM 微调技术，涵盖其发展历程、四类主要方法（经典参数微调、高效参数微调、提示微调和强化学习微调）以及未来研究方向。

4.1.1 微调技术在问答领域中的应用

微调技术允许模型在特定领域或任务上进行针对性的优化，从而更好地理解 and 处理与该领域相关的问题。通过在大规模通用数据上预训练的模型基础上，

利用特定领域的少量标注数据进行微调，问答系统能够快速适应新的领域知识和语言表达方式。如表 2 所示，本文总结了微调技术在问答领域的相关论文。Malladi S 等人^[35]提出 MeZO 优化器，仅用前向传播微调语言模型减少内存消耗，可有效优化跨任务和规模的大型语言模型，能处理非可微目标，在多项任务中表现良好。Lv K 等人^[36]提出 LOMO 优化器可在有限资源下实现大语言模型全参数微调，降低内存使用并在下游任务中展现良好性能。Tanwisuth K 等人^[37]提出了 PromptOriented(PO) 方法，仅使用前向传播实现无监督微调，提升效率且在多种任务中表现良好。陈俊臻^[32]提出 GNN - KBQA 模型，通过大模型微调、模糊集增强逻辑形式、图神经网络推理等步骤，在通用和垂直领域数据集上实验，证明了该方法的有效性。Houlsby N 等人^[38]提出 Adapter Tuning 的迁移学习方法，可在保持模型性能的同时显著减少参数更新数量，提高参数利用效率。Lester B 等人^[39]提出 prompt tuning 方法，通过学习软提示来适配冻结的语言模型以执行特定下游任务，在保持模型性能的同时提高参数利用效率。Dettmers T 等人^[40]提出 QLORA 方法，在不降低性能的前提下，显著降低大语言模型微调的内存需求。

Zhang Q 等人^[41]提出 AdaLoRA，通过自适应分配参数预算，以奇异值分解形式参数化增量更新，并基于重要性评分动态调整秩，在多个任务和模型上实验，表现优于基线方法。Lin X 等人^[42]提出了 DEALRec，通过分层采样选择样本进行少样本微调，避免贪婪选择导致的数据覆盖不足和样本冗余问题，提高数据覆盖率。Han G 等人^[43]提出利用基础模型进行 FSOD，通过预训练的 DINOv2 作为视觉骨干网络和大语言模型进行上下文少样本学习，在多个 FSOD 基准测试中取得了领先性能。Li Z 等人^[44]提出 FlexKBQA 框架，利用大语言模型生成合成数据训练轻量级模型，并通过执行引导的自训练和固有推理增强模型，在少样本甚至零样本场景下表现优异。Mao K 等人^[45]提出 RAG-Studio，通过自对齐生成合成数据来微调 RAG 模型，使其适应特定领域，在多个领域的问答数据集中实验，表现优于基线方法，证明了该方法的有效性和合成数据的潜力。Siriwardhana S 等人^[46]提出了 RAG-end2end 方法，通过联合训练检索器和生成器，并引入辅助训练信号，使 RAG 模型能适应特定领域知识，在三个领域数据集上实验，验证了该方法有效性。

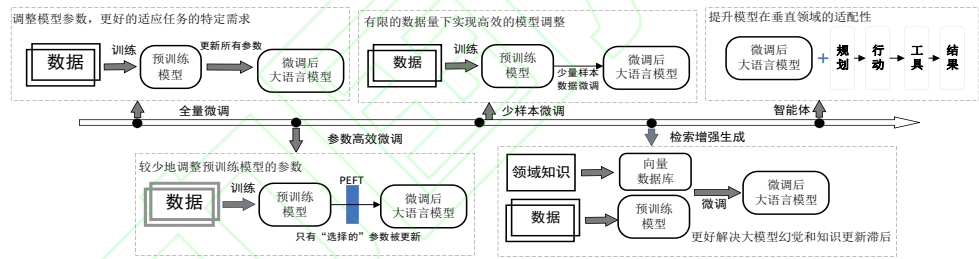


图 5 各类微调范式在问答场景中的演进

Fig.5 Evolution of various fine-tuning paradigms in question-answering scenarios.

表 2 微调技术

Table 2 Fine-Tuning Techniques

微调技术	代表方法	创新点	主要特点	解决问题
全量微调	MeZO ^[35]	实现了与跨多个任务的反向传播微调相当的性能	对整个预训练模型进行微调，预训练模型的所有层和参数都会被更新和优化	适用于任务和预训练模型之间存在较大差异的情况
	LOMO ^[36]	将梯度计算和参数更新融合在一起，以减少内存的使用		
	POUF ^[37]	对模型进行微调或对未标记的目标数据进行提示		
	GNN-KBQA ^[32]	采用高效微调策略方法对开源轻量级大模型进行微调		
参数高效微调	Adapter Tuning ^[38]	原始网络的参数保持不变，实现了高度的参数共享	较少地调整预训练模型的参数	需要快速调整模型以适应新任务或新数据的情况
	prompt tuning ^[39]	用于学习“软提示”来调节冻结的语言模型以执行特定的下游任务		
	QLORA ^[40]	冻结的 4 位量化预训练语言模型将梯度反向传播到 LoRA		
	AdaLoRA ^[41]	以奇异值分解的形式将增量更新参数化		
少样本微调	DEALRec ^[42]	识别为 llm 的少样本微调量身定制的代表性样本	有限的数据量下实现高效的模型调整	只用少量标记示例学习新任务的问题
	FSOD ^[43]	全类输入和查询图像建议的情境化 few-shot learning		
	FlexKBQA ^[44]	引入了一种执行指导的自我训练方法来迭代利用未标记的用户问题		

其他	RAG-Studio ^[45]	利用合成的数据对 RAG 系统进行微调，使检索器可以引入更精确	利用领域内相关知识对模型进行微调	解决大模型幻觉和知识更新之后的问题
	RAG-end2end ^[46]	引入了辅助训练信号来注入更多的领域特定知识		

4.2 基于提示学习的问答系统

大模型具有广泛的知识 and 强大的生成能力，问答任务重表现非常出色，但在特定领域问题和任务中，生成的回答往往不够准确，在复杂对话和多轮交互中，通过提示学习可以根据不同的领域和专业设置专门的提示，在问答任务中获取更加准确、专业的回答。提示学习（Prompt Learning）是一种基于预训练模型的线索，使得模型能够更好地理解人类的问题。通过将提示信息与预训练模型相结合，提高模型在特定领域或任务上的性能。在问答系统中，可以为模型提供更有针对性的引导，问答系统能够更好地适应相关领域的需求，从而能够为用户提供专业化的回复。提示学习包括，（1）提示设计，作为提示学习的核心，设计有效的提示可以引导模型生成准确且相关的回答；

（2）上下文理解，大型语言模型能够通过提示理解上下文，在处理复杂和开放式问题时提供更加符合用户的回答；（3）自适应性：提示学习具有高度的自适应性，可以在不需要大规模数据标注的情况下，通过调整提示来应对不同类型的问答任务。

4.2.1 提示学习方法

提示学习是一种通过设计和使用提示来引导预训练语言模型执行特定任务的方法，提示是一些预定义的文本片段，插入在输入数据中，引导模型生成所需的输出。大模型在问答系统中的发展非常迅速，在生成自然语言文本方面表现优异，能够生成流畅且连贯的回答，具有强大的上下文理解能力，能够根据上下文生成符合逻辑的回答，根据对话历史，与用户展开多轮对话交互。

表 3 提示词设计示例

Table 3 Examples of Prompt Design

任务描述：如何预防枸杞的常见病虫害？			
提示词设计类型	示例提示词	目的与应用	输出
简单指令提示	如何预防枸杞常见病虫害？	直接回答事实性问题	枸杞的常见病虫害预防包括：定期检查植株、及时清除病叶、施用有机肥料等方法。
增强上下文提示	请根据枸杞的生长环境，解释如何预防其常见病虫害？	在问题中加入特定背景信息	在宁夏等干旱地区，枸杞受到病虫害的威胁。通过适当灌溉、使用生物农药等手段。
角色扮演提示	作为一个植保专家，如何预防枸杞的常见病虫害？	设定角色，增强模型输出的专业性	作为植保专家，预防枸杞的常见病虫害应注重病害的早期诊断和综合治理。
多轮对话提示	如何预防枸杞的常见病虫害？那如果使用化学农药，有哪些注意事项？	设计多轮问答，便于处理复杂问题或获取详细信息	第一轮回答：“定期检查植株，及时清除病叶，使用生物农药。” 第二轮回答：“使用化学农药时应遵循标签上的使用剂量，并注意选择合适的农药种类。”

如上 3 表所示，给定任务描述，可以通过不同类型的提示词设计，引导模型生成更加符合有效、符合场景的回答。提示词的设计会引导模型生成更加准确的回答，因为精心设计的提示可以帮助模型更好地理解任务的上下文、识别关键信息，并从预训练知识中提取相关内容。在智能问答系统中，通过提供明确、结构化的提示词，能够减少模型的推理误差，确保其回答更加贴合用户需求。例如，通过在提示中指定角色或情境，模型可以更准确地模拟专业领域的语言风格和表达方式；而通过明确问题类型或任务目标，模型则能够聚焦于问题的核心，避免生成无关或不精确的回答。因此，良好的提示词设计不仅能提升模型的准确性，还能提高其在特定领域任务中的专业性和实

用性。

4.2.2 提示学习在问答领域中的应用

本小节分别从人工模板构建、离散模板构建、连续模板构建三种模板构建方法应用于智能问答进行介绍。如表 4 所示，对三种模板构建方式的优缺点及适用场景进行总结归纳。手工提示方法适用于简单明确的场景，而离散和连续方法更适合需要复杂提示优化的任务。离散方法在自然语言提示生成方面表现出色，而连续方法则更注重性能优化和自动化。三者在目标上是一致的：通过提示优化问答模型性能，但在实现方式、资源需求和适用场景上各有差异。

表 4 模板构建方法比较

Table 4 Comparison of template construction methods

模板构建方法	优点	缺点	适用场景
人工构建模板	通过直观的模板引导，语言模型能够在少量示例的基础上学习和推理。	耗时费力且难以适应复杂任务，限制模型的通用性。	适用于小规模、明确的任务。
离散构建模板	无需手动构建，提升效率准确性。	受限于文本语料库的质量和覆盖范围，可能无法找到所有的模板。	多任务或多样化场景。

连续构建模板	允许在模型的嵌入空间中执行连续提示，提高了灵活性。	需要更多的参数和计算资源，增加了模型的复杂性和训练时间成本。	复杂或开放域问答任务。
--------	---------------------------	--------------------------------	-------------

如下表 5 所示，本文总结了提示学习在问答领域相关的论文，根据不同的提示模板，将论文进行分类总结。

(1)手工构建的模板，旨在为特定任务或场景提供有效的指导和框架，利用模型的潜力，使其在处理问答任务时更加高效、准确。Mo T 等人^[47]提出了本文提出了一种创新的特定领域小样本表格问题回答方法，利用 Vicuna-13B 语言模型结合结构化提示模板与对比范例选择策略，有效提升了在资源受限条件下生成高质量 SQL 查询语句的能力。少样本场景下展现出卓越性能，为特定领域表格 QA 任务提供了高效且强大的解决方案。Wang L 等人^[48]提出了 GenRE 框架，通过融合生成性多轮问答与对比学习技术，创新性地解决了实体关系抽取任务中的难题。该方法利用模板化问题提示引导多轮问答，有效提升了抽取准确性和全面性，并通过对比学习优化模型性能。

(2)连续提示，是一种灵活且可调整的提示模板构建方法，其放宽了对模板内容的约束，允许模板内容具有自己独立的参数，这些参数可以根据下游任务的训练数据进行自动调整和优化。Jung M 等人^[49]引入了提示迁移 (Prompt Transfer, PoT) 的概念，旨在通过源任务的训练提示来更好地初始化目标任务的提示，通过一系列实验系统地探讨了问答任务中软提示 (soft prompt) 迁移的有效性。Liu Y 等人^[50]提出了一种创新的基于声明的提示调优 (DPT) 方法，通过将视觉问答 (VQA) 任务重新制定为与预训练目标一致的掩码语言模型 (MLM) 和图像文本匹配 (ITM) 任务，显著提升了预训练视觉语言 (VL) 模型在 VQA 任务上的泛化能力和性能。Liu H 等人^[51]提出了一种创新的异构图提示学习方法，用于提高社区问题回答 (CQA) 任务的效率和性能，将多种信息源构建为异构图作为图提示，直接引导预训练语言模型的生成过程，从而避免了对模型结构的复杂调整。Baek J 等人^[52]提出了 KAPING 框架，通过直接从知识图谱中检索相关知识并作为提示附加到问题输入中，显著提升了大型语言模型在零样本知识图谱问答任务上的性能。Ossowski T 等人^[53]提出了一种多模态提示检索 (MPR)

框架，利用预训练的 CLIP 模型和 KNN 搜索从数据集中检索相关图像-文本对作为多模态提示，显著提升了生成式视觉问答 (G-VQA) 模型在医疗领域的性能。Wu X 等人^[2]提出了一种段落特定提示调整方法，通过微调少量学习参数和结合段落特定知识，显著提升了大型语言模型在开放域问答任务中段落重新排序的性能。Cui C 等人^[54]提出了一种提示增强生成的模型，针对多模态开放问题回答任务，通过提示学习为图像生成补充描述，结合视觉与语言联合编码提升编码和检索性能，并利用前缀调优技术挖掘预训练模型背景知识，为答案生成提供额外信息。王喆等人^[55]将提示学习融入洪涝灾害应急决策的自动问答模型，通过构建问答数据集并引入 GPT2 与自动连续前缀提示，有效缓解了数据稀缺导致的过拟合问题，提升了模型在复杂灾害情景下生成富含决策信息答案的能力。Zhang Y 等人^[56]研究了机器人手术中视觉问题回答的双模态提示学习模型，通过引入视觉补充提示器和文本补充提示器，实现了在预测答案的同时精确定位图像中相关区域。Zhong W 等人^[57]提出了一种创新的基于结构化提示的统一问答预训练方法 ProQA，旨在通过单一模型解决多样化的问答任务，该方法利用结构化提示作为桥梁，将复杂的输入组件统一组织，同时建模各任务间的知识泛化及特定任务的知识自定义。

(3)离散提示，是在自然语言处理任务中，特别是指需要生成或补全文本的任务中，使用预定义的模板来引导模型产生预期的输出。Ma Z^[58]提出了 HybridPrompt，该方法是一种创新的视觉问答方法，旨在解决现有视觉语言预训练 (VLP) 方法在 VQA 任务中存在的局限性。Lazaridou A 等人^[59]提出了一种创新的少样本提示方法，利用互联网作为知识源来增强大型语言模型在开放领域问答任务中的性能。Lu P 等人^[60]提出了一种基于思维链 (CoT) 的多模态推理方法，通过训练语言模型生成详细的讲座和解释作为推理过程，显著提升了在科学问题回答任务上的性能。Mitra C 等人^[61]提出了 RetLLM-E 方法，通过结合文本检索与特定于课程内容的提示策略，显著提升了大型语言模型在学生论坛中回答课程相关问题的质量。

表 5 提示学习在问答系统中的应用

Table 5 The application of prompt learning in question answering system

提示类型	提示方法	解决问题	预训练模型	模型效果
人工设计的模板	Few-Shot Table Prompt ^[47]	在特定的领域中，表有很多列，对应的问题也复杂	Vicuna13B	ESQL 和 SMI-SQL 数据集 Acc (LF) :23.3% Acc (EX) :35.8% Acc (LF) :22.4% Acc (EX) :36.2%
				WebNLG 数据集 Prec:87.6% Rec:86.9% F1:87.2% NYT 数据集 Prec:93.7% Rec:91.6% F1:92.6%
	基于模板的问题提示 ^[48]	实体关系无法充分捕捉到实体间的复杂性和丰富性	GLM	自定义数据集 Prec:91.6% Rec:89.9% F1:90.7%
连续提示	Prompt transfer ^[49]	提示初始化会导致性能差异	T5	
	Declaration-based	预训练和微调目标不一致，需要	VinVL	GQA 和 VQAv2 数据集上:

离散提示	Prompt Tuning ^[50]	大量的标记数据进行微调		Acc:63.13% Acc:74.50%
	图像提示学习 ^[51]	充分利用外部知识库进行问答	BART	AntQA 数据集 BLUE:16.20% MSMplus 数据集 BLUE:9.47%
	检索增强提示 ^[52]	问答任务中存在知识不足和错误的问题	T5/GPT-3	WebQSP 数据集 Acc:47.94% Mintaka 数据集 Acc:32.27%
	多模态提示 ^[53]	在标记数据有限的低资源领域容易过拟合	T5	VQA-RAD 数据集 SLAKE 数据集
	特定段落提示 ^[2]	硬提示的性能受提示质量的影响,而微调模型则计算量大、耗时长	Llama-2-Chat-7B	NQ:36.89% SQuAD:46.04% TriviaQA:42.63% NQ:45.73% SQuAD:45.27% TriviaQA:42.53%
	提示增强生成 ^[54]	忽略了信息源中图像-文本对的语义对齐	OFA	WebQA 数据集 (All) :37.8% MultimodalQA 数据集(All):63.8%
	连续型前缀提示 ^[55]	提高洪灾处理的时效性和科学性	GPT-2	
	双模态提示学习 ^[56]	无法给出图像中相关内容的位置	CAT-ViL	EndoVis-2018 Acc:69.53% F-score:51.37% EndoVis-2017 Acc:49.57% F-score:37.17% PSI-AVA Acc:32.22% Recall:79.76%
	结构化提示 ^[57]	系统无法对任务之间的共性进行建模	ProQA	
	HybridPrompt ^[58]	上游预训练任务和下游问答任务不兼容		VQAv2 数据集 Acc(dev):76.12% Acc (std) :76.30%
	few-shot prompting ^[59]	LSLM 容易产生幻觉,即回答不正确或不准确	GOPHER LM	
	思维链提示 ^[60]	现有数据集规模小,领域多样性有限	UnifiedQA GPT-3	SCIENCEQA 数据集 Acc (GPT-3) :75.17%
	利用检索到的内容设计提示 ^[61]	知识库并非针对特定课程内容进行训练	LLaMA-2-13B-chat	100 个问题-答案对, ROUGE_1:19.3% ROUGE_2:4.5% BERTScore83.8%

4.3 基于知识图谱的问答系统

知识图谱 (Knowledge Graph, KG) 是一种基于图的数据结构,以符号的形式来描述现实中的知识信息及其之间的关系,是结构化的语义数据库,其组成的基本单位是“实体-关系-实体”三元组,以及实体机器相关属性值通过关系相互连接,构成知识的网状结构。知识图谱问答 (Knowledge Graph Question Answering, KGQA) 是根据用户提出的自然语言问题,从图谱中搜寻与其相关的实体,将找到的知识答案返回给用户。目前基于知识图谱的问答方法主要分为两类,第一类是基于语义解析的方法,通过符合逻辑形式表示问题,在知识图谱上执行获得最终的答案。第二类是基于信息检索的方法,该方法是指从整个知识图谱中抽取一个子图,随后基于子图进行问题的推理,并选取子图中排序较高的实体作为答案。

4.3.1 大模型与知识图谱在问答任务中的通用框架

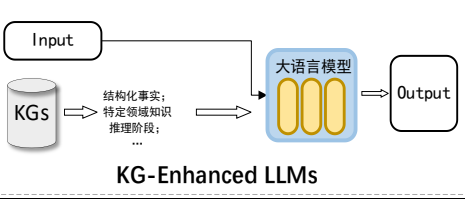
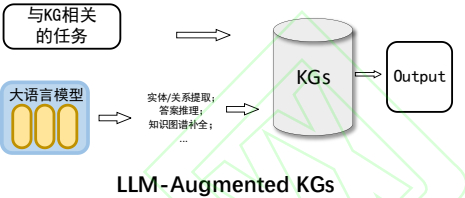
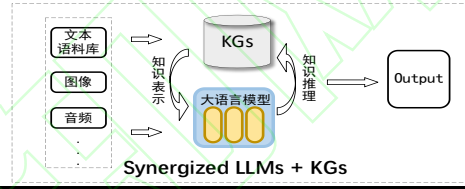
在大规模的语料库中进行预训练的大语言模型如 (BERT^[62]、T5^[63]等) 在自然语言处理 (NLP) 任务中表现的非常出色,具有强大的语言理解能力和广泛的知识储备,能够处理各种自然语言任务,在许多复

杂的任务中表现出巨大的潜力。但大语言模型无法有效捕捉到事实信息,经常通过生成事实不正确的陈述而产生幻觉^[64],大模型缺乏特定领域的知识和训练数据,导致在通用语料库上训练的 LLM 无法有效泛化到特定领域。为了解决上述问题,可将知识图谱与 LLM 合并在一起,利用知识图谱结构化的知识表达形式,通过提供外部知识来增强 LLM。但 KG 在本质上难以构建,目前的构建 KG^[65]方法不足以满足现实世界中动态变化的性质,其次 KG 中现有的方法是根据特定的领域知识所制定的,其通用性不能够满足所有领域的任务。因此利用 LLM 解决 KG 中所存在的问题是必要的,因此在表 6 中总结了 LLM 与 KG 通用的框架图。

KG-Enhanced LLM:利用 KG 来增强 LLM,在预训练阶段,将知识图谱融入大型语言模型增强模型在特定领域的知识能力。在推理阶段结合 KG 与 LLM,将用户查询首先通过大模型进行理解和解析,然后利用知识图谱进行精确的查询和推理。LLM-Augmented KG:LLM 用来解决与 KG 相关的任务。利用 LLM 的处理信息的能力对知识库中的文本语料进行处理,LLM 可用来处理文本语料库并提取关系和实体以构建 KG; Synergized LLM+ KG^[66]:将 LLM 和 KG 进行协同的融合,使用一个统一的框架结合两者的优势。

表 6 LLM 与 KG 在问答任务中的通用框架图

Table 6 General Framework Diagram of LLM and KG in Question Answering Tasks.

	框架名	描述	图示
LLM+KG	KG-Enhanced LLM	在 LLM 的预训练和推理阶段引入 KG，或为了增强 LLM 对学习的知识的理解	 KG-Enhanced LLMs
	LLM-Augmented KG	利用 LLM 完成不同的 KG 任务，如嵌入、完成、构建、图到文本生成和问题回答	 LLM-Augmented KGs
	Synergized LLM+KG	协同的方式增强 LLM 和 KG	 Synergized LLMs + KGs

4.3.2 知识图谱在问答领域中的应用

如表 7 所示，总结了知识图谱和大模型在问答任务中的相关研究成果，两者在问答任务中的结合有效提升了问答系统的鲁棒性和智能性，使其在面向特定领域或常识推理方面的表现更为出色。

(1)知识图谱增强大模型:大模型能够同时处理多个复杂任务，具有较好的泛化能力，但是大模型在处理一些问题是仍然存在一些不足之处，例如大模型的性能依赖于训练数据的质量和多样性，如果训练数据不足或者存在偏差，可能会学习到错误的知识和特征表示，知识图谱可以提供结构化信息来提升大模型在特定任务上的性能，提供更加准确的信息检索和知识发现。Sun J 等人^[67]提出了了 Think-on-Graph (ToG) 框架，该框架通过结合 LLM 和知识图谱，在特定领域内提升了问答任务的深度推理能力。Tan X 等人^[68]提出了 STRUCT-X 框架，一个通过“读取-模型-填充-反射-推理”五阶段流程增强大型语言模型在结构化数据上推理能力的创新方法。该框架利用图嵌入、知识检索模块、自监督模块和辅助模块，有效地整合了结构化数据，提高了 LLMs 在复杂推理任务中的表现。Wang Y 等人^[69]提出了一种新的知识图谱增强大语言模型的方法，称为“高效知识路径推理”(RoK)，旨在优化从知识图谱中选择推理路径的流程，RoK 通过链式思维 (CoT) 和 PageRank 算法来生成最可能包含答案的知识路径，从而提高了 LLM 在领域问题回答任务中的表现。Guo T 等人^[70]提出了 KnowledgeNavigator，通过引入外部知识显著提高了 LLM 在处理复杂任务时

的性能。其核心优势在于能够有效地结合 LLM 的自然语言理解和推理能力，通过优化提示和知识表示来提升整体性能。Jiang J 等人^[71]介绍了 StructGPT 框架，通过迭代阅读和推理框架，构建了专门接口以高效访问和过滤知识图谱、数据表 and 数据库等结构化数据，使 LLM 能基于收集的信息进行推理。Agrawal G 等人^[72]提出了一个新颖的方法 CyberGen，它结合了大型语言模型和网络安全领域知识图谱来自动生成问题和答案。这种方法通过三种提示技术 (Zero-Shot (ZS)、Few-Shot (FS) 和 Ontology-Driven (OD)) 来增强 LLMs 的输出，减少错误陈述，并提高事实推理能力。Zhang Q 等人^[73]介绍了一个名为 KnowGPT 的框架，它通过将知识图谱与大型语言模型结合，来提高 LLMs 在零样本 (zero-shot) 知识图谱问答任务中的性能。Baek J 等人^[52]介绍了一个名为 KAPING 的框架，它通过在大型语言模型的输入中直接增强知识图谱中与问题相关的知识，来提升零样本知识图谱问答任务的性能。Wang J 等人^[74]提出了 LPKG 框架，该框架通过从知识图谱中提取模式实例并将其转化为自然语言问题来增强大型语言模型在复杂问答任务中的规划能力。实验结果显示，在多个数据集上优于传统基线方法，尤其在较小的 LLMs 上效果显著。Wang F 等人^[75]介绍了 InfuserKI 框架，该框架旨在将知识图谱中的新知识高效整合到大型语言模型中，同时避免对已有知识的遗忘。通过检测 LLMs 中未知的知识，并利用知识适配器编码新知识能够在保留原始模型参数的同时，减少对现有知识的干扰，缓解知识遗忘问题。

Moiseev F 等人^[76]提出了 SKILL 方法,它通过在知识图谱的事实三元组上直接训练 T5 模型,使模型能够从结构化数据中学习知识,直接从知识图谱向语言模型注入知识的有效性,并展示了其在提高模型事实性方面的潜力。

(2)大模型增强知识图谱:知识图谱通过图的形式将实体和关系结构化表示,可以整合来自不同来源的数据,形成统一的知识库,但还存在一些不足之处,例如结构化表示方式在处理自然语言的复杂性和多样性方面存在局限,难以捕捉知识的变化,无法及时更新知识。大模型能够提供深度的文本理解和语义分析能力,帮助知识图谱更好地理解 and 利用文本信息。Xia T 等人^[77]提出了 LACT 框架,通过将复杂的一阶逻辑查询分解为二叉树结构,并结合逻辑感知的课程学习策略来提升大语言模型在知识图谱上进行复杂逻辑推理的能力。

Zhang Y 等人^[78]提出了 KnowPAT 框架,它通过构建两种偏好集合来提升大型语言模型在特定领域问答任务中的应用效果,确保回答既符合用户需求也能合理利用领域知识库。Gao Y 等人^[79]提出了 GenTKGQA,一个创新的两阶段生成式时间知识图谱问答框架,它利用大语言模型来处理涉及时间约束和动态结构化知识的问题。Zhang Y 等人^[80]提出了一种知识前缀适配器(KoPA),它能够有效地结合预训练的知识图谱结构嵌入和 LLM,通过将结构信息融入 LLM,可以显著提高其对 KGC 任务的推理能力。Agarwal D 等人^[81]提出了介绍了 BYOKG 系统,是一个通用的知识图谱问答系统,能够在任何知识图谱上进行操作,不需要人类标注的训练数据。Luo H 等人^[82]提出了 ChatKBQA 框架,通过结合精细调整的大型语言模型和无监督检索方法,显著提高了知识库问答任务的性能和效率。实验结果表明,该框架在处理复杂问题时表现出色。

(3)大模型与知识图谱协同:知识图谱与大模型协同在问答任务中取得了一定的成果,这种协同可以造就执行知识表征和推理的强大模型。Feng C 等人^[83]提出了一个 Knowledge Solver (KSL)的新方法,旨在通过利用大型语言模型的通用能力来搜索外部知识图谱中的领域知识。Wang H 等人^[84]提出了一种名为“Knowledge-tuning”的方法,旨在提高大模型在医疗领域的可靠性和准确性,通过结合准确的知识检索,知识调整模型能够更好地应对复杂医疗问题,并提供高质量的回答。Zhou T 等人^[85]提出了 CogMG 框架通过明确定义和完成相关知识三元组,解决了大型语

言模型与知识图谱集成中的两大问题:知识覆盖不完整和知识更新不同步。实验结果证明了 CogMG 在提高问答系统准确性和可靠性方面的有效性。Wen Y 等人^[86]提出的 MindMap 方法通过结合知识图谱和大型语言模型,有效提升了模型的推理能力和透明度,实验结果显示,该方法在复杂的医疗问答任务中表现出色,证明了其有效性和鲁棒性。

4.4 基于检索增强的问答系统

4.4.1 检索增强生成方法

检索增强生成(RAG)是一种结合了信息检索和生成模型的技术,用于提升自然语言处理任务的性能。将检索到的内容与用户问题一起输入到大模型中,在外部知识的加持下,大模型可以生成更加准确、高质量的回复^[87]。RAG 技术的优势在于两方面:首先,通过结合信息检索与生成,RAG 显著增强了回答的准确性和可靠性。传统生成模型(如 GPT-3)依赖训练数据,容易生成不准确或虚假的信息,而 RAG 通过从大规模知识库或文档中检索相关信息,再结合这些信息生成回答,提供了更为精确的结果。此外,RAG 的信息检索模块能够访问实时更新的数据库或互联网资源,确保提供及时和最新的信息,解决了纯生成模型在时效性方面的不足。

4.4.2 大模型与检索增强生成的通用框架

大语言模型在语言理解和生成方面展现了显著的能力,但仍然具有局限性,如幻觉和无法有效检索到关键信息,在预训练大模型时,并没有明确使 LLM 学习到如何高质量利用检索文本完成生成任务,对于长而复杂的检索文本,LLM 无法准确提取到关键信息。检索增强生成技术通过整合外部数据源的信息来丰富和细化 LLM 的生成能力,允许 LLM 在处理查询或生成文本时,动态地从外部数据库中检索相关内容。如下表 8 所示,本文总结了检索增强生成与 LLM 的通用框架图。

(1)retriever-then-read:从外部数据存储中检索相关文档,Shi W 等人^[88]提出了 REPLUG,是一种检索增强的语言建模框架,直接将检索到的文档作为输入添加到冻结的语言模型中。Ram O 等人^[89]提出了一种简化版的检索增强语言建模方法 In-Context RALM,通过将相关文档直接添加到输入中,无需修改 LM 架构或进一步训练,来提升语言建模性能。

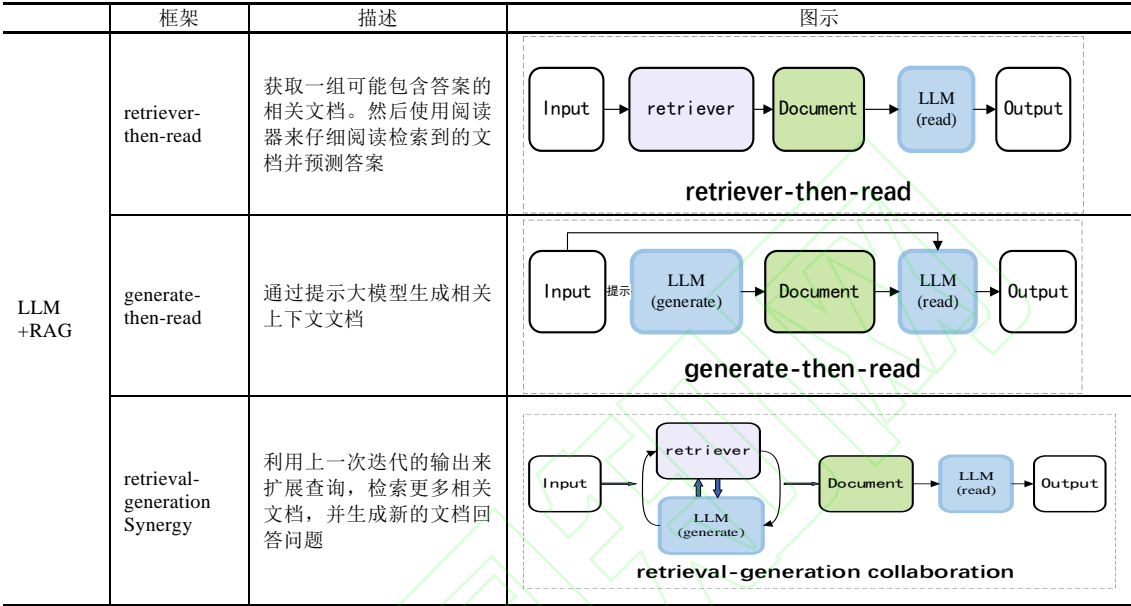
表 7 知识图谱在问答系统中的应用

Table 7 Application of knowledge graph in question answering system

分类	方法	解决问题	方法	模型效果(数据集、指标)
知识图谱增强大模型	TOG ^[67]	LLM 的幻觉问题	从 KGs 中提取多种多跳推理路径, 作为 LLM 推理的基础, 增强 LLM 对知识密集型任务的深度推理能力	CWQ(GPT-4) Hits@1:67.6% WebQSP Hits@1:82.6% GrailQA Hits@1:81.4% Simple Questions Hits@1:66.7%
	STRUCT-X ^[68]	表示结构化数据中嵌入的复杂知识的能力有限	利用结构化数据来增强 LLM 的交互和复杂推理能力	WebQSP Acc:75.13% Prec:73.40% Rec:77.25% F1:75.29% MetaQA Acc:79.63% Prec:78.27% Rec:77.53% F1:76.61% TriviaQA:EM(XXL + KELM):42.95% NaturalQuestions:EM(XXL + KELM):32.74%
	RoK ^[69]	多跳推理问题中难以获得高质量的答案	高效地从 KG 中选择知识路径, 充分激发 LLM 的推理能力	GenMedGPT-5k Pre:81.7% Rec:81.3% F1:81.5% CMCQA Pre:94.1% Rec:92.9% F1:93.4% WebQuestions Acc:80.5%
	Knowledge Navigator ^[70]	处理复杂推理和扩展逻辑序列任务中, 会出现幻觉	迭代检索和过滤知识图中的候选实体和关系, 提取相关的外部知识并转换为提示	MetaQA Acc:99.5% WebQSP Acc:83.5%
	StructGPT ^[71]	解决结构化数据的问答任务	结合专门的接口来操作结构化数据记录	WebQSP Hits@1:72.6% MetaQA Hits@1:97.1% TabFact Acc:52.2% WTQ Acc:65.6% Spider EX:77.8% Spider-syn EX:64.0% SpiderRealistic EX:72.0%
	CyberGen ^[72]	大模型在特定领域问答时, 会产生不准确的陈述	从知识图中增加结构化知识, 可以提高 LLM 的事实推理能力	CyberQ (OD) BLEU:10% ROUGE:39% CyberQ (ZS) BLEU:17% ROUGE:43% CyberQ (FS) BLEU:4% ROUGE:32%
	KnowGPT ^[73]	特定领域时出现的幻觉问题	从 KG 中提取知识, 将提取的知识自动转换为有效的提示。	CommonsenseQA Acc:81.8% OpenBookQA Acc:92.4% MedQA Acc:78.1%
	KAPING ^[52]	LLM 产生事实上错误的答案	检索与问题相似的 top-K 个三元组, 增强为提示的形式	WebQuestionsSP Acc:47.94% Mintaka Acc:32.27%
	LPKG ^[74]	分解复杂问题方面面临挑战	使用知识图衍生的规划数据来提高 LLM 的规划能力	HotPotQA EM:37.6% 2WikiMQA EM:37.2% MuSiQue EM:29.6% Bamboogle EM:30.4%
	InfuserKI ^[75]	LLM 在知识密集型任务中表现不佳	利用 transformer 内部状态来确定是否使用附加信息增强原始 LLM 输出	UMLS F1:88% PubMedQA F1:58% MetaQA F1:67%
大模型增强知识图谱	SKILL ^[76]	无法确定对知识的表达能力	提出了一种针对大模型的知识注入方法, LLM 直接从知识三元组中学习	TriviaQA:EM(XXL + KELM):42.95% NaturalQuestions: EM(XXL + KELM):32.74% FreebaseQA: EM(XXL + KELM):45.9% Wikidata-answerable QA EM(XXL + KELM):26.26%
	LACT ^[77]	无法共享知识来提高逻辑推理	将 KG 中包含的知识纳入训练语料库, 以激活 LLM 的相应知识, 并补充 LLM 在微调中缺失的相关知识	FB15K MRR:71.9% FB15K-237 MRR:44.4% NELL995 MRR:32.0%
	KnowPAT ^[78]	难以适应用户实际场景 QA	解决实际场景 LLM 应用程序的特定领域的 QA 任务	CPQA BLEU-1:21.87% RJUA-QA BLEU-1:25.61%
	GenTKGQA ^[79]	知识挖掘问题中的时间约束	子图检索和答案生成两个阶段指导 LLM 回答时态问题	CronQuestion Hits@1:97.8% Hits@10:98.3%
	KoPA ^[80]	结构信息和文本信息之间存在明显的语义区分	将 KG 嵌入到 LLM 中旨在实现 LLM 中的结构感知推理	UMLS Acc:92.56% CoDeX-S Acc:82.74% FB15K-237N Acc:77.65%
	BYOKG ^[81]	缺乏训练数据	将经典的逻辑编程语言集成到 LLM 中, 解决 KGQA 任务	GrailQA F1:46.47% MetaQA F1:73.45% Hits@1:80.57%
	ChatKBQA ^[82]	低效率的知识检索, 检索错误对语义解析的不利影响	微调 LLM 和候选逻辑形式中的实体和关系的无监督检索来生成	WebQSP F1:83.5% Hits@1:86.4% Acc:77.8% CWQ F1:81.3% Hits@1:86.0% Acc:76.8%
	KSL ^[83]	推理过程产生幻觉	LLM 利用自身强大的泛化能力从外部知识库中搜索基本知识	CommonsenseQA Acc:79.6% OpenbookQA Acc:81.6% MedQA-USMLE Acc:58.4%
	Knowledge-tuning ^[84]	领域知识有限而产生对医学事实的幻觉反应	用结构化的医学知识库为 LLM 有效地掌握领域知识并促进可靠的响应生成	cMedKnowQA Acc:86.7%
	CogMG ^[85]	知识不完整和知识更新错位	超出当前 KG 的知识范围, LLM 显式分解所需的知识三元组	KQA Pro Direct Answer Acc:40%
知识图谱与大模型协同	MindMap ^[86]	大模型存在幻觉难以吸收新知识	LLM 能够理解 KG 输入, 并结合隐性和外部知识进行推断	GenMedGPT-5k Pre:79.3% Rec:79.7% F1:79.5% ExplainCPE Pre:93.3% Rec:93.7% F1:93.5% CMCQA Pre:94.1% Rec:93.2% F1:93.6%

表 8 LLM 与 RAG 通用框架图

Table 8 LLM and RAG General Framework Diagram



(2)generate-then-read: 提示一个大型语言模型根据给定的问题生成上下文文档，接着读取生成的文档以生成最终答案。Yu W 等人^[90]提出了一个解决知识密集型任务的新视角，利用 LLM 取代文档检索器。Sun Z 等人^[91]提出了 RECITE，通过采样从 LLM 记忆中背诵一个或多个相关段落，然后产生最终答案。

(3)retrieval-generation Synergy: 旨在结合信息检索与文本生成的优点，以提高生成文本的真实性和相关性。Feng Z 等人^[92]提出了 ITRG，一个迭代检索生成协同框架，用于增强大型语言模型解决知识密集型任务的能力。Shao Z 等人^[93]提出了 ITER-RETGEN，将检索和生成过程迭代进行，利用上一轮生成的输出作为检索的上下文，从而获取更相关的知识，进而生成更好的输出。

4.4.3 检索增强生成在问答领域中的应用

如下表 9 所示，总结了 RAG 技术在问答领域中的论文，使得 RAG 技术在问答任务中展现出了强大的性能与灵活性。

(1) 先检索后阅读(retriever-then-read):从庞大的知识库中高效地检索出与问题相关的信息或文档，将这些信息用作大语言模型的上下文输入，以增强模型对特定问题的理解和回答能力。RALLM^[94]结合了传统信息检索技术和大语言模型，这种架构不仅保留了 LLM 强大的语言理解和生成能力，还通过外部信息源增强准确性。Hei Z 等人^[95]提出了一种创新的 RAG 框架，旨在解决多跳问答等复杂任务中的不足，如低效和不完备问题。该方法通过挖掘文档间的动态相关性来提高检索召回率和答案准确率，同时优化整体效率。

He X^[96]提出了框架 G-Retriever，旨在让用户能够像聊天一样与他们的文本图进行交互，用户可以提出关于图的问题，将提供文本回复并突出显示图中的相关部分。Yang D^[97]提出了 IM-RAG，利用内部独白学习进行多轮检索增强生成的新方法，旨在解决大型语言模型在生成幻觉和静态知识库方面的问题。它将信息检索系统与 LLM 集成，并通过学习内部独白(IM)来支持多轮 RAG。Roy K 等人^[98]提出了一种名为 QA-RAG 的新方法，通过将查询和检索到的上下文转换为问答形式，帮助 LLM 更好地理解 and 处理信息。Wu R 等人^[99]提出了一种名为 MSRAG 的多源检索问答框架，旨在解决传统 RAG 范式容易受到错误检索信息影响的问题。Ye L 等人^[100]提出了 ConvRAG，通过整合 LLM 和 RAG 技术，有效地解决了会话问答中的问题表示和知识获取挑战，展现出优异的性能。Ma X 等人^[101]提出了 Rewrite-Retrieve-Read 框架，在检索器之前添加了一个查询重写步骤，利用 LLM 或可训练的小型语言模型（重写器）生成更准确的查询，从而缩小输入文本和所需知识之间的差距。Zou W 等人^[102]提出了 Adaptive-RAG 是一种有效的自适应 RAG 策略，能够根据查询复杂度动态选择最合适的策略，平衡准确性和效率，提高问答系统的性能。Pan F 等人^[103]提出了 T-RAG 的新型端到端表格问答模型，将非参数化密集向量索引与 BART 序列到序列模型联合训练。Mao Y 等人^[104]提出 FIT-RAG，通过利用事实信息、LLM 偏好和自我知识识别，有效地提高了 RAG 的效率和效果，使其在知识密集型任务中具有广泛的应用潜力。Su W 等人^[105]提出了一种有效且高效的动态框架

DRAGIN, 能够显著提升大型语言模型的知识密集型生成能力, 为构建更可靠的人工智能应用提供了新的思路。

(2) 先生成后阅读(generate-then-read): 基于检索到的内容, 运用自然语言处理技术生成准确、流畅的回答。Chen W 等人^[106]提出了多模态检索增强生成器 MuRAG, 旨在解决开放式问答中语言模型知识不足的问题。它通过访问外部非参数多模态记忆库, 将图像和文本中的知识融入到语言生成中, 从而实现更准确的问答。Fatehkia M 等人^[107]提出了 T-RAG, 通过引入树形结构来增强上下文信息, 从而提高了企业内部文档问答的准确性和可靠性。Levonian Z 等人^[108]提出了一种检索增强生成系统, 通过将高质量开源数学教科书的内容整合到 LLM 提示中, 以提高回答质量。Alawwad H A 等人^[109]提出将 LLM 和 RAG 技术结合起来可以显著提升 TQA 任务的准确率, 并有效解决跨域问题。Siriwardhana S 等人^[46]提出了 RAG-end2end, 在训练过程中更新检索器组件 (DPR) 和外部知识库的编码, 从而提高模型在特定领域的适应性。

(3) 检索生成协同(retrieval-generation Synergy): 将

检索技术和生成模型的优势相结合, 以提高问答系统的性能。检索模块可以从大规模知识库中找到与问题相关的精确信息, 为生成模型提供丰富的上下文支持; 生成模块则利用这些信息进行自然语言生成, 从而输出流畅且内容丰富的答案。这种协同方式既保证了回答的准确性, 又增强了系统的灵活性, 特别是在处理复杂或开放性问题时表现出色。因此, 检索-生成协同为提升问答系统的鲁棒性和智能性提供了一种有效的解决方案。Feng Z 等人^[92]提出了 ITRG 框架, 通过迭代的方式结合检索和生成, 利用参数知识和非参数知识, 并通过检索-生成交互帮助找到正确的推理路径。Shao Z 等人^[93]提出了 ITER-RETGEN (迭代检索-生成协同) 的方法, 旨在增强检索增强型大型语言模型的性能, 通过迭代地结合检索和生成来提高模型在处理复杂信息需求时的表现。Mao Y 等人^[110]提出了 (Generation-Augmented Retrieval, GAR) 的方法, 用于回答开放域问题, 通过文本生成来扩充查询, 添加通过启发式发现的相关上下文, 而不需要外部资源作为监督。

表9 RAG在问答领域中的应用

Table 9 Application of RAG in question answering field

方法/框架	解决问题	改进	模态	模型效果	
				数据集	指标
retriever-then-read	DR-RAG ^[95]	检索相关文档效率低	通过一个两阶段检索框架, 挖掘查询与文档之间的相关性	文本	MuSiQue HotpotQA 2Wiki EM:26.97% F1:38.9% Acc:34.03% EM:48.54% F1:62.87% Acc:55.68% EM:49.60% F1:55.62% Acc:55.18%
	G-Retriever ^[96]	实现高效图问答	通过软提示进行微调以增强图的理解	图像	在多个领域的文本图任务中超越基线
	IM-RAG ^[97]	多轮检索解决复杂的问答问题	将 LLM 和信息检索模块连接起来, 实现上下文感知的多轮 RAG	文本	HotPotQA EM:68.4% F1:82.5%
	QA-RAG ^[98]	检索的信息包含许多错误信息	在检索增强过程中利用基于问题和答案的检索块来提高响应质量	文本	明显改善了跨各种数据集的 RAG 管道结果
	MSRAG ^[99]	检索到错误的信息	GPT 检索与 WEB 检索结合的多源检索方法	文本	2WikiMultiHopQA HotpotQA StrategyQA EM:50.8% F1:56.4% EM:30.3% F1:30.6% Acc:86.3%
	ConvRAG ^[100]	RAG 无法适应复杂的会话环境	融合了细粒度检索增强方法	文本	中文 CQA 数据集 seen unseen BLUE-1:32.96% ROUGE-1:40.53% BLUE-1:33.90% ROUGE-1:47.69%
	Rewrite-Retrieve-Read ^[101]	输入的文本与所需的知识之间存在差距	提示 LLM 生成查询, 使用 web 搜索引擎检索上下文	文本	HotPotQA AmbigNQ PopQA EM:30.47% F1:41.34% EM:45.8% F1:58.5% EM:43.2% F1:47.53%
	Adaptive-RAG ^[102]	无法确定查询的复杂性	通过分类器来预测传入查询的复杂程度	文本	TriviaQA HotpotQA EM:52.2% F1:60.7% Acc:58.2% EM:42% F1:53.82% Acc:44.4%
	T-RAG ^[103]	传统信息技术准确率低	利用统一的管道自动搜索表语料库, 从表单元中定位正确答案	表格	NQ-TABLES E2E_WTQ EM:43.06% F1:50.92% MRR:59.23% Hit@1:50.65%
	FIT-RAG ^[104]	忽略事实信息	通过构建双标签文档计分器来利用事实信息	文本	TriviaQA NQ PopQA Acc:75.2% Acc:54.0% Acc:54.4%
	DRAGIN ^[105]			HotpotQA	EM:30.4% F1:39.31%

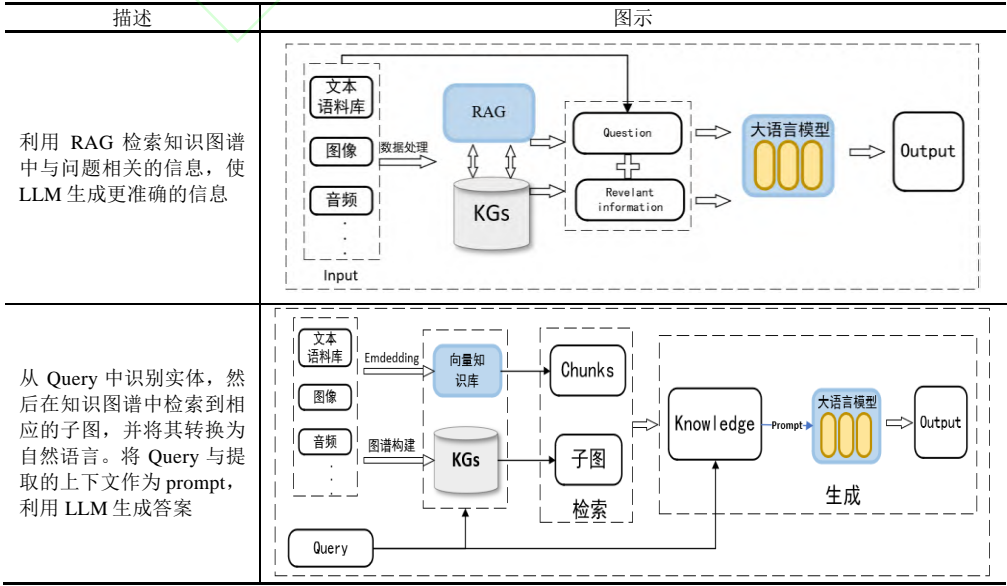
generate-then-read		检索的策略过度依赖于静态规则	基于 LLM 信息需求的动态检索增强生成	文本	2WikiMultihop QA	EM:31.4% F1:42.38%
	MuRAG ^[106]	忽略了其他模态的知识	访问外部非参数多模态存储器来增强语言生成	多模态	WebQA MultimodalQA	Acc:54.6% EM(All):51.4%
	T-RAG ^[107]	RAG 面对不完整的上下文会导致幻觉	使用实体树来进一步增强向量数据库检索到的上下文	文本	与实体相关的问题	正确答案数目:100% 正确答案数目:68.2%
	Levonian Z ^[108]	针对数学领域的回答误差	增加了检索和使用高质量数学内容的提示	文本	—	—
	Alawwad H A ^[109]	模型推理能力弱	利用迁移学习处理上下文提高推理能力	文本	CK12-QA	Acc:82.40%
retrieval-generation Synergy	RAG-end2end ^[46]	没有针对特定领域的知识进行优化	通过神经检索器对模型进行微调、引入了更多领域内的知识	文本	COVID-19 QA News QA ConversationQA	EM:8.32% F1:19.57% EM:14.08% F1:23.7% EM:25.95% F1:37.96%
	ITRG ^[92]	难以处理需要大量知识的知识密集任务	大模型和原始问题生成的伪文档来扩展查询。理解检索到的文档，以生成用于回答问题的新文档。	文本	Natural Questions TriviaQA HotpotQA	EM:37.6% EM:77.0% EM:31.0%
	ITER-RETGEN ^[93]	检索器难以捕捉相关性	将检索和生成过程迭代进行，利用上一轮生成的输出作为检索的上下文，	文本	HotPotQA 2WikiMultiHop QA	EM:45.1% F1:60.4% Acc:72.9% EM:35.5% F1:47.4% Acc:58.4%
	GAR ^[110]	通过检索提高事实性问题的解答	通过文本生成来增强查询，而不需要外部监督或耗时的下游反馈。	文本	Natural Questions TriviaQA	Acc:70.7% Acc:76.0%

4.4.4 大语言模型与检索增强生成、知识图谱协同

LLM+RAG-KG 是一种将大型语言模型、检索增强生成和知识图谱集成在一起的技术。如下表 10 所示，这种集成方法通过利用 KG 的结构化知识来增强 RAG 的能力，从而提高模型的知识表示和推理能力。KG

提供了关于实体和关系的丰富信息，可以帮助 LLM 理解文本中的语义，并生成更准确的答案。Ravi R 等人^[111]提出 PRAGyan1，以检索增强生成为关键组件的混合 KG-RAG-LLM，将检索机制与生成模型相结合，通过提供相关的上下文信息来增强生成过程。Gutiérrez B J 等人^[112]提出了 HippoRAG，允许大型语言模型构建并利用类似的关联图来解决知识整合任务。

表 10 LLM+KG-RAG 图示
Table 10 LLM + KG-RAG Diagram



4.5 基于 Agent 的问答系统

4.5.1 Agent 介绍

代理 (Agent) 是计算机科学和人工智能领域中的一个关键概念, 代表能够感知其环境、做出决策并采取行动的计算实体^[113]。代理技术在各种应用中发挥着重要作用, 包括自动化系统、机器人、智能助手、推荐系统和多代理系统等。大型语言模型的兴起为任务型对话 Agent 的设计和开发带来了新机遇。LLM 强

大的语言理解和生成能力, 能够有效提高对话系统的准确性和用户体验。得益于这些特点, 本文有机会进一步简化任务型对话 Agent 的开发流程, 并显著提高开发效率。由几个部署各种重要功能的关键组件提供支持, 包括规划、计划和工具使用。大型语言模型的崛起为任务导向型对话代理的设计与开发带来了新的契机。凭借 LLM 卓越的语言理解与生成能力, 对话系统的精确度和用户交互体验得以显著提升。如下表 11 所示。

表 11 LLM Agent 组件

Table 11 LLM Agent Components

组件	Agent 制定	工具	记忆			规划	
			长期记忆	短期记忆	感知记忆	子目标和解	反思和完善
内涵	确定 Agent 角色	工具使 LLM 能够通过外部环境来获取信息或完成子任务。	存储可长期保留的信息, 通常存储在外部数据库中。	存储当前意识到的信息, 用于执行复杂的认知任务。	存储通过感官接收到的信息的印象。	大型任务分解为更小的、更可控的子任务, 从而能够有效完成复杂的任务。	通过 LLM 对完成的子任务进行反思, 从错误中吸取教训, 并完善未来的步骤。
能力	提示工程/大模型	大模型短板、需借助外部力量		提示工程		通过 LLM 提示工程, 为智能体赋予规划思维模式	
路径	解析和执行提示模板的指令	Function Calling、函数描述	外部向量存储和快速检索	上下文窗口限制的实现	Prompt	思维链、思维树	ReAct、环境反馈

4.5.2 Agent 在问答领域中的应用

大语言模型 (LLM) 智能体是一种高度先进的机器人或软件, 它具备了自我管理、自我学习、自我适应和自我决策的能力。与传统的自动化程序相比, 大语言模型智能体能够在没有人工直接干预的情况下独立工作, 从而展现了更高的灵活性和适应性。如果将大语言模型比作一个强大的底层操作系统, 那么智能体 (Agent) 就如同这个系统上的高级应用程序。LLM 提供了智能体所需的基础语言处理能力和知识库, 而智能体则利用这些能力来执行各种复杂的任务, 包括理解自然语言、生成自然语言文本、进行逻辑推理等。通过这种方式, 大语言模型智能体不仅能够在没有人工参与的情况下自主工作, 还能够不断地自我提升和改进, 以适应不断变化的环境和任务需求。如表 12 所示, 本文通过回顾和分析多篇关于 Agent 在问答领域应用的论文, 旨在全面综述该领域的发展现状。

(1) 单智能体 (Single-Agent) 指的是只有—个智能体在环境中感知、决策并采取行动的系统。该智能体独立执行任务, 没有其他智能体的协作或竞争。Zong C^[3]等人提出了一个统一的框架 Triad, 是一个基于多角色 LLM 代理的框架, 旨在通过多个角色来解决 KBQA 任务, 分为三个角色:G-Agent 用于学习子任务, D-Agent 用于选择候选实体、关系和 SPARQL 查

询, 顾问代理 A-Agent 用于最终答案生成。Niu C 等人^[114]提出了一个名为 LUAS(LLM-backed User-Agents Simulation) 的算法, 旨在利用大语言模型增强对话状态跟踪(DST)模型, 利用 LLM 生成大量带有 DST 标签的模拟对话数据, 从而减少对话数据的收集和标注成本, 并提高 DST 模型的性能。Xu Y 等人^[115]提出了一种名为 Generate-on-Graph(GoG)的新方法, 用于解决不完全知识图问答问题, 将 LLM 视为代理和知识图谱, 有效整合内部和外部的知识, 并通过选择、生成和回答框架动态扩展子图和生成新知识, 从而更好地回答 IKGQA 问题。Liu N^[116]提出了 RAISE 架构, 通过增强记忆系统和微调大型语言模型, 提升了对话代理在复杂对话中的上下文感知和适应性。

(2) 多智能体 (Multi-Agent) 指的是由多个智能体组成的系统, 这些智能体在同一个环境中共同存在, 并且它们之间可能存在协作、竞争或混合的关系。Zhang S^[117]提出了一种基于 LLM 辅助的多智能体对话本体对齐模型, 通过智能体之间的协商过程减少了对领域专家的依赖, 提高了本体对齐的效率和可解释性。Patel B^[118]首次提出基于多 LLM 智能体的 EQA 框架, 并引入 CAM 进行答案聚合。Wang K^[119]提出了一个通用框架, 利用语言反馈和非语言奖励信号来训练大模型。

表 12 Agent 在问答领域中的应用

Table 12 Agent in the Application of Question-Answering Systems

Single-	方法/框架	解决问题	效果
---------	-------	------	----

Agent			数据集	效果
Triad ^[3]	基于 LLM 的 Agent 如何通过充当多个角色来解决 KBQA 任务		YAGO-QA	Pre:56.1%Rec:56.8% F1:56.4%
			LCQuAD 1.0	Pre:40.8% Rec:42.5% F1:41.6%
			QALD-9	Pre:69.0% Rec:66.4% F1:67.7%
	用户代理仿真 (LUAS) 算法 ^[114]	对话状态跟踪获取数据成本高昂	MultiWOZ 2.2	JGA:66.25%
			MultiWOZ 2.4	JGA:78.20%
Multi-agent	GoG ^[115]	知识图不完整, 无法涵盖回答问题所需的全部知识	WebQuestionSP	Hits@1:75.2% Hits@1:61.0%
			Complex WebQuestion	Hits@1:84.4% Hits@1:74.8%
	RAISE ^[116]	增强智能体在复杂、多回合对话中的可控性和适应性	----	-----
			----	-----
	LLMA 对话模型 ^[117]	LLM 的可靠性和可解释性不全是可预测的	OAEI 解剖学数据集	Pre:54.4% Rec:59.9% F-measure:66.6%
	CAM ^[118]	EQA 在单智能体上检索既耗时又昂贵	----	-----
	LTC ^[119]	智能体无法同时利用语言反馈和非语言奖励信号进行训练	ALFWorld	success rates:91%
			HotpotQA	EM:35.8%
			Chameleon Wu	winning rates:25%
			GSM8k	Acc:41.3%

5 智能问答系统未来发展方向及讨论

5.1 方法对比分析

如下表 13 所示, 对本文所提出的提示学习、知识图谱、检索增强生成 (RAG)、Agent、微调这几种方法类型进行了多方面的对比分析。在技术原理上, 各方法各有特点, 如提示学习靠设计自然语言提示引导预训练模型推理, 知识图谱以图结构呈现知识等。优点方面, 提示学习具备高效性与灵活性, 知识图谱可解释性强且知识结构化, RAG 能补充信息并提升生成

质量, Agent 可处理复杂动态任务, 微调能快速适应新任务。然而, 它们也分别存在一些问题, 像提示学习依赖人工设计且优化困难, 知识图谱构建更新成本高、动态性差, RAG 依赖检索模块且可能引入冗余信息, Agent 需大量计算资源, 微调则要大量任务特定数据。在适用场景上, 各方法适用于不同情况, 如任务明确且预训练模型基础好的场景适合提示学习, 知识密集型场景适用知识图谱等。针对存在的问题, 各方法也有相应的改进思路, 如提示学习可结合自动化提示生成等方法优化, 知识图谱可建立自动化构建和更新机制等。

表 13 方法对比

Table13 Method Comparison

方法类型	技术原理	优点	存在问题	适用场景	改进思路
提示学习	通过设计合适的自然语言提示引导预训练模型进行任务推理。	高效性	依赖人工设计, 效果	任务明确, 预训练	结合自动化提示生成方法
		灵活性	不稳定, 优化困难	模型已有较好基础	或强化学习优化提示
知识图谱	通过图结构表示实体及其关系	可解释性强 结构化知识	构建与更新成本高, 动态性差	知识密集型场景	自动化构建和更新机制
检索增强生成 (RAG)	结合信息检索	信息补充	依赖于检索模块	动态知识	增强信息
	和生成模型	生成质量高	引入冗余信息	获取场景	检索模型
Agent	与环境互动 并基于反馈进行学习	能够处理复杂动态任务	需要大量计算资源	复杂任务执行场景	通过强化学习提升稳定性
微调	使用任务特定数据进一步调整模型参数	快速适应 新任务	需要大量 任务特定数据	数据标注丰富且任务明确的场景	少量样本微调

5.2 挑战及未来研究方向

基于大语言模型的智能问答系统在近年来取得了显著的进展, 但仍然面临一些局限性, 如图 6 所示, 比如提示模板难以解决复杂的问题、依赖于检索模块等问题。随着大模型技术的不断进步, 未来的发展方

向将会更加多样化和深入, 本节概述了基于大模型的问答系统所面临的主要挑战以及未来的发展方向。

(1) 优化提示模板设计

提示学习通过设计合适的提示词和模板引导大模型回答用户提出的问题, 针对复杂问答场景, 可以通

过多种策略进行优化,例如使用提示组合或链式提示,这种方式能够分步骤或分层次地引导大模型进行推理。此外,动态提示生成技术也能根据问题的具体内容实时生成提示,更具灵活性和适应性,从而克服传统提示方法在处理复杂问答任务中的局限性。通过这些策略的结合,不仅可以显著提高问答系统的响应准确性,还能够多模态场景下增强模型的表现力和处理能力,为用户提供更加智能化和精准的问答体验。

(2) 优化检索算法

RAG 技术通过在生成过程中结合检索到的外部知识,更好地理解用户的意图和需求。针对优化检索算法,向量检索优化可以提升模型在大规模数据集中的检索效率,个性化检索优化针对不同用户的需求进行定制化的检索,通过结合向量检索优化和个性化检索优化,RAG 技术能够更加高效、精准地为用户提供答案。

(3) 动态知识更新

知识图谱中的信息构建完成后,难以快速更新,可以通过自动化知识更新和多步推理解决知识更新问题,利用信息抽取等技术自动地从数据源中抽取实体、关系等信息,将其添加到知识图谱中。通过不断迭代推理,根据已知的事实和关系中推断出新的信息,扩展知识图谱。

(4) 增强交互灵活性

将复杂的任务分解成多个简单的子任务,每个子任务可以更精细化地进行处理,从而提高整体问答的准确性和效率。通过使用不同的 Agent 系统来处理这些子任务,可以根据具体任务的性质选择最适合的 Agent 进行处理。同时,借助交互数据的实时收集和动态反馈调整 Agent 的行为和策略,可以提升系统的灵活性和适应性。

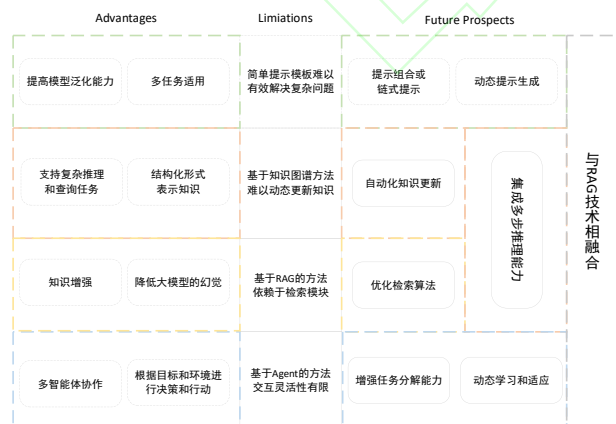


图6 挑战和未来研究方向

Fig.6 Challenges and Future Research Directions

6 总结

近年来,基于大语言模型的智能问答系统得到了广泛关注和快速发展,尤其是通过结合提示学习、知识图谱、检索增强生成(RAG)和智能代理(Agent)等先进技术,极大提升了问答系统的性能和应用范围。提示学习通过优化模型对特定任务的响应,提高了系

统在处理复杂问题时的准确性和灵活性。知识图谱的引入使系统能够在结构化信息的支持下,更好地理解问题背景,实现更精准的知识检索和答案生成。检索增强生成作为一种新兴技术,结合了信息检索与生成模型的优势,能够在动态知识环境中提供及时、相关的答案,从而提升系统的实用性。此外,智能代理技术赋予了问答系统自主决策的能力,使其在复杂交互场景中能够更好地满足用户需求。这一切表明,基于大语言模型的智能问答系统不仅在理论研究上取得了突破,还在实际应用中展现出巨大的潜力和价值。通过本文对相关技术的系统综述与分析,希望能推动问答系统的进一步发展,解决当前面临的挑战,并探索新的研究方向。未来的研究将继续关注问答系统在实际应用中的表现,探索其在各个领域的潜在应用价值,并不断优化和改进现有技术,为用户提供更高效、准确和个性化的问答体验。

参考文献:

- [1] 姚元杰,龚毅光,刘佳,徐闯,朱栋梁.基于深度学习的智能问答系统综述.计算机系统应用,2023,32(4):1-15
YAO Y J ,GONG Y G ,LIU J,et al. Survey on Intelligent Question Answering System Based on Deep Learning. Computer Systems & Applications, 2022, 32(4): 1-15.
- [2] WU X, PENG Z, RAJANALA S, et al. Passage-specific Prompt Tuning for Passage Reranking in Question Answering with Large Language Models[J]. arXiv preprint arXiv:2405.20654, 2024.
- [3] ZONG C, YAN Y, LU W, et al. Triad: A Framework Leveraging a Multi-Role LLM-based Agent to Solve Knowledge Base Question Answering[J]. arxiv preprint arxiv:2402.14320, 2024.
- [4] FRISONI G, COCCHIERI A, PRESEPI A, et al. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering[J]. arXiv preprint arXiv:2403.01924, 2024.
- [5] PAN X, SUN K, YU D, et al. Improving question answering with external knowledge[J]. arxiv preprint arxiv:1902.00993, 2019.
- [6] MENSIO M, RIZZO G, MORISIO M. Multi-turn qa: A rnn contextual approach to intent classification for goal-oriented systems[C]//Companion Proceedings of the The Web Conference 2018. 2018: 1075-1080.
- [7] 王婷,王娜,崔运鹏,等.基于人工智能大模型技术的果蔬农技知识智能问答系统[J].智慧农业(中英文), 2023, 5(4): 105-116.
WANG T, WANG N, CUI Y P, et al. Agricultural Technology Knowledge Intelligent Question-Answering System Based on Large Language Model[J]. Smart Agriculture, 2023, 5(4): 105-116.
- [8] WOODS W A. Lunar rocks in natural English: Explorations in natural language question answering[J]. 1977.
- [9] AO H, TAKAGI T. ALICE: an algorithm to extract abbreviations from MEDLINE[J]. Journal of the American Medical Informatics Association, 2005, 12(5): 576-586.
- [10] SHARMA Y, GUPTA S. Deep learning approaches for question answering system[J]. Procedia computer science, 2018, 132: 785-794.

- [11] LIU Y, HAN T, MA S, et al. Summary of chatgpt-related research and perspective towards the future of large language models[J]. Meta-Radiology, 2023: 100017.
- [12] HOFFMANN J, BORGEAUD S, MENSCH A, et al. Training Compute-Optimal Large Language Models[J]. 2022. DOI:10.48550/arXiv.2203.15556.
- [13] YANG H, LIU X Y, DAN WANG C. FinGPT: Open-Source Financial Large Language Models[J]. FinLLM at IJCAI, 2023.
- [14] CUI J, LI Z, YAN Y, et al. Chatlaw: Open-source legal large language model with integrated external knowledge bases[J]. arxiv preprint arxiv:2306.16092, 2023.
- [15] BORDES A, USUNIER N, CHOPRA S, et al. Large-scale simple question answering with memory networks[J]. arxiv preprint arxiv:1506.02075, 2015.
- [16] MILLER A, FISCH A, DODGE J, et al. Key-value memory networks for directly reading documents[J]. arxiv preprint arxiv:1606.03126, 2016.
- [17] BERANT J, CHOU A, FROSTIG R, et al. Semantic parsing on freebase from question-answer pairs[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 1533-1544.
- [18] KWIATKOWSKI T, PALOMAKI J, REDFIELD O, et al. Natural questions: a benchmark for question answering research[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 453-466.
- [19] GU Y, KASE S, VANNI M, et al. Beyond IID: three levels of generalization for question answering on knowledge bases[C]//Proceedings of the Web Conference 2021. 2021: 3477-3488.
- [20] YANG Z, QI P, ZHANG S, et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering[J]. arxiv preprint arxiv:1809.09600, 2018.
- [21] ZHANG Y, DAI H, KOZAREVA Z, et al. Variational reasoning for question answering with knowledge graph[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
- [22] BAO J, DUAN N, YAN Z, et al. Constraint-based question answering with knowledge graph[C]//Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. 2016: 2503-2514.
- [23] YIH W, RICHARDSON M, MEEK C, et al. The value of semantic parse labeling for knowledge base question answering[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016: 201-206.
- [24] TALMOR A, BERANT J. The web as a knowledge-base for answering complex questions[J]. arxiv preprint arxiv:1803.06643, 2018.
- [25] UNGER C, FORASCU C, LOPEZ V, et al. Question answering over linked data (QALD-4)[C]//Working notes for CLEF 2014 conference. 2014.
- [26] CAO S, SHI J, PAN L, et al. KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base[J]. arxiv preprint arxiv:2007.03875, 2020.
- [27] REDDY S, CHEN D, MANNING C D. Coqa: A conversational question answering challenge[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 249-266.
- [28] CHOI E, HE H, IYYER M, et al. QuAC: Question answering in context[J]. arxiv preprint arxiv:1808.07036, 2018.
- [29] RAJPURKAR P, ZHANG J, LOPYREV K, et al. Squad: 100,000+ questions for machine comprehension of text[J]. arxiv preprint arxiv:1606.05250, 2016.
- [30] YANG Y, YIH W, MEEK C. Wikiqa: A challenge dataset for open-domain question answering[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. 2015: 2013-2018.
- [31] 唐嘉, 庞太威, 刘书铭, 等. 大语言模型微调训练与检索增强生成技术在油气企业制度问答应用中的效果对比研究[J]. 数字通信世界, 2024, (11): 104-106.
- TANG J, PANG D W, LIU S M, et al. A Comparative Study on the Effects of Fine-tuning Large Language Models and Retrieval Enhanced Generation Technology in Oil and Gas Enterprise Regulation Question Answering Application[J]. Digital Communication World, 2024, (11): 104-106.
- [32] 陈俊臻, 王淑莹, 罗浩然. 融合大模型微调与图神经网络的知识图谱问答[J]. 计算机工程与应用, 2024, 60(24): 166-176.
- CHEN J Z, WANG S Y, LUO H R. Combining Large Model Fine-Tuning and Graph Neural Networks for Knowledge Graph Question Answering[J]. Computer Engineering and Applications, 2024, 60(24): 166-176.
- [33] 文森, 钱力, 胡懋地, 等. 基于大语言模型的问答技术研究进展综述 [J]. 数据分析与知识发现, 2024, 8(6): 16-29.
- Wen S, Qian L, Hu M D, et al. Review of Research Progress on Question-Answering Techniques Based on Large Language Models[J]. Data Analysis and Knowledge Discovery, 2024, 8(6): 16-29.
- [34] 张钦彤, 王昱超, 王鹤羲, 等. 大语言模型微调技术的研究综述[J]. 计算机工程与应用, 2024, 60(17): 17-33.
- ZHANG Q T, WANG Y C, WANG H X, et al. Comprehensive Review of Large Language Model Fine-Tuning[J]. Computer Engineering and Applications, 2024, 60(17): 17-33.
- [35] MALLADI S, GAO T, NICHANI E, et al. Fine-tuning language models with just forward passes[J]. Advances in Neural Information Processing Systems, 2023, 36: 53038-53075.
- [36] LV K, YANG Y, LIU T, et al. Full parameter fine-tuning for large language models with limited resources[J]. arXiv preprint arXiv:2306.09782, 2023.
- [37] TANWISUTH K, ZHANG S, ZHENG H, et al. POUF: Prompt-oriented unsupervised fine-tuning for large pre-trained models[C]//International Conference on Machine Learning. PMLR, 2023: 33816-33832.
- [38] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]//International conference on machine learning. PMLR, 2019: 2790-2799.
- [39] LESTER B, AL-ROUF R, CONSTANT N. The power of scale for parameter-efficient prompt tuning[J]. arXiv preprint arXiv:2104.08691, 2021.
- [40] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. Qlora: Efficient finetuning of quantized llms[J]. Advances in Neural Information Processing Systems, 2024, 36.
- [41] ZHANG Q, CHEN M, BUKHARIN A, et al. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning[J]. arxiv preprint arxiv:2303.10512, 2023.
- [42] LIN X, WANG W, LI Y, et al. Data-efficient Fine-tuning for LLM-based Recommendation[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and

- Development in Information Retrieval. 2024: 365-374.
- [43] HAN G, LIM S N. Few-Shot Object Detection with Foundation Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 28608-28618.
- [44] LI Z, FAN S, GU Y, et al. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(17): 18608-18616.
- [45] MAO K, LIU Z, QIAN H, et al. RAG-Studio: Towards In-Domain Adaptation of Retrieval Augmented Generation Through Self-Alignment[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 725-735.
- [46] SIRIWARDHANA S, WEERASEKERA R, WEN E, et al. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1-17.
- [47] MO T, XIAO Q, ZHANG H, et al. Domain-Specific Few-Shot Table Prompt Question Answering via Contrastive Exemplar Selection[J]. Algorithms, 2024, 17(7): 278.
- [48] WANG L, YU K, WUMAIER A, et al. Genre: generative multi-turn question answering with contrastive learning for entity-relation extraction[J]. Complex & Intelligent Systems, 2024: 1-15.
- [49] JUNG M, PARK S, SUL J, et al. Is Prompt Transfer Always Effective? An Empirical Study of Prompt Transfer for Question Answering[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers). 2024: 528-539.
- [50] LIU Y, WEI W, PENG D, et al. Declaration-based prompt tuning for visual question answering[J]. arXiv preprint arXiv:2205.02456, 2022.
- [51] LIU H, QIN Y. Heterogeneous graph prompt for community question answering[J]. Concurrency and Computation: Practice and Experience, 2022: e7156.
- [52] BAEK J, AJI A F, SAFFARI A. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering[J]. arXiv preprint arXiv:2306.04136, 2023.
- [53] OSSOWSKI T, HU J. Multimodal prompt retrieval for generative visual question answering[J]. arXiv preprint arXiv:2306.17675, 2023.
- [54] CUI C, LI Z. Prompt-Enhanced Generation for Multimodal Open Question Answering[J]. Electronics, 2024, 13(8): 1434.
- [55] 王喆,杨栋梁,况星园,等.考虑提示学习的洪涝灾害应急决策自动问答模型研究*[J].中国安全生产科学技术,2022,18(11):12-18.
WANG Z,YANG D L,KUANG X Y,et al.Research on automatic question answering model of flood disaster emergency decision-making considering Prompt-learning[J]. Journal of safety science and technology, 2022,18(11):12-18.
- [56] ZHANG Y, FAN W, PENG P, et al. Dual modality prompt learning for visual question-grounded answering in robotic surgery[J]. Visual Computing for Industry, Biomedicine, and Art, 2024, 7(1): 9.
- [57] ZHONG W, GAO Y, DING N, et al. ProQA: Structural prompt-based pre-training for unified question answering[J]. arXiv preprint arXiv:2205.04040, 2022.
- [58] MA Z, YU Z, LI J, et al. HybridPrompt: bridging language models and human priors in prompt tuning for visual question answering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(11): 13371-13379.
- [59] LAZARIDOU A, GRIBOVSKAYA E, STOKOWIEC W, et al. Internet-augmented language models through few-shot prompting for open-domain question answering[J]. arXiv preprint arXiv:2203.05115, 2022.
- [60] LU P, MISHRA S, XIA T, et al. Learn to explain: Multimodal reasoning via thought chains for science question answering[J]. Advances in Neural Information Processing Systems, 2022, 35: 2507-2521.
- [61] MITRA C, MIROYAN M, JAIN R, et al. RetLLM-E: Retrieval-Prompt Strategy for Question-Answering on Student Discussion Forums[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(21): 23215-23223.
- [62] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arxiv preprint arxiv:1907.11692, 2019.
- [63] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of machine learning research, 2020, 21(140): 1-67.
- [64] PETRONI F, ROCKTÄSCHEL T, LEWIS P, et al. Language models as knowledge bases?[J]. arxiv preprint arxiv:1909.01066, 2019.
- [65] ZHONG L, WU J, LI Q, et al. A comprehensive survey on automatic knowledge graph construction[J]. ACM Computing Surveys, 2023, 56(4): 1-62.
- [66] 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, 9: 176-194.
- [67] SUN J, XU C, TANG L, et al. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph, 2024[J]. URL <https://arxiv.org/abs/2307.2023>, 2023, 7697.
- [68] TAN X, WANG H, QIU X, et al. Struct-X: Enhancing Large Language Models Reasoning with Structured Data[J]. arXiv preprint arXiv:2407.12522, 2024.
- [69] WANG Y, JIANG B, LUO Y, et al. Reasoning on Efficient Knowledge Paths: Knowledge Graph Guides Large Language Model for Domain Question Answering[J]. arXiv preprint arXiv:2404.10384, 2024.
- [70] GUO T, YANG Q, WANG C, et al. Knowledgenavigator: Leveraging large language models for enhanced reasoning over knowledge graph[J]. Complex & Intelligent Systems, 2024, 10(5): 7063-7076.
- [71] JIANG J, ZHOU K, DONG Z, et al. Structgpt: A general framework for large language model to reason over structured data[J]. arXiv preprint arXiv:2305.09645, 2023.
- [72] AGRAWAL G, PAL K, DENG Y, et al. CyberQ: Generating Questions and Answers for Cybersecurity Education Using Knowledge Graph-Augmented LLMs[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(21): 23164-23172.
- [73] ZHANG Q, DONG J, CHEN H, et al. Knowgpt: Black-box knowledge injection for large language models[J]. arXiv preprint arXiv:2312.06185, 2023.
- [74] WANG J, CHEN M, HU B, et al. Learning to Plan for Retrieval-Augmented Large Language Models from

- Knowledge Graphs[J]. arXiv preprint arXiv:2406.14282, 2024.
- [75] WANG F, BAO R, WANG S, et al. InfuserKI: Enhancing Large Language Models with Knowledge Graphs via Infuser-Guided Knowledge Integration[J]. arXiv preprint arXiv:2402.11441, 2024.
- [76] MOISEEV F, DONG Z, ALFONSECA E, et al. SKILL: Structured knowledge infusion for large language models[J]. arXiv preprint arXiv:2205.08184, 2022.
- [77] XIA T, DING L, WAN G, et al. Improving Complex Reasoning over Knowledge Graph with Logic-Aware Curriculum Tuning[J]. arXiv preprint arXiv:2405.01649, 2024.
- [78] ZHANG Y, CHEN Z, FANG Y, et al. Knowledgeable preference alignment for llms in domain-specific question answering[J]. arXiv preprint arXiv:2311.06503, 2023.
- [79] GAO Y, QIAO L, WAN G, et al. Two-stage Generative Question Answering on Temporal Knowledge Graph Using Large Language Models[J]. arXiv preprint arXiv:2402.16568, 2024.
- [80] ZHANG Y, CHEN Z, GUO L, et al. Making large language models perform better in knowledge graph completion[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 233-242.
- [81] AGARWAL D, DAS R, KHOSLA S, et al. Bring your own kg: Self-supervised program synthesis for zero-shot kgqa[J]. arXiv preprint arXiv:2311.07850, 2023.
- [82] LUO H, TANG Z, PENG S, et al. Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models[J]. arXiv preprint arXiv:2310.08975, 2023.
- [83] FENG C, ZHANG X, FEI Z. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs[J]. arXiv preprint arXiv:2309.03118, 2023.
- [84] Wang H, Zhao S, Qiang Z, et al. Knowledge-tuning Large Language Models with Structured Medical Knowledge Bases for Reliable Response Generation in Chinese[J]. arXiv preprint arXiv:2309.04175, 2023.
- [85] ZHOU T, CHEN Y, LIU K, et al. CogMG: Collaborative Augmentation Between Large Language Model and Knowledge Graph[J]. arXiv preprint arXiv:2406.17231, 2024.
- [86] WEN Y, WANG Z, SUN J. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models[J]. arXiv preprint arXiv:2308.09729, 2023.
- [87] GAO Y, XIONG Y, GAO X, et al. Retrieval-augmented generation for large language models: A survey[J]. arXiv preprint arXiv:2312.10997, 2023.
- [88] SHI W, MIN S, YASUNAGA M, et al. Replug: Retrieval-augmented black-box language models[J]. arxiv preprint arxiv:2301.12652, 2023.
- [89] RAM O, LEVINE Y, DALMEDIGOS I, et al. In-context retrieval-augmented language models[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1316-1331.
- [90] YU W, ITER D, WANG S, et al. Generate rather than retrieve: Large language models are strong context generators[J]. arxiv preprint arxiv:2209.10063, 2022.
- [91] SUN Z, WANG X, TAY Y, et al. Recitation-augmented language models[J]. arxiv preprint arxiv:2210.01296, 2022.
- [92] FENG Z, FENG X, ZHAO D, et al. Retrieval-generation synergy augmented large language models[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 11661-11665.
- [93] SHAO Z, GONG Y, SHEN Y, et al. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy[J]. arxiv preprint arxiv:2305.15294, 2023.
- [94] HU Y, LU Y. Rag and rau: A survey on retrieval-augmented language model in natural language processing[J]. arXiv preprint arXiv:2404.19543, 2024.
- [95] HEI Z, WEI W, OU W, et al. DR-RAG: Applying Dynamic Document Relevance to Retrieval-Augmented Generation for Question-Answering[J]. arxiv preprint arxiv:2406.07348, 2024.
- [96] HE X, TIAN Y, SUN Y, et al. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering[J]. arxiv preprint arxiv:2402.07630, 2024.
- [97] YANG D, RAO J, CHEN K, et al. IM-RAG: Multi-Round Retrieval-Augmented Generation Through Learning Inner Monologues[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 730-740.
- [98] ROY K, ZI Y, SHYALIKA C, et al. QA-RAG: Leveraging Question and Answer-based Retrieved Chunk Re-Formatting for Improving Response Quality During Retrieval-augmented Generation[J]. 2024.
- [99] WU R, CHEN S, SU X, et al. A Multi-Source Retrieval Question Answering Framework Based on RAG[J]. arxiv preprint arxiv:2405.19207, 2024.
- [100] YE L, LEI Z, YIN J, et al. Boosting Conversational Question Answering with Fine-Grained Retrieval-Augmentation and Self-Check[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 2301-2305.
- [101] MA X, GONG Y, HE P, et al. Query rewriting for retrieval-augmented large language models[J]. arxiv preprint arxiv:2305.14283, 2023.
- [102] ZOU W, GENG R, WANG B, et al. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models[J]. arXiv preprint arXiv:2402.07867, 2024.
- [103] PAN F, CANIM M, GLASS M, et al. End-to-end table question answering via retrieval-augmented generation[J]. arxiv preprint arxiv:2203.16714, 2022.
- [104] MAO Y, DONG X, XU W, et al. FIT-RAG: Black-Box RAG with Factual Information and Token Reduction[J]. arxiv preprint arxiv:2403.14374, 2024.
- [105] SU W, TANG Y, AI Q, et al. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models[J]. arxiv preprint arxiv:2403.10081, 2024.
- [106] CHEN W, HU H, CHEN X, et al. Murag: Multimodal retrieval-augmented generator for open question answering over images and text[J]. arxiv preprint arxiv:2210.02928, 2022.
- [107] FATEHKIA M, LUCAS J K, CHAWLA S. T-RAG: lessons from the LLM trenches[J]. arxiv preprint arxiv:2402.07483, 2024.
- [108] LEVONIAN Z, LI C, ZHU W, et al. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference[J]. arxiv preprint arxiv:2310.03184, 2023.
- [109] ALAWWAD H A, ALHOTHALI A, NASEEM U, et al. Enhancing textbook question answering task with large language models and retrieval augmented generation[J].

- arxiv preprint arxiv:2402.05128, 2024.
- [110] MAO Y, HE P, LIU X, et al. Generation-augmented retrieval for open-domain question answering[J]. arXiv preprint arXiv:2009.08553, 2020.
- [111] RAVI R, GINDE G, ROKNE J. PRAGyan--Connecting the Dots in Tweets[J]. arxiv preprint arxiv:2407.13909, 2024.
- [112] GUTIÉRREZ B J, SHU Y, GU Y, et al. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models[J]. arxiv preprint arxiv:2405.14831, 2024.
- [113] XI Z, CHEN W, GUO X, et al. The rise and potential of large language model based agents: A survey[J]. arXiv preprint arXiv:2309.07864, 2023.
- [114] NIU C, WANG X, CHENG X, et al. Enhancing Dialogue State Tracking Models through LLM-backed User-Agents Simulation[J]. arxiv preprint arxiv:2405.13037, 2024.
- [115] XU Y, HE S, CHEN J, et al. Generate-on-Graph: Treat LLM as both Agent and KG in Incomplete Knowledge Graph Question Answering[J]. arxiv preprint arxiv:2404.14741, 2024.
- [116] LIU N, CHEN L, TIAN X, et al. From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models[J]. arxiv preprint arxiv:2401.02777, 2024.
- [117] ZHANG S, DONG Y, ZHANG Y, et al. Large Language Model Assisted Multi-Agent Dialogue for Ontology Alignment[C]//Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. 2024: 2594-2596.
- [118] PATEL B, DORBALA V S, BEDI A S. Embodied Question Answering via Multi-LLM Systems[J]. arxiv preprint arxiv:2406.10918, 2024.
- [119] WANG K, LU Y, SANTACROCE M, et al. Adapting LLM Agents with Universal Feedback in Communication[C]//ICML 2024 Workshop on Foundation Models in the Wild.