◎热点与综述◎

基于大型语言模型的检索增强生成综述

刘雪颖[1,2],云 静[1,2+],李 博[1,2],史晓国[1,2],张钰莹[1,2]
1.内蒙古工业大学 数据科学与应用学院,呼和浩特 010080
2.内蒙古自治区大数据软件服务工程技术研究中心,呼和浩特 010080
+ 通信作者 E-mail:yunjing_zoe@163.com

摘 要:最近,智能体代理能在复杂任务中提供高效的解决方案,在工业界备受关注。作为智能体代理的常见范式之一,检索增强生成(retrieval-augmented generation,RAG)旨在结合信息检索和内容生成技术增强生成响应质量,已逐步成为研究的重点。在对国内外检索增强生成方法研究的基础上,阐述了RAG的基本概念及工作流程,归纳了技术现状,分析了现有RAG技术的优缺点,梳理了现有评估指标、数据集和基准。最后探讨了RAG技术在未来应用场景下所面临的挑战,并展望了其未来发展方向。

关键词:大语言模型;检索增强生成;评估基准
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.2410-0088

Survey of Retrieval-Augmented Generation Based on Large Language Models

LIU Xueying[1,2], YUN Jing[1,2+], LI Bo[1,2], SHI Xiaoguo[1,2], ZHANG Yuying[1,2]
1.College of Data Science and Application, Inner Mongolia University of Technology, Hohhot 010080, China
2.Inner Mongolia Autonomous Region Engineering and Technology Research Center of Big Data Software Service, Hohhot 010080, China

Abstract: Artificial intelligence agents provide efficient solutions in complex tasks, which have recently gained attention in industry. As one of the paradigms of artificial intelligence agents, retrieval-augmented generation (RAG), which aims to enhance the quality of generated responses by combining information retrieval and content generation techniques, has gradually become the focus of research. According to the studies on retrieval enhancement generation methods at home and abroad, the basic concept and workflow of RAG are elaborated, the current state of the technology is summarized, the advantages and disadvantages of the existing RAG technology are analyzed, and the existing evaluation indexes, datasets and benchmarks are sorted out. Finally, challenges faced by RAG technology in future application scenarios are discussed and the future development direction of RAG technology is envisioned.

Key words: large language models; retrieval-augmented generation; evaluation benchmarks

近年来,算力的飞跃、数据的爆炸性增长以及深度学习技术的持续进步,使人工智能跃迁到一个全新时代。大型语言模型(large language models,LLMs)由于其强大的表征和泛化能力,在机器翻译、内容生成和情感分析等任务中表现卓越。GPT-4o、GPT-4和LLAMA-3等最新大模型应用程序的相继出现,垂直领域的大模型相继在企业场景中应用,例如,医疗领域的Med-PaLM[1]、法律领域的DISC-LawLLM[2]和教育领域的松鼠AI[3]。需要注意的是,随着模型能力的不断提升,在垂直领域的应用过程中,专业知识不足、数据隐私和安全和模型幻觉等挑战不可避免。例如,在医疗领域,LLMs缺乏医学知识和临床经验,可能无法准确识别复杂疾病的症状

基金项目:国家自然科学基金(62062055);内蒙古高校青年科技英才项目(NJYT24061);内蒙古自治区直属高校基本科研业务费项目(JY20220249)。

作者简介:刘雪颖(2001—),女,硕士研究生,研究方向为检索增强生成;云静(1980—),女,博士,副教授,CCF会员,研究方向为AIGC大模型;史晓国(1999—),男,硕士研究生,研究方向为步态识别;张钰莹(2001—),女,硕士研究生,研究方向为大模型对齐。

收稿日期:2024-10-09 **修回日期:**2025-01-15 **文章编号:**1002-8331(2025)13-0001-25

或药物副作用,导致不符合临床实际的答案,甚至混淆副作用或提供不适合患者的治疗建议,影响医疗决策。同时,金融等敏感领域的数据隐私问题尤为突出,特别是在涉及用户财务信息的微调过程中,LLMs可能无意处理敏感数据,引发泄露或滥用的风险。因此,在学术界和工业界,弥补大模型的上述不足,是研究人员和从业者亟待解决的现实问题。

为解决上述问题,现有工作主要可以分为两类:第一类通过重新训练大模型以适应特定领域的数据,即微调(fine-tuning,FT)预训练模型。例如,Nikdan等人[4]通过微调LLMs以增强其适配能力,从而提高上下文记忆的准确性。第二类是利用检索增强生成技术(retrieval-augmented generation,RAG)来结合外部知识库的信息。例如,MolReGPT[5]通过RAG增强大模型在分子发现中的上下文学习能力,有效在会话任务[6]中更新知识,减少了LLMs产生不正确或无意义的回复。然而,在大模型微调过程中存在以下问题。一是在医疗、金融等敏感领域,数据隐私和安全是至关重要的。微调过程涉及敏感数据处理,可能导致数据泄露等问题。二是在垂直领域内,大模型在内容生成过程会产生幻觉问题。三是在特定领域微调大模型,不仅耗时,而且成本高昂,这在实际应用中需要综合考量。因此,RAG以外挂知识库的方式,能有效避免隐私数据泄露。同时,在内容生成过程中,会检索大规模知识集合,生成的内容更丰富,有效增强大模型的泛化能力。此外,RAG根据实际存在数据做出反应,减少了生成或捏造错误信息的可能性,从而提高生成内容的真实性和可靠性。因此,RAG能够在各种场景下提供更精确且完备的支持。

检索增强生成(RAG)是一种结合信息检索(无监督学习)和内容生成(有监督学习)的混合方法,旨在通过从外部信息源中检索相关信息来增强生成模型的能力,从而提高生成内容的相关性和质量。最近有关RAG的研究主要可以分为两类,第一类是优化检索机制,相关研究致力于改进信息检索的性能,实现从大量信息中快速且高效的检索[7-9]。REALM通过改进相似性度量方法以加快检索速度。第二类是增强生成模型,相关研究试图减少生成模型生成看似合理但实际上错误的信息,即幻觉[10]。例如,Izacard等人[11]通过降低LLMs的规模,在知识密集型任务上实现了出色的性能,Wu等人[12]通过扩展对长上下文的支持,实现了对过往输入内部表示的记忆功能,从而减轻幻觉。除了检索和生成两个主流方向,高级流程优化聚焦于优化检索与生成过程之间的协同作用,即增强生成过程。研究者们提出了多种方案涵盖以检索结果直接作为生成器的增强输入[13-14],以潜在的表示的形式加入到生成过程[15-16],以logits的形式贡献于最终生成结果[17-18],甚至影响或改变某些生成步骤[19-20]。

大量研究优化RAG的检索与生成过程,但在实际应用中,检索与生成的协同仍存在不足,尤其是在处理复杂任务时。现有策略未能有效整合检索和生成,导致生成内容的相关性和准确性仍有很大提升空间。特别是在多领域的应用中,如何高效协调两者的协同效应,依然是RAG技术面临的关键挑战。研究人员提出了多种具体策略挖掘RAG的性能。例如,Zhao等人[21]讨论了AIGC的RAG。Gao等人[22]对LLMs的RAG进行了比较全面的综述,重点关注推理准确性和生成质量。Fan等人[23]侧重于技术角度论述RAG。尽管上述研究为RAG优化提供了有益方向,但在如何系统优化检索与生成的协同工作,尤其是在复杂任务中的应用,仍缺乏深入探讨。此外,Chen等人[24]详细回顾了RAG的能力,但在解决实际应用中的挑战上仍存在许多空白。本研究专注于系统地探讨RAG挑战及其现有的解决方案,特别是如何有效评估其性能、如何优化检索和生成过程的协同作用,以及如何提高模型的鲁棒性等方面。同时,提出创新的优化策略,推动RAG技术在实际应用中的落地与发展。通过为学术界提供更加系统的理论框架,并为工业界提供具体的技术方案,本研究不仅推动了RAG技术在实际场景中的广泛应用,也为其在智能对话、个性化推荐、自动化推理等领域的未来发展奠定了基础。文章的结构详见图1。

本文在全面综述RAG技术现状的同时,特别关注以下四个方面:

(1)提供了RAG技术的详尽概述,包括基本概念、工作流程、技术现状、对比分析及未来发展方向,为读者提供了清晰的技术发展脉络。

(2)特别提出了“检索”“生成”和“增强”这三个RAG核心组件的基础与高级策略,深入探讨了它们如何相互作用以形成一个有效的RAG框架及其原理与应用。

(3)总结了RAG的评估方法,涵盖了近20个任务和指标,提供了对现有评估基准和工具的全面概述。此外,进一步提出了一个创新的评估框架,深入探讨了RAG在实际应用中面临的挑战,并提出了相应的潜在解决方案。

(4)关注了RAG技术在实际应用中的未来研究方向,并提出预测和建议,旨在为学术界和工业界提供实用的解决方案和指导。

1 RAG概述

1.1 RAG范式

由于ChatGPT依赖预训练数据,它缺乏提供最新动态更新的能力。例如,用户向ChatGPT提出特定领域的问题时,若查询内容超出其训练数据的范围,特别是涉及特定领域未知数据或需要最新信息的情况,LLMs可能会表现出较低的生成质量。RAG通过从外部数据库中整合知识,弥合这一信息鸿沟。它通过收集与查询相关的内容,与原始用户问题相结合,形成全面的提示,从

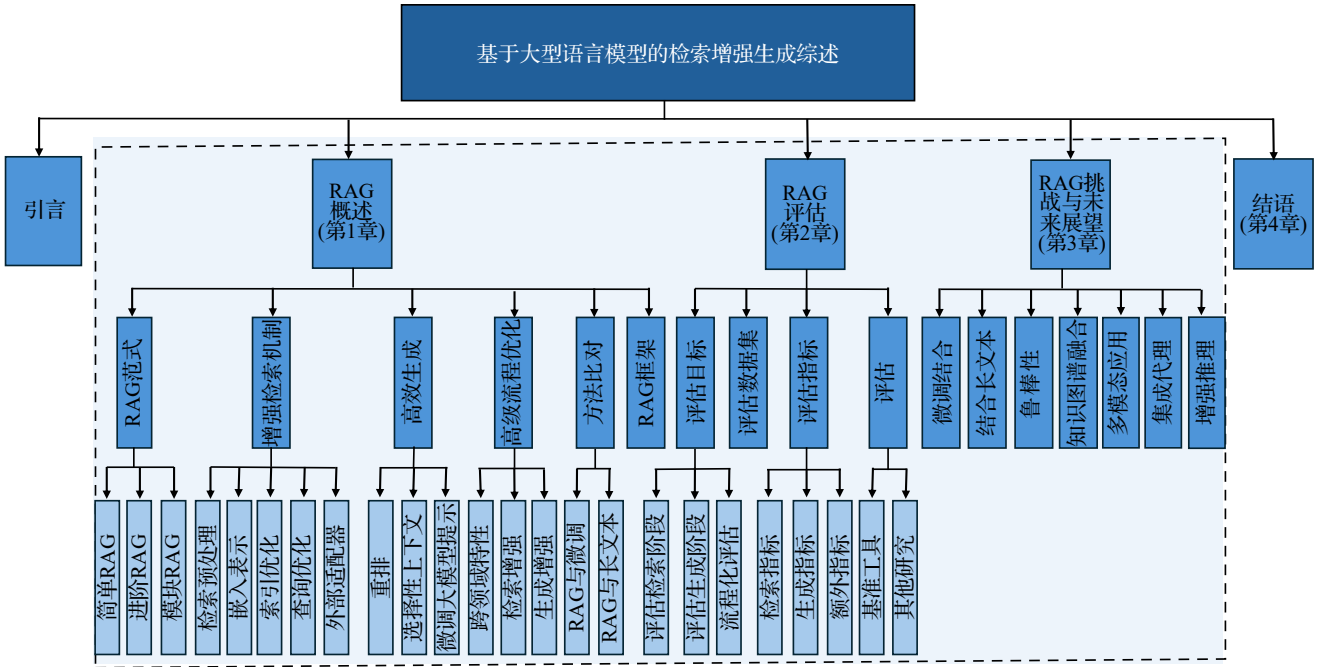

图1　文章结构示意图

Fig.1　Schematic of article structure

而提升LLMs生成答案的准确性。图2运用问题实例展示简单RAG的工作流程，突出其在处理用户查询时的实际步骤：索引、检索和生成。

(1)索引

收集并统一转换文本、音频和视频等不同格式的数据，通过分割数据为更小的、可处理的块(Chunk)，以更好地适应LLMs的语境限制。然后，进行嵌入向量化。它通常使用的是预训练的文本嵌入向量化模型，如BERT、OpenAI的ada-002或其他Transformer架构模型。例如，BERT模型处理最多512个tokens序列。上述模型将文本转换为高维向量，从而捕捉文本的语义信息，并存储结果在向量数据库中，构建索引。

(2)检索

收到用户查询后，RAG使用索引阶段相同的编码模型将查询转换为向量表示。为了精确匹配，RAG通过计算查询向量与索引语料库中各Chunk向量的相似度，识别出与用户查询最相关的文档片段。这一过程通过计算向量之间的余弦相似度，为所有潜在Chunk进行

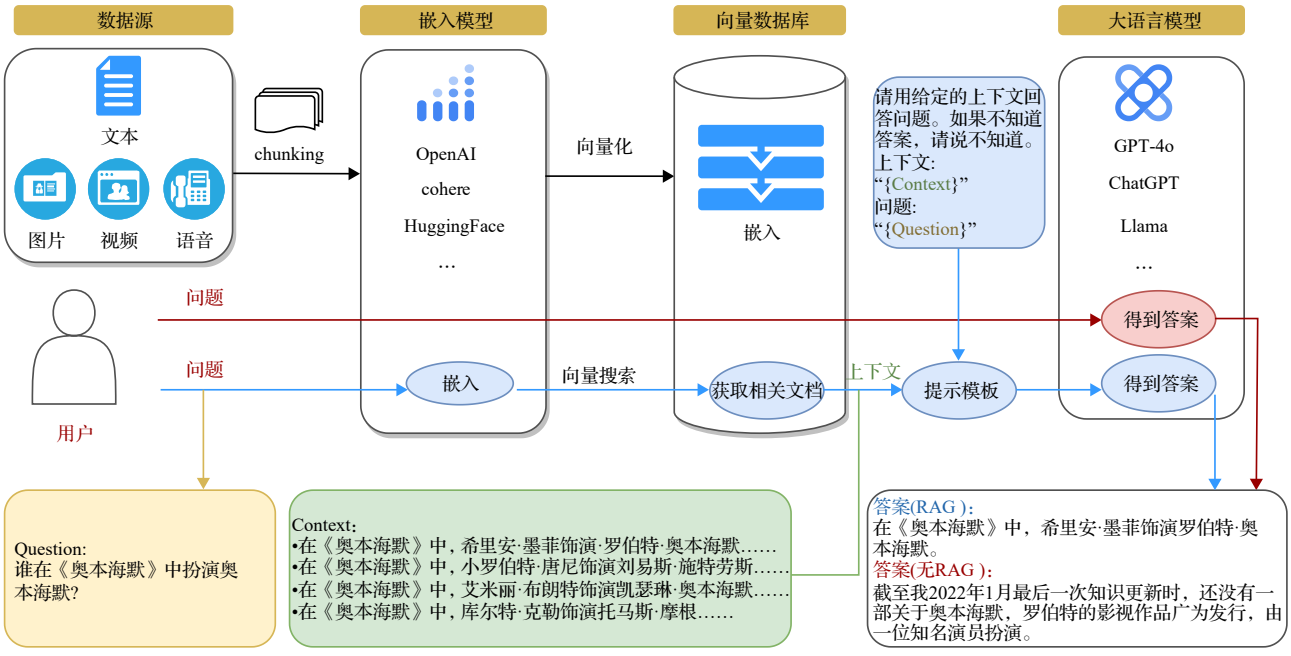

图2　RAG简单工作流程

Fig.2　RAG simple workflow

相似度评分。RAG根据这些评分对Chunk进行排序,挑选出最相关的前 K 个Chunk,并将这些Chunk整合到提示(prompt)中。

(3)生成

在RAG框架中,用户提出的问题转化为嵌入向量,通常由预训练模型如BERT或OpenAI的ada-002完成。随后,这些向量与选定的文档构建prompt,为大型语言模型(LLMs)如GPT-4、ChatGPT或Llama提供生成回答所需的上下文。该过程使LLMs的回答策略会根据不同任务的标准进行调整,提供的答案信息既可以是固有参数,也可以限制于检索文档。此外,在当前对话时,任何现有对话历史都可以集成到prompt,使LLMs进行多回合对话交互。与单独使用LLMs相比,对话式交互显著提升回答的质量和相关性。

RAG在大模型问答任务中提高了检索准确性。然而,对于复杂的多推理问题,仅依靠原始查询的单一检索的RAG可能无法生成理想的响应质量。在这一背景下,RAG的研究范式不断发展,针对检索挑战、生成困难和增强障碍等问题的研究,逐渐从简单RAG向进阶RAG[25]、模块化RAG转换。

图3为进阶RAG在处理用户查询时的演变,其核心工作流程与简单RAG类似,均遵循链式结构。简单RAG通过嵌入模型向量化数据,并在向量数据库中检索相关信息,提供了清晰的框架。以此框架为基础,进阶RAG引入预检索和后检索的两个关键环节,提高检索准确性和生成答案质量。预检索对应于RAG工作流程的索引阶段,涵盖数据的收集、处理和向量化及构建索引,以支持高效的相似性搜索。后检索策略对应于RAG的检索和生成阶段,检索到的相关信息被用于构建提示,并由LLMs生成最终回答。

预检索关注优化索引结构和原始查询,提升被索引内容的质量。涉及的具体索引结构的分布策略有:文本解析、Chunk分割、优化索引结构、添加元数据、对齐优化和混合检索。同时,用户查询的清晰嵌入也对索引质量有显著影响。用户查询优化的目标是使原始问题更清晰,适合检索任务,常用方法包括查询重写、查询变换和扩展等技术[25-26]。例如,分层索引检索、创建假设性问答对、使用LLMs进行信息去重、测试和找寻最优分块大小等方法。同时,嵌入模型的使用,如FlagEmbedding的LLMs,使RAG在性能和大小之间取得平衡。

后检索策略关注于有效集成检索到的上下文与查询,主要方法包括重排和选择性上下文。重新排序是指将最相关的信息重新定位到提示符的边缘,以提高提示符的相关性,已在LlamaIndex2等框架中实现。为避免信息过载,后检索工作集中在选择、压缩关键内容并过滤不相关信息,以提高效率并减少token使用。进阶RAG利用现有的框架和工具,如LlamaIndex等,提供了强大的功能支持摘要和融合过程,实现从大量数据中快速准确地检索并生成答案。

图4展示模块化RAG范式的多功能性和适应性,该范式不局限于顺序检索和生成,还引入迭代和自适应检索等增强RAG流程方法。它在多个具体功能模块的引入体现得非常明显。模块化RAG框架引入新组件,以增强检索和处理能力。例如,Predict模块旨在通过LLMs直接生成上下文,从而减少冗余和噪声,确保相关性和准确性[9]。内存模块利用LLMs的记忆能力优化检索过程,通过构建无界内存池和迭代增强,使文本和数据分布更加紧密对齐[27]。RAG路由通过不同的数据源进行导航,为查询选择最佳路径。这些方法简化检索过程,提高信息的质量和相关性,以更高的精度和灵活性迎合广泛的查询。

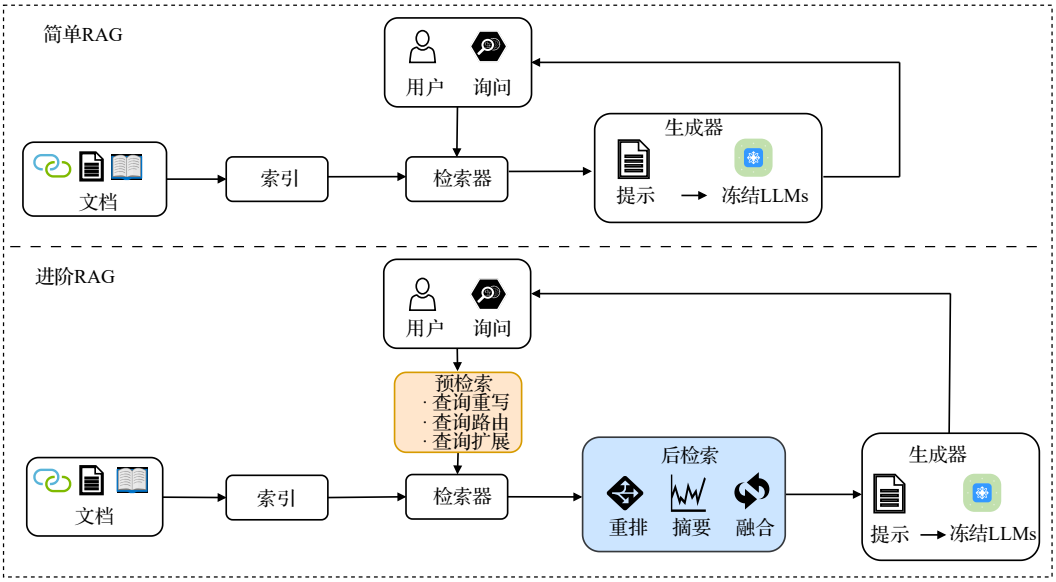

图3　进阶RAG范式的演变

Fig.3　Evolution of advanced RAG paradigm

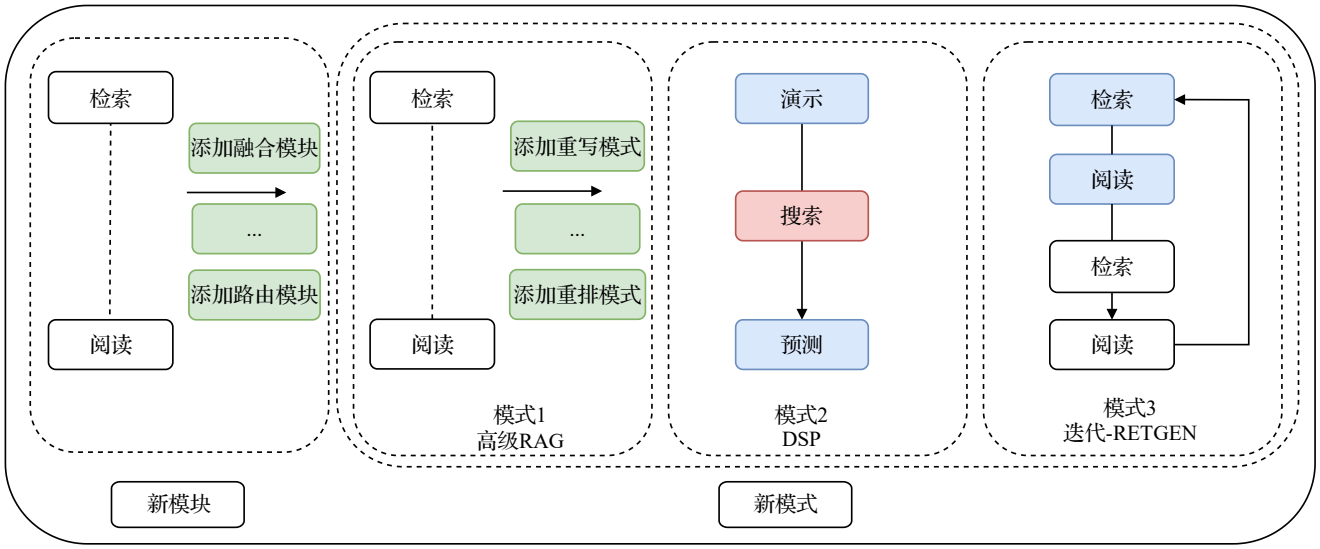

图4 模块化RAG范式

Fig.4 Modular RAG paradigm

模块化RAG通过允许模块替换或重新配置来解决特定的挑战。Rewrite-Retrieve-Read[28]利用LLM的迁移能力,通过重写模块和LLMs反馈优化检索查询,以此更新重写模型。演示-搜索-预测(DSP)框架和I-TER-RETGEN[29]的迭代检索-读取-检索-读取流程,表示模块输出的动态使用,说明模块安排和交互的调整。因此,围绕检索器、生成器和RAG的高级流程优化的三个核心部分在1.2~1.4节回顾了重要技术研究,整理了不同阶段的RAG所使用的不同方法,由表1所示。

1.2 增强检索机制

1.2.1 检索预处理

(1)文本解析

文本解析是指将非结构化或半结构化的文本数据转换为结构化信息的过程。无论是处理用户询问还是检索数据源,文本解析的质量对LLMs的生成效果具有关键影响。部分研究侧重于改进对用户问题的解析,以

表1 RAG方法总结

Table 1 Summary of RAG methods

RAG过程	文献	检索增强生成方法
检索预处理	SKR[30]、Adaptive-RAG[31]	基于问题精准的文本解析
检索预处理	DenseX[32]、TableGPT[33]、ISEEQ[34]、knowledgpt[35]、G-Retriever[36]、	基于外部检索源的文本解析
检索预处理	LLAMAi[37]、Small2Big[38]、Meta-Chunking[39]、GLMwCFiCR[40]、RAAT[41]	Chunk分割
嵌入表示	RALM[42]、DAPR[43]、SURE[44]、SBRR-RAG[45]	混合检索
嵌入表示	C-Pack[46]、RetroMAE-2[47]、REPLUG[48]、ARl2[49]、Promptagator[50]、BGE[51]、AnglE[52]、ListT5[53]	微调嵌入
索引	HyDE[54]	分层索引
索引	RAPTOR[55]	树形索引
索引	KGP[56]	知识图谱(KG)索引
查询	RAG-Fusion[57]	多响应扩展
查询	LTMP-CLR[58]	子查询
查询	RRR[28]、BEQUE[59]	查询转换/重写
查询	Step-back prompt[26]、FCTG-ICRA[60]、DPR[61]	查询路由
外部适配器	UPRISE[62]	轻量级提示检索器
外部适配器	Search-Adaptor[63]、PRCA[64]、CRAG[65]	特定于任务的检索器
外部适配器	GenRead[66]、PKG[67]	引入LLMs生成器
检索信息生成	RAG-Fusion[57]、Filter-Reranker[68]、G-RAG[69]、RankRAG[70]	重排
检索信息生成	Lingua[71]、LongLLMLingua[72]、FILCO[73]、RECOMP[74]、ACD-RAG[75]	选择上下文/上下文压缩
检索信息生成	RAAT[39]、IBP-ENFoRAG[76]、Chatlaw[77]	信息过滤
提示输出	URR-RAIC[78]	微调数据
提示输出	SANTA[79]、DFK-TOD[80]、RA-DIT[81]	强化学习/对齐
高级流程优化	RQ[82]、SIB-SLLM-IR[83]、ERLLMwIRGS[84]、Efficient RAG[85]、SMBI-SSPM[86]、RAFV-SCA[87]、RAT[88]	迭代检索
高级流程优化	IRCoT[89]、ToC[90]、CK-GLMwSKB[91]	递归检索
高级流程优化	Graph-ToolFormer[92]、WebGPT[93]、Flare[94]、Self-RAG[95]、RACG[96]、Self-Reasoning[97]、IM-RAG[98]	自适应增强

提升模型的文本理解能力。SKR[30]将问题解析为已知或未知,并结合RAG方法处理未知问题,从而提高了LLMs的生成答案精准度。Adaptive-RAG[31]设计了小型问题分类的解析器,依据问题复杂度动态调整解析策略,兼顾效率和处理复杂任务的能力。针对RAG中外部检索源的解析难度,DenseX[32]通过细粒度的文本解析和命题抽取,将检索文档内容转化为原子表达式,使每个命题代表一个独特的事实片段。这种方法能够提高检索精度和相关性,使外部信息更精确地传达关键内容,展示文本解析在处理复杂外部信息中的核心作用。对于表格数据,TableGPT[33]通过文本解析将结构化的表格内容转换为SQL查询语句,并利用LLMs的代码生成能力,根据用户输入的自然语言问题生成相应的SQL查询,从而在数据库中执行操作。这种方法使得结构化表格数据的查询高效化。另一方面,使用文档加载器(如LlamaParsePDF提取器)等文本解析技术将多样化的知识源转换为纯文本数据。文本解析器将表格数据转换为文本格式,使得这些数据能够通过基于文本的方法保证解析准确性的同时,还能处理更为复杂的任务。此外,知识图的处理[34]提升了文本解析的精确度。具体来说,knowledgpt[35]通过构建知识库,优化文本解析过程中对复杂知识的提取和利用,增强RAG模型的知识丰富性。近期针对LLMs在理解和回答文本图问题方面的局限性,G-Retriever[36]解析图神经网络,结合RAG进行目标图检索。这种集成改进了图结构信息的处理,提升了结构化数据库的构建、验证和维护效率,展示文本解析技术在知识图谱和结构化数据处理中的关键作用。

(2)Chunk分割

对于复杂查询,答案往往分散在多个文本块(Chunk)中,单独的Chunk无法提供完整上下文。常见的文本切分策略是将文档按照固定数量的token(例如,100、256、512)进行分割[37]。这种方法在处理时会面临两个主要挑战:大Chunk提供更多上下文但可能引入噪声,且处理成本高;小Chunk噪声少但上下文不足。

为了解决上述问题,Yang[38]提出了Small2Big模型,将句子(小)作为检索单元,并结合前后句子的上下文提供给LLMs,从而提升检索器效率。Meta-Chunking[39]提出边际采样分块和困惑度分块两种策略,动态合并实现细粒度与粗粒度分块的平衡。此外,为增强文本块的有效性,Chunk中添加页码、文件名、作者、类别和时间等元数据,这些附加信息增强LLMs上下文学习能力,确保其知识新鲜度,避免信息过时。相反,GLMwCFiCR[40]提出一种无需分块的上下文检索方法。这种方法将用户查询和长文档直接输入LLMs,通过预测token的概率来识别最相关的段落。Liang等人[41]同样通过不分块架构在多阶段训练实现高效的表征编码。这些方法在提升RAG增强LLMs的效果方面表现出色,但对计算资源的需求更高。

1.2.2 嵌入表示

嵌入模型是将文本转换为向量形式,并在向量空间中评估其关系,视作嵌入表示。例如,“苹果是一种水果”和“香蕉是黄色的”会被转换为向量进行比较。

在通用RAG框架中,嵌入模型包括稀疏编码器和一个密集检索器。它们分别以不同的方式进行语义表示。稀疏编码器依赖于词频统计,而密集检索器通过深度学习模型提供更加细致的语义表示。为了避免单一嵌入表示的知识鸿沟,RALM[42]结合了稀疏和密集检索,强调从模型权重中检索,增强处理知识密集型任务的能力,并提高LLMs鲁棒性。嵌入表示整合了关键字、语义检索和向量搜索的混合检索方法,以满足多样化的查询需求。Wang等人[43]通过采用混合检索方法,对长文档的检索源进行语义增强,从而提升RAG的检索性能。Kim等人[44]提出的SURE框架结合多种检索方法生成摘要文档,有效支持问答任务。同时,研究表明[45],将RETRO模型的传统检索替换为基于表面层相似度的嵌入方式,显著降低检索困惑度。此外,对于上下文明显偏离预训练语料库的情况,特别是在医疗保健、法律实践等高度专业化的学科中,实现微调嵌入模型至关重要[46]。例如,Liu等人[47]提出DupMAE模型来预训练模型的所有上下文化嵌入,提高语义表示的质量提升嵌入表示。除了补充领域知识外,微调嵌入表示的另一个目的是对齐检索器和生成器,以提升嵌入表示准确性。REPLUG[48]通过LLMs监督信号来对齐检索器,无需特定交叉注意机制的需求,从而提升嵌入长尾知识的精准度。如Zhang等人[49]通过对比学习和成对的logistic损失微调检索器训练,使嵌入模型适应不同的下游应用。Dai等人[50]利用LLMs生成少量任务进行特定查询,解决了监督式微调嵌入的挑战,特别是在数据稀缺的领域。特别地,受强化学习(reinforcement learning with human feedback,RLHF)的启发,基于LLMs的反馈也通过RLHF强化嵌入表示[67]。例如,使用LLMs阅读器的反馈强化学习训练[8]。上述方法说明微调嵌入模型有利于提升检索性能和适应性。

最近的研究通过选择嵌入模型,增强了LLMs长上下文的结构表示。常见的RAG嵌入模型包括BGE[51]、AnglE[52]、Voyage和ListT5[53]等。它们在多个应用场景中显著改善了处理实际任务的表现。

1.2.3 索引优化

索引质量决定向量数据库的构建,从而影响检索上下文的准确性。RAG利用分层索引[61]多次检索、合并上下文相关信息,并动态调整Chunk的大小以优化句子的截断效果,从而为数据库构建出精确的向量表示。文档索引采用子查询和假设文档嵌入(HyDE)[54]的方法,通过提高答案和真实文档之间的嵌入相似性来加快相关数据的检索,减轻由块提取引起的语义完整性和长上下文不平衡性。

在处理多层嵌入、分群和概括的文本内容时,RAPTOR[55]则提出一种树型组织索引方法,通过构建分层摘要树逐层精炼信息,形成结构化的检索语料库,适用于摘录式提取。此外,知识图谱(KG)索引通过构建文档的层次结构,描绘不同概念和实体之间的联系,减少了错觉。对于进一步捕捉文档内容和结构之间的逻辑关系,KGP[56]利用知识图谱(knowledge graph,KG)在多个文档间建立索引,通过节点(表示文档中的段落或结构,如页面和表)和边(表示段落之间的语义或结构关系),有效地解决多文档环境的知识边缘检索和推理问题。

近期的RAG研究,常用的向量数据库如Faiss、Milvus和HNSW等,通过优质的索引构建,在处理大量数据时迅速返回相关的文档片段。这些数据库的索引技术为RAG的有效检索提供强大支持。

1.2.4 查询优化

有效的查询不仅加速信息返回,还减少了检索过程中的噪音,从而显著提升用户体验。RAG的主要挑战之一是它直接依赖用户查询检索。有时,复杂或模糊的问题会限制查询的效果。例如,LLMs在处理专业词汇或具有多重含义的缩写时可能难以辨别“LLMs”指的是“大语言模型(large language models)”还是法律语境中的“法律硕士(legal master)”。

查询优化的方法包括多响应扩展等策略,用于增强用户查询。基于多响应扩展,LLMs通过提示工程将复杂查询扩展为多个并行执行的子查询,以得到更好的答案相关性输出。同样,分解单响应变为子查询也是一种合理的手段,利用逐一在上下文中添加必要的子问题并生成回答。具体来说,LTMP-CLR[58]应用最小到最大提示法,将复杂问题分解成更简单的子问题,从而使每个子问题都能在上下文中得到充分回答。这一过程通过单个查询扩展为多个查询来丰富查询内容,提供了进一步的上下文以解决缺乏具体细节的问题,从而确保生成答案的相关性。

基于涉及检索库中的复杂语义,查询转换是优化策略之一。例如,RRR[28]提示LLMs查询重写的方法在淘宝上的实现,被称为BEQUE[59],显著增强了长尾查询的召回效率。HyDE[54]通过构造假设文档,关注答案间的嵌入相似度而不是问题本身,进一步增强查询的相关性和精确度。此外,查询路由为RAG提供了额外的灵活性。通过动态选择不同的检索文档或查询路径,查询路由使RAG能够适应各种场景中的长上下文和复杂语义处理。RAG根据查询与检索内容的相关性来双重界定路径,从而引导至最合适的RAG管道。无论涉及摘要生成、特定数据库搜索还是合并不同的信息流,查询路由都能选择最佳路径[60]。最后,Step-back prompt方法[26]为查询优化提供了抽象策略。基于这种方法,原始查询被抽象为更高级的概念问题,并进行更广泛的检索。具体而言,RAG技术同时利用退步问题(fallback queries,FQ)和原始查询的结果来生成最终答案。通过结合这两种查询方式,扩展查询经过LLMs的验证,相比原始查询通常具有更高的可靠性[57],从而减少幻觉,提高大模型的答案质量。

1.2.5 外部适配器

微调模型可能受限于本地计算资源,因此,一些方法通过API集成,选择合并外部适配器来帮助校准。

UPRISE[62]训练了轻量级提示检索器,自动从预先构建的提示池中检索适合给定零任务输入提示。除了传统使用的零样本设置之外,LLMs利用来自查询-语料库来数据配对也能进一步提高黑盒LLMs Embedding的能力。Search-Adaptor[63]利用其有效性和鲁棒性的特点定制LLMs的信息检索。该适配器修改由黑盒LLMs生成的表征,并调用API等接口,任何LLMs都能与它集成。PRCA[64]则添加一个可插拔的奖励驱动上下文适配器,旨在提升特定任务的生成性能。CRAG[65]训练了T5-large模型,评估检索文档的整体质量,筛选关键信息与RAG的框架无缝耦合,同样增强LLMs生成能力。

GenRead[66]引入LLMs生成器取代传统的文档检索器,实现RAG增强。通过生成文档直接获取答案,说明生成器与检索器进行结合可以达到更好的表现,并说明了充分利用LLMs内部知识的挑战。PKG[67]则通过替换检索器模块,并利用指令微调将知识集成到白盒模型中,从而优化特定的搜索任务。这种方法有助于解决在微调过程中遇到的困难,并增强模型性能。

1.3 高效生成

选择最相关的检索段落信息有利于提高生成答案的准确性。在LLMs充分利用检索信息的情况下,保证最小化计算资源的消耗,最大化相关知识与答案相似性。因此,系统回顾了生成如何利用检索信息进行高效生成。

1.3.1 重排

重排的定义是优化与检索问题高度相关的前 k 个背景知识段落的顺序,并对其进行评分,从而有效减少无关文档的影响。重排执行使用基于规则的方法,通常依赖于预定义的指标,如多样性、相关性及平均倒数排名(mean reciprocal rank,MRR)。另一方面,采用基于模型的方法。例如BERT系列中的编码器-解码器模型(如SpanBERT),以及专门的重新排序模型,包括一些通用的大型语言模型,如GPT等。例如,Ma等人[68]提出的“Filter-Reranker”范式,结合了LLMs和小型语言模型(small language model,SLM)的优势。在这个范例中,SLM充当过滤器,LLMs充当重新排序模型。研究表明,由SLM识别的具有挑战性的样本指导LLMs重新排列,可以显著改善信息提取任务的效率。

重排既是增强器又是过滤器,它为语言模型处理提供了精细的输入。G-RAG[69]结合文档图和AMR图优化

RAG,减少计算资源消耗,并为解决排名得分平局问题提供了新思考。因此,重排RAG[66]已被引入应对具体检索问题。在面对业务场景时,RankRAG[70]进行排列评分并将相关性最强的检索段落置于前列,从而在RAG中同时实现上下文排序和答案生成的双重目标。在搜索任务中,RAG提出的演示-搜索-预测[31]框架展示了模块输出的动态支持另一个模块,表明RAG重排机制发挥了灵活的协同作用。知识GPT则结合RAG检索内容,使用LLMs进行特定的排序评分,并生成代码和查询语言[35],从而实现了跨多种数据源(如搜索引擎、数据库和知识图谱)的直接搜索。此外,Rackauckas等人[57]通过应用多响应扩展策略,从不同角度扩展用户查询,结合并行向量搜索和智能重排,发现显性和变革性的知识边缘,这一过程有助于克服传统搜索的局限性。

1.3.2 选择性上下文

RAG在管理长文档和扩展对话时,导致内存和推理时间的计算需求显著增加。当输入超过LLMs的固定上下文长度时,可能出现上下文截断。因此,选择性上下文通过识别和修剪输入上下文中的冗余信息,使输入更加紧凑,从而提高LLMs的推理效率。Lingua[71],LongLLMLingua[72]利用SLM的优势,用GPT-2 small、LLaMA-7B和LLaMA-8B等,检测并删除不重要的段落,并转换为LLMs更好地理解的语言形式。这种方法为快速选择适合的上下文段落提供了一种直接且实用的途径,使RAG在平衡语言完整性和压缩比的同时,消除了对LLMs的额外训练要求。PRCA[64]通过训练一个信息提取器来解决上述问题,其优势在于能够适应多种下游任务。FILCO[73]训练小型上下文过滤模型,基于词汇和信息论方法识别有用的上下文并测试。它强调了检索段落的关键部分,缩短了处理的上下文长度。类似的,RECOMP[74]使用对比损失进行训练编码器,以选择最有用的上下文来完成LLMs的提示。而Kim等人[75]则在存在噪声上下文的情况下,使用自适应对比解码(ACD)策略,有效提升上下文利用效率。

过多的上下文也会引入更多的噪声。Zhu等人[76]提出了信息瓶颈理论过滤检索段落的噪声使上下文互信息最大化,从而得到更好的地面输出结果。这种方法有效地减少上下文的嘈杂信息,提升了LLMs的生成提示的正确性。RAAT[39]设计一种噪声分类损失自适应对抗性训练LLMs,通过多任务学习来确保模型内在的识别噪声,选择输出上下文的有用信息。合理检索并过滤相关文档的数量也有助于减少上下文的噪声。最直接有效的方法之一是让LLMs在生成最终答案之前评估检索到的内容,使LLMs通过自身评论过滤掉相关性差的文档。如Chatlaw[77]提示LLMs引用法律条款进行自我建议,评估其相关性。

1.3.3 微调LLMs提示

LLMs适应于特定的数据格式(如问答格式),并按照指示[89]生成特定风格的响应。微调LLMs提示(fine-tuning LLMs prompts,FT-LLMs)指的是根据特定任务(如问答任务)或应用场景的需求,对LLMs的输入输出提示(prompts)进行优化和调整的过程。当LLMs缺乏特定领域的数据,通过微调提示提供给LLms额外的知识。Huggingface的微调数据可以作为这一过程的初始步骤。例如,在图像标题生成任务中,模型容易被出现较多的token误导,将错误token复制到输出。URR-RAIC[78]利用Huggingface数据处理,采取多样化的标题集合,从而防止模型复制多余的token,有效地增强数据特征。对于涉及结构化数据的检索任务,SANTA框架[79]有效地封装上下文结构和语义上的细微差别,实现了LLMs的有效输出。另外,LLMs通过强化学习将生成输出与人类的偏好对齐。例如,手动注释最终生成的答案,然后通过强化学习提供反馈。

微调模型和检索器的偏好同样使输出保持一致[80]。当无法访问强大的专有模型或更大参数的开源模型时,一种简单而有效的方法是提取更强大的模型(例如:GPT-4)中提取知识提示输出。LLMs的微调也可以与搜索引擎的微调相协调,以对齐偏好输出。一种典型的方法,如RA-DIT[81]使用KL散度来对齐retriver和Generator之间的评分函数。

1.4 高级流程优化

高级流程优化是整合检索与生成的高级策略。通过提升检索精度与生成相关性,流程优化显著提高RAG的跨领域特性。它确保RAG检索与生成协同增强方面的灵活性,通过在检索阶段优化相关信息的选择,并结合生成阶段的上下文理解与内容质量提升,实现两者协同工作,提升了RAG在复杂任务中的适应性与输出效果[82]。高级流程优化详解如图5所示。

高级流程优化分为检索增强与生成增强。具体而言,RAG采用迭代和递归两种检索形式,使得检索过程能持续优化信息获取。迭代检索通过多次循环,逐步精炼查询条件和结果,而递归检索则在生成过程中根据需要动态调整检索深度和广度,二者均有助于更有效地收集和利用背景信息。同时,生成增强则强调LLMs智能决策能力,通过自适应增强机制和特殊令牌控制检索与生成的时间点,提高信息源的相关性和效率。

1.4.1 检索增强

(1)迭代检索

迭代检索旨在逐步增强生成答案的鲁棒性。每次迭代都会引入新的相关背景知识,从而提高最终生成内容的质量。然而,不连续的语义和不相关的信息积累可能会干扰这一过程。

为解决这一问题,Feng等人[83]提出一种引入多样化检索策略的模型(retrieval model,RM),包括多模态检索、反馈循环机制和语义增强等,与LLMs协同工作,以

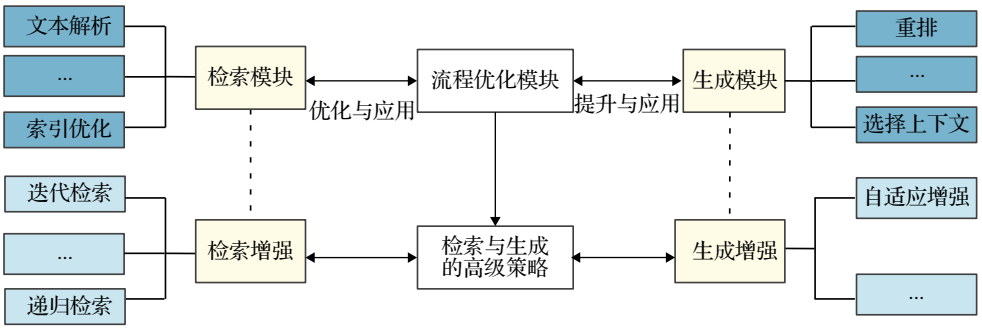

图5 高级流程优化详解

Fig.5 Detailed explanation of process optimization

优化信息获取效率和准确性。Shao等人[84]则利用“检索增强生成”和“生成增强检索”的方法,确保在迭代中通过检索相关上下文知识,提升生成响应的准确性。EfficientRAG[85]在迭代中自主生成新查询,有效剔除无关信息,成为处理多跳问答任务的高效检索器。与上述方法不同,Tan等人[86]训练小型语言模型(language model, LM)作为评估器,评判每一个问题的回答是否正确,并决定是否进行额外检索。研究指出,评估器在迭代过程中的引入,使得模型能够快速识别需要额外信息的场景,从而提升了迭代检索的效率。为了获取上下文的真实信息,Yue等人[87]强调了检索阶段增加文档重要性评估。它运用粗排和重排相关文档来迭代检索,在每次检索后实施重排序策略,优化上下文信息真实度。在处理长文本结构中存在错误分层语义信息的情况中,Wang等人[88]精炼思维链的结构,确保在每次迭代中,生成的回答不仅具备逻辑性,还能有效应对复杂的上下文信息。

(2)递归检索

为了提高搜索结果的深度与相关性,递归检索逐步细化用户查询,将问题分解为子问题。IRCoT[89]利用链式思维(chain-of-thought, CoT)来指导检索过程,并通过检索到的结果进一步细化CoT。ToC[90]采用澄清树结构,优化查询中的歧义部分。该方法在复杂的搜索场景中尤为有效,特别是当用户需求最初不明确,或者搜索高度专业化的信息时。因此,递归特性广泛应用RAG的流程,使RAG不断学习和适应用户需求,从而显著提高用户对搜索结果的满意度。

在递归使用中,RAG通过不断从历史信息中提取结果与精炼搜索查询。具体来说,基于更复杂的多步骤问答任务,简单问答直接生成答案,多步骤问答则通过递归检索相关文档并追溯问答对,验证并修改答案中的错误。递归流程增强为RAG打开了新思路。在特定的数据场景下,RAG将递归检索和多跳检索技术结合使用。递归检索涉及分层与知识图谱(KG)索引时,首先对文档或冗长PDF的部分进行摘要,然后,利用文档二次检索细化搜索,彰显其递归特性[91]。递归增强旨在更深入地研究图结构,以提取相互关联的信息。

1.4.2 生成增强

自适应增强作为生成增强的一种重要形式,侧重于使RAG能够自主地决定是否需要外部知识检索以及何时停止检索和生成,通常利用LLMs生成的特殊令牌进行控制。这种方法使LLMs主动确定检索的最佳时刻和内容,从而提高信息源的效率和相关性,进一步完善RAG框架。正如在AutoGPT、Toolformer和Graph-ToolFormer等先进的模型代理(Agent)中观察到的趋势显示,LLMs在其运作过程中正逐渐融入主动判断的机制,这标志着一个更广泛的技术演进方向。Graph-ToolFormer[92]将检索过程划分为不同步骤,LLMs应用Self-Ask技术主动使用检索器,并通过少量提示启动查询。这种主动学习的姿态使得LLMs能够自主决定何时检索必要的信息,类似于智能Agent(代理)在需要时如何有效使用工具。WebGPT[93]集成了强化学习框架,在文本生成过程中使用搜索引擎自主训练GPT-3模型。它利用特殊的令牌来导航,促进搜索引擎查询、浏览结果和引用操作,从而扩展GPT-3的功能。

Flare[94]和Self-RAG[95]则根据反馈结果决定是否需要调用外部检索增强数据源。Flare通过实时监控生成过程的置信度,自动执行定时检索[94]。例如,调整生成项的概率,当概率低于某一阈值时,会激活检索系统收集相关信息,从而优化检索周期。Self-RAG[95]引入了“反射令牌”机制,允许模型自省其输出,分为两种形式:“检索”和“批评”。模型自主决定何时激活检索或达到预定义阈值时触发检索过程。在检索阶段,生成器跨多个段落进行搜索,派生出最连贯的序列。评论家分数用于更新细分分数,并在推理过程中灵活调整这些权重,定制模型行为。Self-RAG的设计不需要额外的分类器或依赖于自然语言推理(natural language inference, NLI)模型,从而简化检索决策过程,并提高模型在生成准确响应的自主判断能力。RACG[96]使检索器能够从生成器的反馈中自增强学习,从而检索出更有助于提升注释质量的示例。百度推出的Self-Reasoning框架[97],通过相关意识,证据意识选择和轨迹分析过程,有效增加LLMs自身生成的推理轨迹,提高RAG模型的可靠性和

可追溯性。此外,IM-RAG[98]通过强化学习实现端到端优化,解锁AI内心独白,提供多轮检索的可解释性,为未来处理高度复杂、抽象或创造性推理任务方面的学习与应用带来思考。

1.5　RAG对比

本节将RAG与微调和长文本LLMs的对比。这一对比帮助全面理解RAG在不同应用场景下的

独特价值及局限性。表2总结了RAG、微调和长文本LLMs的基本原理、优缺点以及适用场景,使读者能够直观地比较这些方法的特点和效果。

1.5.1　RAG对比微调

研究者们越来越关注如何优化大模型的性能。在众多优化方法中,RAG经常被拿来与微调(fine-tuning,FT)和提示工程(prompt engineering,PE)进行比较。这三种技术各有优势,它们在提升模型性能方面展现出不同的潜力和应用场景。例如,面向生物科学研究领域,BIORAG[99]通过动态检索最新的生物信息,创建高效的生物问答系统。它通过问题解构,结合搜索引擎的迭代检索进行分步推理问答。而在推荐系统应用[100]中,微调可以针对小样本学习完成数据修剪,快速扩展大模型适应新的推荐数据集。提示工程则充分利用模型的固有能力,在不依赖外部知识的情况下进行有效的文本生成。

图6通过象限图来说明三种方法在外部知识需求和模型适应需求两个维度上的差异。提示工程将外部知识和模型适应的必要性降至最低;RAG可看作为信息检索提供一个带有定制教科书的模型,非常适合精确的信息检索任务,如聊天机器人实时回答用户问题或法律咨询系统查询最新法规等。相比之下,FT随着时间的推移内化知识,适用于需要复制特定结构、风格或格式的场景,例如使用特定的情感标注数据集进行情感分析。

在多次评估中,RAG在处理知识密集型任务时,相较于无监督微调,展现出在整合现有知识和吸收新知识方面的优势。如文献[101]深入分析RAG以及微调增强LLMs处理长尾知识的能力差异。研究验证,LLMs通过RAG检索知识弥补了LLMs无监督微调中学习新信息方面的不足。尽管RAG在动态环境中表现出色,但在数据检索方面可能会导致更高的延迟,并引发伦理考虑。另一方面,FT是一种更为静态的优化方式,在模型行为和风格的定制上提供深度控制,但可能在面对未知数据时遇到挑战。

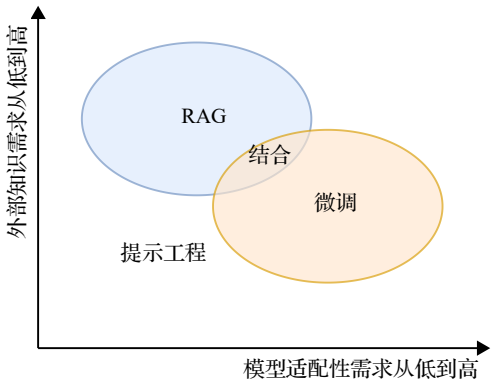

图6　RAG与提示工程、微调对比

Fig.6　Comparison of RAG with prompt engineering and fine-tuning

RAG和FT之间的选择取决于应用程序上下文中对数据动态、自定义和计算能力的特定需求。研究表明,RAG和FT并非相互排斥,而是可以在不同层次上互补,以增强模型的能力。文献[102]给出搜索示例:“欲望之岛的导演是谁?”,发现FT和RAG结合能展现出强大的泛化能力。然而,优化过程中可能需要多次迭代,才能获得理想结果。

1.5.2　RAG对比长文本LLMs

随着相关研究的不断深入,长上下文语言模型(long context language model,LCLM)的应用范围也在持续扩大[103]。目前,像Claude-3和Gemini-1.5-pro等模型,已经能够轻松处理数十万甚至数百万个token的序列。面对长上下文任务,LCLM可以轻松完成长篇研究报告自动化撰写。另一种范式,检索增强生成(RAG),作为这些长上下文LLMs的有力补充,采用流水线方法,使用检索器动态选择用于查询的上下文,从而减轻生成器直

表2　RAG对比

Table 2　RAG comparison

维度	RAG(检索增强生成)	微调(fine-tuning)	长文本大模型(long context language model)
原理	结合检索与生成,通过外部知识库增强生成内容	调整模型参数以适应特定任务	处理长上下文,整合复杂管道为单一模型
优点	实时访问最新信息 提高生成内容的准确性 易用性和时效性	深度控制模型行为和风格 能够内化知识	处理长文本能力强 改善级联误差 适应多轮对话
缺点	可能导致较高的检索延迟 存在伦理问题	在未知数据上可能表现不佳 需要大量计算资源	依赖于模型的上下文长度 可能无法保证时效性
应用场景	知识密集型任务,动态环境	需要特定结构或风格的场景	多轮对话、需要大量上下文的信息
应用案例	医疗问答系统(实时检索医学文献) 聊天机器人(法律咨询)	情感分析(使用情感标注数据集) 推荐系统(扩展推荐数据集)	自动撰写长篇研究报告 客服对话(确认信息准确性)

接处理长上下文的需求。如医疗问答系统中[104],RAG可以实时检索最新的医学文献,确保生成的回答不仅准确而且符合最新的研究成果。

长上下文大模型具备超大的内存,实现全新的任务和应用,同时消除上下文长度限制带来的复杂工具和管道的依赖。LCLM通过将复杂的管道整合为统一模型,改善级联误差[105]和优化等问题,为模型提供了一种精简的端到端方法。在多轮对话场景,LCLM通过上下文理解增强,来提升模型在对话中的表现。这些技术包括意图识别、槽位填充、状态管理策略决策等,通过蒙特卡洛方法和LLMs生成训练数据集,高效微调多种语言模型,从而构建任务型对话Agent。LCLM通过简化架构,提高处理效率,并减少错误传递的可能性。另一方面,RAG的特点是易用性、事实性和时效性。通过RAG的方式,将原有大模型的元素变成多维标签,甚至将RAG本身设计为端到端的向量或标签化实体,以防信息损失。RAG为LLMs开启一个外部知识库,类似添加一块移动硬盘,使得大模型能够快速访问最新的知识。对于一些严肃的场景中,如法规条文、保险或教育等,RAG将法规条文、政策文件纳入数据库,通过RAG生成精准的法规解答,避免产生错误解读。此外,对于软件工程领域,涉及代码的补全、翻译或重构时,输入token数量往往非常庞大。仅依靠滑动窗口处理会导致理解障碍,而使用RAG可以保证私人数据隐私的同时,提升知识生成的效率。如开发人员在代码重构时,结合信息检索技术和神经生成模型,从而提供更精确的建议[96]。在某些情况下,将LCLM和RAG结合使用,通过向量库检索文本并集中召回,随后大模型的整合可以提升生成效果。

1.6 RAG框架

针对日益深入的RAG研究,RAG框架为实践提供了支持。表3汇总目前主流的RAG通用框架,并说明每个框架都有独特的应用场景。

LlamaIndex是一个专为检索增强生成(RAG)设计的强大工具集,通过“摄取-索引-查询”流程来管理结构及非结构化数据。它能够从多种来源(如企业内部系统)中导入数据,并使用先进的向量索引技术将信息转换成易于搜索的形式,有效关联语义相似但表述不同的内容。例如,当查询是“皇室成员”时,国王和王后是高度相关的;但若查询是“性别”,两者的相关性则较低。另外,数据的优化方式包括列表索引、树索引和关键词索引。用户经过索引处理后就可以利用自然语言界面来进行直观而高效的查询操作。这使得无论是构建客户服务聊天机器人还是开发知识助手,都能高效访问与利用大规模数据资源。对于医疗领域而言,集成LlamaIndex可以帮助医生参与到眼科疾病治疗方案的设计当中,从而显著提升诊疗质量和效率[106]。相比而言,HayStack[107]和LLAMA_Index2也致力于优化信息检索[1-39],但它们各自侧重不同的方式结合传统检索技术和最新

表3 RAG开源框架

Table 3 RAG open source framework

框架名称	项目地址	简介
LlamaIndex2	https://github.com/run-llama/llama_index	LLM应用程序的数据框架
FastGPT	https://github.com/labring/FastGPT	基于LLMs的知识库问答系统,提供开箱即用的数据处理、模型调用等能力
Langchain-Chatchat	https://github.com/chatchat-space/Langchain-Chatchat	LLMs与Langchain等应用框架实现,可离线部署的RAG大模型知识库项目
RAGFlow	https://github.com/infiniflow/ragflow	基于深度文档理解构建的开源RAG引擎
网易QAnything	https://github.com/netease-youdao/QAnything	支持任何格式文件或数据库的本地知识库问答系统,可断网安装使用
MaxKB	https://github.com/1Panel-dev/MaxKB	基于LLMs的知识库问答系统
open-webui	https://github.com/open-webui/open-webui	可扩展、功能丰富且用户友好的自托管WebUI,旨在完全离线操作
HayStack	https://github.com/deepset-ai/haystack	端到端LLMs框架,允许构建由LLMs、Transformer模型、向量搜索等
langraph	https://github.com/langchain-ai/langgraph	用于使用LLMs构建有状态的多角色应用程序,创建代理和多代理工作流
RAGFoundry	https://github.com/ntelLabs/RAGFoundry	RAG Foundry是一个库,旨在通过微调模型来提高LLM使用外部信息的能力
RAG-GPT	https://github.com/gpt-open/rag-gpt	RAG-GPT从用户定制的知识库中学习,为各种查询提供上下文相关的答案,确保快速准确地检索信息
智谱 RAG	https://github.com/THUDM/ChatGLM3	ChatGLM3是智谱AI和清华大学KEG实验室联合发布的开源双语对话语言模型
SpRAG	https://github.com/SuperpoweredAI/spRAG	适用于非结构化数据的高性能检索引擎
TableGPT	https://github.com/ZJU-M3/TableGPT-techreport	统一表、自然语言和命令的微调GPT模型的报告
Dify	https://github.com/langgenius/dify	Dify是结合了AI工作流程、RAG管道、代理功能、模型管理、可观测性功能的LLMs应用程序开发平台
LLAMA_Index	https://github.com/run-llama/llama_index	一个用于构建上下文增强LLM应用程序的框架

神经网络模型,提高搜索的相关性和准确性。

Langchain-Chatchat 将 LLMs 和 Langchain 框架结合,旨在创建高效的检索增强生成管道。工作流程通过加载文件、文本分割、向量化处理及匹配最相似的文本片段来生成上下文丰富的回答。相比传统对话系统,LangChain-Chatchat的设计注重模块化和灵活性,允许用户根据具体需求选择嵌入模型、向量存储等组件,适用于多种场景,如对话、信息检索或数据库查询。针对专业领域,如法律咨询,LangChain-Chatchat可以快速定位相关法规并给出证据支持,极大地提高了工作效率和服务质量[108]。FastGPT 是为问答和对话任务设计的语言模型。通过大量预训练数据,FastGPT能理解和生成自然语言文本,优化QA结构以应对非典型问题。它结合向量模型增强表达能力,确保高效沟通。FastGPT提供可视化工作流,清晰展示问题到答案的路径,便于调试;API架构支持无缝集成至现有应用,无需修改源代码。调试工具包括搜索测试、引用修改及完整对话预览,提升其开发效率和用户体验。除了支持主流LLMs,FastGPT还计划引入自定义向量模型,展现出高度的灵活性,适用于客服问答、专利撰写等领域,推动行业智能化发展[109]。与此同时,网易QAnything通过整合多源数据并利用语义理解和检索技术来提供精准问答;MaxKB采用知识图谱和深度学习相结合的方式构建智能知识库,以支持复杂查询与推理;RAG-GPT则结合检索增强生成模型,从大规模文档中检索相关信息来提升语言模型的回答质量。上述框架均旨在通过先进的数据处理和深度学习技术提高信息检索的准确性和效率[3]。

RAGFlow 在需要利用外部知识或数据时,结合RAG技术增强LLMs的上下文来提升特定任务表现。它采用多路召回策略(如关键词搜索和嵌入向量搜索)获取相关信息,重排序后,并最终利用这些信息块增强LLMs的上下文,以生成更精准的回答。RAGFlow的核心优势在于深度文档理解能力,能够从复杂格式的非结构化数据(文本、图片和表格)中提取关键信息。此外,它还提供多种可控可解释的文本切片模板,如问答、简历、论文等,满足多应用场景需求。为降低幻觉,RAGFlow通过引用原文链接和内容快照增强答案的可信度,兼容多种文件类型,包括Word、PDF和图像文件,广泛适用于企业知识管理和个人学习研究[59]。SpRAG(特定领域检索增强生成)通过结合领域特定的知识库和大型语言模型,利用精准的检索技术来提高企业的信息获取效率[71]。

Dify是一个优化LLMs应用程序开发与部署的平台,涵盖数据处理(ETL)管理提示(Prompts IDE)、交互控制(Dify Agent DSL)、请求处理(BaaS)、缓存优化、组件协调、插件集成和内容审核。它支持多种LLMs和持久化存储。与LangChain等工具箱相比,Dify提供完整的生产级解决方案,支持私有化部署,为特定领域的聊天机器人及AI助理设计[110]。Open-webui是一个开源界面框架,提供灵活可定制的组件,简化Web应用开发,帮助开发者轻松创建交互性强、响应迅速的用户界面,无需深入复杂的前端技术[79]。Langraph通过将文本转换为图形结构,利用图神经网络捕捉表示文本中的语义依赖和关联,以更好地理解和生成复杂的语言关系。它在医学领域的图谱构造中表现突出[111],实现快速且准确的搜索引擎。此外,智谱RAG结合外部知识库来提升语言模型回答质量;TableGPT则专注于理解和生成表格数据,利用专门的模型处理结构化信息,提供精准的财务报表和市场数据问答,应用于金融领域。

2 RAG评估

RAG在大模型中的广泛应用推动RAG评估方法的研究进展。本章论述了统一的RAG评估框架,关注RAG基准的两个关键问题:评估目标是什么?如何评估?框架涉及评估目标、数据集和指标,旨在向读者深入理解RAG模型在多样化应用场景的表现。

2.1 评估目标

RAG评估目标需从三个关键步骤考虑:检索、生成和流程评估。表4总结了评估目标,并列出相关评估工作,区分评估目标的多维度,并提供了多个评估不同原始目标的基准和工具。

检索组件对于获取通知并生成相关信息至关重要。评价检索指标时,需要在给定查询的上下文中有效衡量检索到的文档精确率、召回率和相关度[119]等。检索组件构建两个成对关系,相关文档-查询之间用评估相关性,相关文档-候选文档之间用于评估准确性。相关性评估衡量检索文档与查询中所需要信息的匹配程度,反映了检索过程的精准性和特异性。准确率评估则依据检索文档相对于候选文档的准确性,衡量系统识别和评分相关文档的能力。

生成部分由LLMs驱动,根据检索到的内容产生连贯且符合语境的回应。评估目标在于评估生成内容对输入数据的相关性、忠诚度、正确性。相关性衡量生成的响应与初始查询在意图和内容的一致性程度,确保响应与查询主题相关,并满足查询的具体需求。忠诚度评估依据生成的响应,判断是否准确反映相关文档中的信息,并衡量生成内容与源文档之间的一致性。正确性与检索组件中的准确率类似,衡量生成响应相对于样本响应的准确率。作为基本真值,它检查响应在事实信息方面的正确,以及在查询上下文中的适当性。例如,在创造性内容生成或开放式问答评估任务中,引入了“正确”或“高质量”回应的标准可变性[115]。

检索和生成组件之间的相互作用表明,无法仅通过

表4 RAG评估基准及其侧重的评估目标

Table 4 RAG evaluation benchmarks and objectives

文献	时间	基准	工具	研究	评估目标																				
					上下文相关性	答案相关性	事实基础性	检索准确性	生成准确性	忠实度	余弦距离	检索质量	生成质量	噪声鲁棒性	反事实鲁棒性	信息整合	负拒绝	响应质量	创建读取	更新删除	幻觉	多样性	清晰度	延迟
TruEraRAGTriad[112]	2023.10		√		√	√	√						√											
LangChainBench[113]	2023.11		√		√			√		√	√	√		√				√						
DatabricksEval[114]	2023.12		√						√	√						√			√					
RAGAs[115]	2023.09	√			√	√				√		√				√			√			√		
RECALL[116]	2023.11	√						√	√	√								√	√	√				
ARES[117]	2023.11	√				√	√			√			√					√				√		
RGB[24]	2023.12	√							√					√		√					√	√	√	√
MultiHop-RAG[118]	2024.01	√						√	√								√			√	√			
CRUD-RAG[119]	2024.02	√														√	√	√						
MedRAG[120]	2024.02	√						√						√	√	√						√	√	
FeB4RAG[121]	2024.02	√							√				√										√	
CDQA[122]	2024.03	√				√	√	√	√							√					√			
DomainRAG[123]	2024.06	√							√	√														
ReEval[124]	2024.06	√																			√			
FiD-Light[125]	2023.07			√									√	√										
DiversityReranker[106]	2023.08			√	√	√						√		√	√									√
power_of_noise[126]	2024.01			√		√		√	√						√	√								
eRAG[127]	2024.06			√	√			√	√	√								√	√		√			
RAGEval[128]	2024.08	√				√		√	√								√	√						
RAGChecker[129]	2024.08	√					√	√	√						√	√							√	
R-Eval[130]	2024.08		√					√	√					√	√				√	√				

孤立地评估[131]来全面了解整体的组件性能。评估过程除了涉及检索信息提高响应质量的能力之外,还包括衡量检索组件对生成过程的其他条件。例如延迟、多样性、噪声鲁棒性、负面拒绝和反事实鲁棒性等,用于确保RAG在现实世界场景中的实际适用性,并与人类偏好保持一致。此外,响应延迟、处理模糊和复杂查询的能力等实际考虑,对于评估RAG的整体有效性和可用性[131-132]至关重要。

2.2 评估数据集

不同的基准采用不同的数据集构建策略,例如,可以利用现有资源或生成特定评估维度的全新数据。一些基准数据集借鉴知识密集型语言任务的评估标准,包括自然问题、HotpotQA[133]、FEVER[134]和其他已建立的数据集[135],主要用于问答任务(QA)。HotpotQA强调多跳推理,即模型需要跨越多个证据来源得出正确答案。SuperGLUE[131]使用MultiRC[136]用于多选择问答任务,而ReCoRD[137]数据集则用于评估常识推理的任务,测试模型从长文本中提取关键信息并回答相关问题。然而,这类数据集的使用局限是在解决动态现实场景中存有挑战。类似的情况可以从RAGAs[115]构建的2022年后的维基百科页面中观察到。

功能强大的LLMs的出现彻底改变了数据集的构建过程。使用大模型的框架能够为特定评估目标设计查询和背景真值的能力,作者现在可以轻松地创建所需格式的数据集。RGB、MultiHop-RAG、CRUD-RAG和CDQA[24,118-119,122]等基准测试集采用了这种方法,使用在线新闻文章构建自己的数据集,测试RAG对真实世界信息的感知能力。最近,DomainRAG[123]将多种类型的问答数据集与单文档、多文档单轮和多轮对话相结合,这些数据集是由高校招生网站每年变化的信息生成的,目的使LLMs提供和更新信息。RAGEval[128]通过收集特定领域的种子文档,构建问题-参考-答案三元组数据集,评估RAG模型的有效性。因此,数据集的创建和选择为特定指标或任务量身定制,提高了评估精度,并指导开发适应真实世界信息需求的RAG。评估数据集如表5所示。

表5 RAG评估数据集

Table 5 RAG evaluation dataset

基准	数据集
RAGAs[115]	维基百科WikiEval
RECALL[116]	EventKG[138]、UJ[139]
ARES[117]	Hotpot[133]、FEVER[134]、WoW[135]、MultiRC[136]、ReCoRD[137]、NQ[140]
RGB[24]	新闻生成数据集
MultiHop-RAG[118]	新闻生成数据集
CRUD-RAG[119]	新闻生成数据集UHGEval[141]
MedRAG[120]	MIRAGE
FeB4RAG[121]	FeB4RAG、BEIR[142]
CDQA [122]	新闻生成数据集、Labeller
DomainRAG[123]	高校录取信息生成数据集
ReEval[124]	RealTime QA[143]、NQ[144]
RAGEval[128]	DragonBall数据集
RAGChecker[129]	人工注释的偏好数据集

2.3 评估指标

本节将介绍几种常用的检索和生成的评估指标,评估RAG整体流程的指标,也可以在常用指标中找到。详细的指标简介通过表6作为参考进行探索。

2.3.1 检索指标

检索评估着重于捕捉检索信息在响应查询时的相关性、准确性、多样性和鲁棒性等指标。文献[123]的基准部署的误导率、错误重现率和错误检测率等指标,强调了对RAG内在复杂性的高度认识。具体来说,误导率衡量错误响应的比例,适用于高可信度的问答场景。错误重现率评估错误在后续查询的重复,错误检测率衡量模型纠错能力,适用于对话生成任务。此外,将MAP、MRR和标记化与F1集成到RAG基准[118,122],反映了对传统检索的多维度理解。MAP衡量平均查准率,适用于问答任务;MRR评估第一个相关项的排名,适用于实时性要求高的应用;标记化优化文本单位处理,提升大规模数据检索效果;F1平衡精度与召回率,适用于平衡两者的任务,如文档检索。Zheng等人[145]强调,基于排序的评估方法应该引入更多专门针对RAG的检索评价指标。它不仅反映查准率和查全率,而且考虑了检索的多样性与相关性,并符合RAG信息需求的复杂性和动态性。因此,引入LLMs作为评判者,进一步强调检索评价的适应性和多功能性[146],并将其应用到多任务评估。

(1)准确率(accuracy,ACC)

ACC用于衡量模型整体的预测正确性。如在生成问答时,ACC衡量检索到的文档与问题的相关性,以及生成的答案与标准答案的匹配度。由以下公式(1)所示:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

其中,TP(true positive)表示真阳性,即正确地被分为正类的样本数;TN(true negative)表示真阴性,即正确地被分为负类的样本数;FP(false positive)表示假阳性,即错误地被分为正类的样本数;FN(false negative)表示假阴性,即错误地被分为负类的样本数。

(2)精确率(precision,Prec)

Prec用于衡量模型检索出的相关信息的准确性,并在问答中反映检索回答中准确匹配用户问题的比例,精确率越高,检索性能越优。由以下公式(2)所示:

$$Preccision=\frac{TP}{TP+FP} \tag{2}$$

其中,TP表示真阳性,即正确地被检索或推荐的相关文档的数量;FP表示假阳性,即错误地被检索或推荐为相关的文档的数量。

表6 RAG评估指标

Table 6 RAG evaluation metrics

指标分类	简介
ACC	衡量检索结果的相关性
Prec	衡量相关实例在检索实例中的比例
EM	评价预测中匹配到正确答案的百分比
HR	相关上下文在答案中的命中率
Recall@ K	在只考虑前 K 个结果时,检索到的相关实例占相关实例总数的百分比
MRR	一组查询中的第一个正确答案的倒数排名的平均值
MAP	每个查询的平均查准率得分的平均值
NDCG	衡量文档的相关性和它们在排名中的位置
FN@ K	在某个截断点 K 时,错误排名的文档数量与前 K 个文档总数的比例
ACC@ K	前 K 个检索结果之间的相关性
ROUGE-N/ROUGE-L	摘要与人工生成的参考摘要比较来评估摘要质量的度量
BLEU	计算生成文本相对参考文本的 n 元词串精度,可以用来双语评估文本质量
Bert Score	使用上下文嵌入计算令牌级别的相似度,并生成精确率、召回率和F1分数
METEOR	同义词和句法结构,是对BLEU的改进,更能反映人类翻译质量评价
Perplexity	评估语言模型性能的一个指标,衡量模型预测样本的能力
F1	准确率和召回率的调和平均,是一个综合指标
BERTScore	利用预训练的BERT模型的内部表示来计算生成文本与参考文本之间的相似度
余弦相似度	计算检索到的文档或生成的响应的嵌入
执行时间	评估完成一次查询响应所花费的时间
BLEURT	一个学习型的评估指标,可以训练评估文本的质量
CFR	衡量模型在面对反事实变化时稳定性和一致性的指标

(3)召回率@ K(Recall@K,R@K)

R@ K 用于多步推理任务,评估每步检索的效果,确保每步能有效地缩小搜索范围并接近最终答案。召回率 @K 的值越高,说明在前 K 个检索结果中找到的相关文档的比例越高,RAG检索效果越好。由以下公式(3)所示:

$$Recall@K = \frac{RD \cap Top_{kd}}{|RD|} \tag{3}$$

其中,在RAG的检索任务中, RD 表示给定的查询相关文档的集合; Top_{kd} 表示模型检索出的前 K 个文档的集合。

(4)平均倒数排名(mean reciprocal rank,MRR)

MRR用于计算所有查询的平均倒数排名,其中排名是通过RAG返回的相关项列表的顺序来确定的。如果相关项出现在列表前,倒数排名会更高,MRR越高,表示RAG检索的性能越好。如案例检索时,MRR可以评估重要判例的排名。由以下公式(4)所示:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{4}$$

其中, $|Q|$ 表示查询集合的数量,即有多少个查询需要被评估; $rank_i$ 表示为第 i 个查询的第一个相关项在列表中的排名。

(5)平均查准率(mean average precision,MAP)

MAP用于计算所有查询的平均精度,其中精度是指在相关性后,检索结果的排名质量。MAP值越高,表示RAG检索的性能越好。该指标特别适用于要求高质量排名的任务,评估查询高度相关的句子排在前列。由以下公式(5)所示:

$$MAP = \frac{1}{|Q|}\sum_{q=1}^{|Q|}\frac{\sum_{k=1}^{n}(P(k)\times rel(k))}{|RD_q|} \tag{5}$$

其中, $|Q|$ 表示查询集合数, q 表示单个查询索引, n 表示对于单个查询 q 返回的检索结果的数量, RD_q 表示对于查询 q ,即所有相关文档的总数; $P(k)$ 表示在排名为 k 的检索结果的查准率(Prec), $rel(k)$ 表示为指示函数,当排名为 k 的检索结果是相关的时候,它的值为1,否则为0。

(6)精确匹配(exact match,EM)

EM用于衡量模型输出与标准答案之间的一致性,是RAG一种严格的评估标准。由以下公式(6)所示:

$$EM = \begin{cases} 1 \\ 0 \end{cases} \tag{6}$$

其中,一个开放域问题回答任务,当用户提问时,模型输出判断是否完全匹配标准答案。如果生成答案与人工标注答案匹配,则EM得分为1,否则得分为0。

(7)排名质量(normalized discounted cumulative gain,NDCG)

NDCG用于衡量排名中每个文档的相关性评分,并根据文档的位置进行折扣,以奖励高相关性文档在更靠前位置。NDCG的值范围从0到1,值越高表示排序越接近理想状态。在搜索引擎中,NDCG帮助确保高相关性网页出现在搜索结果前面,从而提高用户的检索体验。由以下公式(7)所示:

$$NDCG_k=\frac{DCG_k}{IDCG_K} \tag{7}$$

其中,DCG表示计算实际排名中每个文档的相关性评分。IDCG表示计算理想情况下每个文档的相关性评分。位置折扣后,确定在理想排序下可以获得的最大DCG。

(8)错误发现率(false negatives at K,FN@K)

在文档检索中,FN@K用于衡量前K个检索结果中遗漏的相关文档的数量。具体来说,当用户输入查询词或问题时,检索系统从文档库中返回前K个最相关的文档。FN@K计算模型未能检索到的相关文档的数量,从而评估检索系统的召回能力和对用户需求的满足程度。FN@K越高,表示检索遗漏的相关文档越多,性能越差。由以下公式(8)所示:

$$FN@K=\frac{TRD-RDTOP\cdot K}{K} \tag{8}$$

其中,TRD表示查询相关的总文档数,$RDTOP\cdot K$表示在前K个检索结果中检索到的相关文档的数量。

2.3.2 生成指标

评估不仅注重生成的准确性,同时也倾向于重视文本的连贯性、相关性,以及与人类评估标准的一致性等方面,全面考察文本的整体质量。因此,要求采用细致入微的评估指标来衡量语言生成的质量。指标需要涵盖摘要类等下游任务的事实准确性,并包括文本的可读性及用户对生成内容的满意度。

(1)概要评估指标(recall-oriented understudy for gisting evaluation,ROUGE-N)

ROUGE-N属于ROUGE家族的指标。用于评估自动文摘或机器翻译输出的质量指标,它通过比较生成摘要与参考摘要之间的相似度来工作。N表示在计算相似度时考虑的连续单词的最大数目。ROUGE-N包括ROUGE-N Precision、ROUGE-N Recall和F1分数等多个指标。

① 词精确率(ROUGE-N precision,ROUGE-NP)衡量系统生成的摘要有多少比例连续的N(n-gram),与参考摘要的匹配。例如,$n=1$代表单个词匹配,$n=2$代表二元组匹配。该指标适用于新闻摘要任务,通过确保生成的摘要尽可能包含参考摘要中的关键信息,以提高摘要的质量。由以下公式(9)所示:

$$ROUGE\text{-}NP=\frac{\sum_{n\text{-}gram\in Refer}Count_{\text{match}}(n\text{-}gram)}{\sum_{n\text{-}gram\in Output}Count(n\text{-}gram)} \tag{9}$$

其中,$Count_{\text{match}}(n\text{-}gram)$表示系统生成的摘要与参考摘要的匹配数量,$Count(n\text{-}gram)$表示系统生成的摘要中$n\text{-}gram$的数量。

② 词召回率(ROUGE-N recall,ROUGE-NR)表示量化参考摘要连续的N个词有多少比例也出现在生成摘要。对于翻译任务,ROUGE-N Recall用来评估机器翻译输出与人类翻译的参考文本之间的相似度。由以下公式(10)所示:

$$ROUGE\text{-}NR=\frac{\sum_{s\in ReferSum}\sum_{gram_n\in S}Count_{\text{match}}(gram_n)}{\sum_{s\in\{ReferSum\}}\sum_{gram_n\in S}Count(gram_n)} \tag{10}$$

其中,$\sum_{s\in ReferSum}$表示所有参考摘要求和,$gram_n\in S$表示摘要中提取的n-gram,其中S是摘要集合。$Count_{\text{match}}(gram_n)$表示生成摘要与参考摘要相匹配的$n$-gram的数量。$Count(gram_n)$表示参考摘要中$n$-gram的总数。

③ F1是ROUGE-N Precision和ROUGE-N Recall的调和平均数,平衡查准率和查全率之间的关系,用于评估自动文摘或机器翻译任务输出的质量。由以下公式(11)所示:

$$ROUGE\text{-}N=F_\beta(P,R)=(1+\beta^2)\cdot\frac{Prec\cdot R}{(\beta^2\cdot Prec)+R} \tag{11}$$

其中,$Prec$表示词查准率,即生成摘要中出现在参考摘要的比例,R表示词查全率,即参考摘要相似于系统生成摘要的比例,β表示为一个调整$Prec$和R的重要参数。

(2)最长公共子序列评价指标(ROUGE-longest common subsequence,ROUGE-L)

ROUGE-L用于衡量摘要中词语顺序和匹配程度。该指标用于评估生成任务,帮助评估生成的答案是否准确地保留输入问题相关的核心信息,并保持语序的合理性。由以下公式(12)所示:

$$ROUGE\text{-}L=F_{1\beta}(P_{\text{L}},R_{\text{L}})=(1+\beta^2)\cdot\frac{P_{\text{L}}\cdot R_{\text{L}}}{(\beta^2\cdot P_{\text{L}})+R_{\text{L}}} \tag{12}$$

其中,$F_{1\beta}$表示ROUGE-L的F1分数,P_{L}表示ROUGE-L Precision,R_{L}表示ROUGE-L Recall。

(3)双语评估指标(bilingual evaluation understudy,BLEU)

BLEU用于衡量机器翻译输出与人类翻译的相似度,针对双语评估研究中[147]一个或多个参考译文来评估机器翻译文本质量。例如,在多轮问答任务中,通过计算n-gram的重叠度,BLEU能够反映生成答案与参考答案之间的语义和表达一致性。由以下公式(13)所示:

$$BLEU=BP\cdot\exp\left(\sum_{n=1}^{N}w_n\log_a P_n\right) \tag{13}$$

其中,$\log_a P_n$表示对数概率,即机器翻译输出中,第n个n-gram与参考翻译中相同n-gram的匹配概率对数,W_n表示为权重系数,即用于为不同长度n-gram设置不同权重,N表示最大n-gram长度,BP表示为惩罚因子,即用于惩罚候选句子长度过短输出。

(4)LLMs评估

LLMs作为一种评估性判断工具,提供了一种通用且自动化的质量评估方法,适用于非传统的事实真值衡

量的情况。LLMs能够依据连贯性、相关性和流畅性等标准,对生成的文本进行评分[117]。LLMs可以通过微调以适应人类的判断标准,从而预测未见过文本的质量,或者在零样本或少样本的情况下生成评价。该方法的优势在于它结合预测驱动推理(pretrained predictive intercoder,PPI)和上下文相关性评分,从而有效评估LLMs输出。正如Saad-Falcon等人[117]阐述,通过策略性地使用提示模板,可以确保与人类偏好保持一致,同时有效地规范不同内容维度的评估。Dong等人[112]利用LLMs进行评估,标志着自动化和上下文响应的评估框架取得显著进展,减少对传统参考的比较依赖。此外,误导率、错误再现率和错误检测率等指标的出现,突显对RAG面临独特挑战的深化理解。

2.3.3 额外指标

RAG的流程评估需要考虑一系列额外的要求,如延迟、多样性和噪声鲁棒性等,以确保其在真实世界场景中的实用性和符合人类偏好。

(1)延迟

延迟用于衡量RAG完成一次查询响应所花费时间的指标。它是影响用户体验的关键因素,特别是在聊天机器人或搜索引擎等需要即时反馈的交互式应用中[118]。在应用RAG进行大模型推理时,生成延迟过高可能导致用户流失,因此需要优化查询处理速度以确保实时响应。

(2)余弦相似度

多样性的评估目标用于衡量RAG检索和生成信息的种类和广度,确保系统能够提供多样化的视角,避免响应冗余。RAG通过计算不同检索结果或生成文本之间的余弦相似度评估多样性。较低的余弦相似度分数表示更高的多样性,表明RAG能够从广泛的信息源中提供多样化的响应[121]。由以下公式(14)所示:

$$Cosine\ Similaity=\frac{A\cdot B}{\|A\|\|B\|} \tag{14}$$

其中,A和B是检索或生成文本的向量表示。

(3)噪声鲁棒性

噪声鲁棒性用于衡量RAG处理不相关或误导信息时,维持响应质量的能力[116]。噪声鲁棒性确保系统在面对不完美或有噪声数据时,仍能提供准确和有用的回答。在医疗问诊时,用户可能输入模糊或拼写错误的症状描述,系统需要在这种情况下仍能返回准确且相关的诊断建议。

(4)负拒绝(negative rejection,NR)

负拒绝是Chen等人[24]提到的假阴性指标,评估RAG在可获得的信息不足或模棱两可的情况时,选择不提供答案的能力。它指RAG拒绝产生响应的速率。如应用在法律评估基准时[148]可以凸显模型在高可信度场景下的稳健性和可靠性。

(5)反事实鲁棒性(counterfactual robustness,CFR)

反事实鲁棒性评估RAG识别和忽略检索文档中不正确或反事实信息的能力。在信息检索与问答任务中,如果RAG在应对不准确或反事实信息时,生成的内容准确反映查询的实际需求,具有较低的错误响应比例,则表明RAG具备较高的反事实鲁棒性。它能更好地衡量模型输出的可信度。

2.4 评估

为了能够精确评估目标,RAG中的每个组成部分都需要一种定制化的评估,以反映其不同的功能和目标。许多研究工作采用了综合评估方法,同时考虑了多个方面。eRAG[127]通过评估每个检索文档在大模型推理的下游性能,以此作为相关性标签,并利用多种任务指标聚合分析,优化检索评估方法。FeB4RAG[121]在文献[145]的基础上,提出一致性、正确性、清晰度和覆盖程度4个标准。其中,正确性相当于检索准确性,而一致性与生成组件的忠实性对应。检索准确性不仅评估了检索信息的正确性,而且它与覆盖率和多样性紧密相关。因此,通过覆盖率与多样性结合,同时考虑检索过程和生成过程的额外优化标准,全面提升系统性能。其他工具和基准程序也做了类似的综合处理方法。工具提供了一个通用框架,文献[112,149]构建完整的RAG应用程序和评估管道以高效评估。R-Eval[130]是一个可扩展Python工具包以评估RAG工作流程。它通过自定义测试数据揭示不同任务和领域LLMs的差异,指导了选择合适的RAG和LLM组合。而基准测试则侧重于RAG评估的不同方面,特别强调检索输出或生成目标。RAGAs[115]和ARES[117]专注于评估检索文档的相关性,而RGB[24]则优先考虑准确性。Yu等人[124]关注的幻觉问题实际上是忠实性和正确性的结合。RAGEval[128]通过生成高质量样本并引入完整性、幻觉性和不相关性三项指标,显著提升RAG系统评估的全面性和与人工评估的一致性。细粒度评估框架RAGChecker[129]提供一套全面的指标和工具,揭示RAG架构设计中有见地的模式与权衡点,帮助研究人员和实践者优化和改进现有RAG。上述这些多维度的评估方法有助于全面理解RAG的性能,确保它们在实际应用中能够满足用户的需求和期望。

3 挑战与未来展望

3.1 LLMS微调结合RAG

RAG与微调相结合正在成为领先的策略之一。确定RAG和微调的最佳集成方式,研究者们正致力于探索如何通过顺序式、交替式,或端到端联合训练,并高效利用两者的参数化。与此同时,非参数化方法在优化检索效率和减轻计算负担方面显示出了优势[92],尤其结合深度学习模型时,能够开辟新的研究路径。另一个趋势

是将具有特定功能的SLM引入RAG框架,并根据RAG的检索结果进行微调。例如,RAG[65]训练了一个轻量级的检索评估器来评估查询检索文档的整体质量,并基于置信度触发不同的知识检索动作。这一方向启示了在推理任务中引入专门知识模块,以优化推理路径和知识检索。

未来研究可以探索不同集成方式的平衡,特别是在复杂任务中通过端到端训练提升微调与RAG的结合效果。同时,结合非参数化方法与深度学习,以及多领域应用(如医疗、金融、法律等)中的策略互补,也是值得关注的研究方向。

3.2 长文本与RAG

在处理长文本时,RAG一次性为LLMs提供大量上下文会显著影响其推理速度,因此采用分块检索和按需输入显得尤为重要,这可以有效提高计算效率。另一方面,基于RAG的生成能够快速定位LLMs的原始引用,帮助用户验证生成结果,并保持整个推理过程的可追踪性。与此相比,依赖长上下文的生成依旧存在黑盒问题,缺乏透明性。扩展上下文为RAG的发展提供了新机遇,使其能够应对更复杂的问题,尤其是在需要阅读大量材料回答的综合性或总结性问题上。因此,在超长上下文环境下开发新的RAG方法,已成为未来研究的重要趋势。

最近的研究表明了RAG和LCLMs的联合任务测试,例如“Needle-in-a-Haystack”任务[150],要求从大文档中提取关键信息。尽管该任务的难度尚未达到最新一代大模型所能完美处理的复杂性,但一些最先进的模型已接近完美性能。因此,如何评估这些系统并确保它们在实际应用中的可靠性和可扩展性,仍是一个开放性挑战。

未来的研究应聚焦于如何在长文本处理中实现更高效的分块检索、更精准的上下文整合策略,并探索如何在多领域任务中优化RAG与LCLMs的联合应用。

3.3 RAG鲁棒性

RAG检索时,噪声或矛盾信息的存在会影响RAG的输出质量。这种情况被描述为“错误信息可能比没有信息更糟糕”。因此,提高RAG的鲁棒性获得关注,并逐渐成为关键的性能指标。Cuconasu等人[126]分析了检索文档的类型,评估了文档与提示的相关性、文档在上下文的位置以及其数量。研究结果表明,意外地引入不相关的文档可能反而提高准确性超过30%,这一发现挑战了传统认为不相关文档会降低质量的假设。结果强调了在RAG中开发专门策略,以有效地将检索和生成模型结合的重要性,并突显加强RAG鲁棒性研究的必要性。

未来的研究应关注改进检索策略、增强噪声处理能力及优化文档排序算法,同时探索深度学习与强化学习的结合的潜力,特别在多模态数据和复杂任务中。此外,研究如何通过反馈机制动态调整检索文档质量,以应对信息不完全或模糊的情境,也是一个关键方向。

3.4 知识图谱与RAG的融合:Graph RAG

在检索过程中,LLMs常常因为通用语义的偏差而引入“幻觉”。即指生成的内容缺乏意义或者不忠实于原始的信息源。GraphRAG[151]提供了一种有效的解决方案,通过引入图形索引和丰富的文本注释来减轻这一问题。未来的研究可以进一步改进GraphRAG方法,探索更本地化的RAG操作方式,例如基于嵌入的图形注释匹配来提高准确性。此外,混合型的RAG方案也是一个值得关注的方向,这类方案在映射-归约摘要机制应用之前,结合基于嵌入的匹配来处理社区报告,从而优化信息提取与总结过程。这种“汇总”操作不仅可以扩展到更高层次的社区结构,还能作为一种“深入挖掘”机制,沿着高层次社区摘要中的信息线索进行追踪。未来应研究如何在知识图谱和RAG框架下,提升LLMs在长尾知识处理、跨领域任务中的推理能力。

3.5 多模态应用:Multimodal RAG

RAG已经超越最初基于文本的问答限制,扩展至多模态数据,催生了创新的多模态模型,并将RAG概念集成到各个领域。例如,在图像领域,RA-CM3[152]作为RAG的开创性多模态模型,开启了新篇章。BLIP-2[153]利用冻结图像编码器和LLMs进行有效的视觉语言预训练,实现了零镜头图像到文本的转换。此外,“在你写之前可视化”的方法[154],通过图像生成引导LLMs的文本生成,在开放式文本生成任务中展现出前景。在音频和视频领域,GSS方法检索并缝合音频片段,将机器翻译的数据转换为语音翻译的数据[155]。UEOP通过结合外部离线策略进行语音到文本转换,标志着端到端自动语音识别技术的重大进步[156]。基于KNN的注意力融合通过音频嵌入和语义相关的文本嵌入来改进LLMs,从而加速领域自适应。Vid2Seq通过引入专门的时间标记增强语言模型,促进了统一输出序列中事件边界和文本描述的预测[157]。在代码领域,RBPS[158]通过编码和频率分析检索与开发者目标一致的代码示例,在小规模学习任务中表现出色。这种方法已经在测试断言生成和程序修复等任务中证明了有效性。对于结构化知识,CoK方法[91]首先从知识图中提取与输入查询相关的事实,然后将这些事实作为提示集成到输入中,从而提高了知识图问答任务中的性能。

未来的研究应优化不同模态间的融合,提高图像、文本和音频等模态的互操作性和数据对齐能力。同时,领域适应、个性化推荐和实时处理能力将成为重要方向。此外,端到端训练、计算效率和可解释性将是多模态RAG发展的关键。

3.6 RAG与Agent集成:AI Agent

大语言模型如今已落地应用,AI Agent成为发展趋

势之一。AI Agent不仅具备大模型强大的语义理解与推理能力,还具备任务规划能力,并调用外部工具来执行任务。RAG视为Agent的一种简化形式,尤其是在知识库作为检索工具的应用方面。此外,当前也看到RAG对记忆和规划能力的集成诉求(例如RAT[88]、RoG等),这一趋势正逐步走向模块化RAG。同时,Agent自身所需的长期记忆存储也将反向依赖RAG的知识库,这种互补关系将推动二者的共同发展。一个典型应用中[110]描述了模拟人类行为的多Agent协同工作模型。该模型利用自然语言存储Agent的经历记录,在一个交互沙箱环境中生成可信的个体和群体行为。为了实现生成Agent,论文描述了一种类似于扩展型RAG的外部架构,该架构通过自然语言存储的Agent的完整经历记录,并在动态检索这些记忆的基础上进行高层次的反思,从而模拟人类的行为计划。

为应对这些挑战,未来应进一步研究如何优化RAG与Agent的集成,提升记忆存储的可扩展性、查询效率以及如何加强记忆与推理能力之间的协同作用。

3.7 Agent赋能:促进LLMs推理能力

尽管LLMs在执行复杂推理任务时已经取得一定进展,但仍面临半幻觉、推理视角错误和信息不一致等缺陷,限制其在高阶推理能力。为解决这些问题,Wang等人[88]提出将RAG与思维链结合的方法,利用外部知识库修正LLMs的推理路径并增强信息一致性。这一方法既修正了推理过程中出现的错误,还在实时推理中动态地引入更多信息。基于此,Debate[159]、MAD[160]和reconciliation[161]等多AI智能体协作框架为LLMs带来新方案。这些框架通过模拟人类讨论过程和多轮交互来增强LLMs的推理能力,避免常见的推理错误。多智能体框架利用强提示帮助代理保持对上下文的敏感性,确保推理路径的准确性,优于传统的Chain-of-Thought[162]方法。此外,任务特定示例(如演示)帮助代理在多轮讨论中调整推理路径,特别在复杂任务如交互问答中,逐步优化推理深度和广度。结合现有的RAG框架,这些多智能体机制通过动态检索和多轮推理提升了推理准确性和鲁棒性。在RAG支持下,智能体能够迅速切换和交互不同信息源,确保推理过程中的一致性。

未来应进一步探索如何在RAG框架中结合多智能体讨论机制,优化推理路径,特别是在需要反复推敲和验证的复杂任务(如法律推理和医学诊断)中。随着这些技术的发展,LLMs在实际应用中的表现将得到显著提升,从而为更广泛的智能任务提供支持。

4 结语

本文强调了RAG通过将语言模型中的参数化知识与外部知识库中的大量非参数化数据相结合,在提升LLMs能力方面取得的重要进展。本研究回顾与分析了RAG技术进展及其在许多不同任务上的应用。尽管RAG技术取得了显著进展,但仍有研究机会提高其鲁棒性和处理扩展上下文的能力。RAG的应用范围正在扩展到多模态领域,未来将继续扩展至医学、工业、教育等实际应用场景,为大型语言模型带来更广泛的适用性和实际效益。这一应用突出了RAG对人工智能部署的重要实践意义,吸引了学术界和工业界的兴趣。随着RAG应用范围的扩大,确保准确且具有代表性的性能评估,将对全面展示RAG在人工智能研究与开发中的贡献至关重要。

参考文献:

[1] QIAN J L, JIN Z Y, ZHANG Q, et al. A liver cancer question-answering system based on next-generation intelligence and the large model med-PaLM 2[J]. International Journal of Computer Science and Information Technology, 2024, 2(1): 28-35.

[2] YUE S B, CHEN W, WANG S Y, et al. DISC-LawLLM: fine-tuning large language models for intelligent legal services[J]. arXiv:2309.11325, 2023.

[3] 房晓楠. 松鼠AI的“AI+智适应教育”之路该如何走?[J]. 机器人产业, 2019(1): 80-84.
FANG X N. How should squirrel AI take the road of “AI+intellectual adaptation education”?[J]. Robot Industry, 2019(1): 80-84.

[4] NIKDAN M, TABESH S, CMCEVIC E, et al. RoSA: accurate parameter-efficient fine-tuning via robust adaptation[C]// Proceedings of the 41st International Conference on Machine Learning, 2024.

[5] LI J T, LIU Y Q, FAN W Q, et al. Empowering molecule discovery for molecule-caption translation with large language models: a ChatGPT perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 6071-6083.

[6] SHUSTER K, POFF S, CHEN M Y, et al. Retrieval augmentation reduces hallucination in conversation[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg: ACL, 2021: 3784-3803.

[7] ZHAO T, WALLACE E, FENG S, et al. Calibrate before use: improving few-shot performance of language models[C]//Proceedings of the International Conference on Machine Learning, 2021: 12697-12706.

[8] CHENG X, LUO D, CHEN X, et al. Lift yourself up: retrieval-augmented text generation with self-memory[C]//Advances in Neural Information Processing Systems, 2024.

[9] GAO L, MADAAN A, ZHOU S, et al. PAL: program-aided language models[C]//Proceedings of the International Conference on Machine Learning , 2023: 10764-10799.

[10] HUANG L, YU W J, MA W T, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[11] IZACARD G, LEWIS P, LOMELI M, et al. Atlas: few-shot learning with retrieval augmented language models[J]. Journal of Machine Learning Research, 2023, 24(251): 1-43.
[12] WU Y, RABE M N, HUTCHINS D L, et al. Memorizing transformers[C]//Proceedings of the International Conference on Learning Representations, 2022.
[13] GUU K, LEE K, TUNG Z, et al. REALM: retrieval-augmented language model pre-training[C]//Proceedings of the International Conference on Machine Learning, 2020: 3929-3938.
[14] LEWIS P, PERES E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[C]//Advances in Neural Information Processing Systems, 2020: 9459-9474.
[15] BORGEAUD S, MENSCH A, HOFFMANN J, et al. Improving language models by retrieving from trillions of tokens[C]//Proceedings of the International Conference on Machine Learning, 2022: 2206-2240.
[16] IZACARD G, GRAVE E. Leveraging passage retrieval with generative models for open domain question answering[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Stroudsburg: ACL, 2021: 874-880.
[17] KHANDELWAL U, LEVY O, JURAFSKY D, et al. Generalization through memorization: nearest neighbor language models[C]//Proceedings of the International Conference on Learning Representations, 2019.
[18] HE J X, NEUBIG G, BERG-KIRKPATRICK T. Efficient nearest neighbor language models[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2021: 5703-5714.
[19] HE Z Y, ZHONG Z X, CAI T L, et al. REST: retrieval-based speculative decoding[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2024: 1582-1595.
[20] BANG F. GPTCache: an open-source semantic cache for LLM applications enabling faster answers and cost savings [C]//Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software, 2023: 212-218.
[21] ZHAO P H, ZHANG H L, YU Q H, et al. Retrieval-augmented generation for AI-generated content: a survey[J]. arXiv:2402.19473, 2009.
[22] GAO Y F, XIONG Y, GAO X Y, et al. Retrieval-augmented generation for large language models: a survey[J]. arXiv:2312.10997, 2023.
[23] FAN W Q, DING Y J, NING L B, 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. New York: ACM, 2024: 6491-6501.
[24] CHEN J W, LIN H Y, HAN X P, et al. Benchmarking large language models in retrieval-augmented generation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16): 17754-17762.
[25] ILIN I. Advanced RAG techniques: an illustrated overview[EB/OL]. [2024-09-25]. https://github.com/NirDiamant/RAG_TECHNIQUES.
[26] ZHENG H S, MISHRA S, CHEN X, et al. Take a step back: evoking reasoning via abstraction in large language models [C]//Proceedings of the 12th International Conference on Learning Representations, 2024.
[27] WANG S H, XU Y C, FANG Y W, et al. Training data is more valuable than you think: a simple and effective method by retrieving from training data[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 3170-3179.
[28] MA X B, GONG Y Y, HE P C, et al. Query rewriting in retrieval-augmented large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2023: 5303-5315.
[29] KHATTAB O, SANTHANAM K, LI X L, et al. Demonstrate-search-predict: composing retrieval and language models for knowledge-intensive NLP[J]. arXiv:2212.14024, 2022.
[30] WANG Y L, LI P, SUN M S, et al. Self-knowledge guided retrieval augmentation for large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.
[31] JEONG S, BAEK J, CHO S, et al. Adaptive-RAG: learning to adapt retrieval-augmented large language models through question complexity[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2024: 7036-7050.
[32] CHEN T, WANG H W, CHEN S H, et al. Dense X retrieval: what retrieval granularity should we use?[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2024: 15159-15177.
[33] ZHA L Y, ZHOU J L, LI L Y, et al. TableGPT: towards unifying tables, nature language and commands into one GPT [J]. arXiv:2307.08674, 2023.
[34] GAUR M, GUNARATNA K, SRINIVASAN V, et al. ISEEQ: information seeking question generation using dynamic meta-information retrieval and knowledge graphs[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10672-10680.
[35] YANG L Y, CHEN H Y, LI Z, et al. Give us the facts: enhancing large language models with knowledge graphs for fact-aware language modeling[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3091-3110.
[36] HE X, TIAN Y, SUN Y, et al. G-Retriever: retrieval-augmented generation for textual graph understanding and question

answering[J]. arXiv:2402.07630, 2024.

[37] TEJA R. Evaluating the ideal chunk size for a RAG system using LlamaIndex[EB/OL]. [2024-10-01]. https://www.llamaindex.ai/blog/evaluating-the-ideal-chunk-size-for-a-ragsystem-using-llamaindex-6207e5d3fec5.

[38] YANG S. Advanced RAG 01: small-tobig retrieval[EB/OL]. (2023-11-05)[2024-10-01]. https://towardsdatascience.com/advanced-rag-01-small-to-big-retrieval-172181b396d4.

[39] QIAN H J, LIU Z, MAO K L, et al. Grounding language model with chunking-free in-context retrieval[J]. arXiv:2402.09760, 2024.

[40] ZHAO J H, JI Z Y, FENG Y C, et al. Meta-chunking: learning efficient text segmentation via logical perception[J]. arXiv:2410.12788, 2024.

[41] LIANG Y, JIANG Z X, YIN D, et al. RAAT: relation-augmented attention transformer for relation modeling in document-level event extraction[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2022: 4985-4997.

[42] SUN Z, WANG X, TAY Y, et al. Recitation-augmented language models[J]. arXiv:2210.01296, 2022.

[43] WANG K X, REIMERS N, GUREVYCH I. DAPR: a benchmark on document-aware passage retrieval[J]. arxiv:2305.13915, 2023.

[44] KIM J, NAM J, MO S, et al. SuRe: summarizing retrievals using answer candidates for open-domain QA of LLMs[J]. arXiv:2404.13081, 2024.

[45] DOOSTMOHAMMADI E, NORLUND T, KUHLMANN M, et al. Surface-based retrieval reduces perplexity of retrieval-augmented language models[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 521-529.

[46] XIAO S T, LIU Z, ZHANG P T, et al. C-pack: packed resources for general Chinese embeddings[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2024: 641-649.

[47] LIU Z, XIAO S T, SHAO Y X, et al. RetroMAE-2: duplex masked auto-encoder for pre-training retrieval-oriented language models[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 2635-2648.

[48] SHI W J, MIN S, YASUNAGA M, et al. REPLUG: retrieval-augmented black-box language models[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2024: 8371-8384.

[49] ZHANG L X, YU Y, WANG K, et al. ARL2: aligning retrievers for black-box large language models via self-guided adaptive relevance labeling[J]. arXiv:2402.13542, 2024.

[50] DAI Z, ZHAO V Y, MA J, et al. Promptagator: few-shot dense retrieval from 8 examples[J]. arXiv:2209.11755, 2022.

[51] LUO K, LIU Z, XIAO S T, et al. BGE landmark embedding: a chunking-free embedding method for retrieval augmented long-context large language models[J]. arXiv:2402.11573, 2024.

[52] LI X M, LI J. AnglE-optimized text embeddings[J]. arXiv:2309.12871, 2023.

[53] YOON S, CHOI E, KIM J, et al. ListT5: listwise reranking with fusion-in-decoder improves zero-shot retrieval[J]. arXiv:2402.15838, 2024.

[54] GAO L Y, MA X G, LIN J, et al. Precise zero-shot dense retrieval without relevance labels[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 1762-1777.

[55] SARTHI P, ABDULLAH S, TULI A, et al. RAPTOR: recursive abstractive processing for tree-organized retrieval[J]. arXiv:2401.18059, 2024.

[56] WANG Y, LIPKA N, ROSSI R A, et al. Knowledge graph prompting for multi-document question answering[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(17): 19206-19214.

[57] RACKAUCKAS Z. RAG-Fusion: a new take on retrieval augmented generation[J]. International Journal on Natural Language Computing, 2024, 13(1): 37-47.

[58] ZHOU D, SVHARLI N, HOU L, et al. Least-to-most prompting enables complex reasoning in large language models[J]. arXiv:2205.10625, 2022.

[59] PENG W J, LI G Y, JIANG Y, et al. Large language model based long-tail query rewriting in Taobao search[C]//Companion Proceedings of the ACM Web Conference 2024. New York: ACM, 2024: 20-28.

[60] LI X, NIE E, LIANG S. From classification to generation: insights into crosslingual retrieval augmented ICL[J]. arXiv:2311.06595, 2023.

[61] KARPUKHIN V, OGUZ B, MIN S, et al. Dense passage retrieval for open-domain question answering[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020: 6769-6781.

[62] CHENG D X, HUANG S H, BI J Y, et al. UPRISE: universal prompt retrieval for improving zero-shot evaluation[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2023: 12318-12337.

[63] YOON J, CHEN Y F, ARIK S, et al. Search-adaptor: embedding customization for information retrieval[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 12230-12247.

[64] 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]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2023: 5364-5375.

[65] YAN S Q, GU J C, ZHU Y, et al. Corrective retrieval augmented generation[J]. arXiv:2401.15884, 2024.

[66] YU W, ITER D, WANG S, et al. Generate rather than retrieve: large language models are strong context generators[J]. arXiv: 2209.10063, 2022.

[67] LUO Z Y, XU C, ZHAO P, et al. Augmented large language models with parametric knowledge guiding[J]. arXiv:2305.04757, 2023.

[68] MA Y B, CAO Y X, HONG Y, et al. Large language model is not a good few-shot information extractor, but a good reranker for hard samples![C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: ACL, 2023: 10572-10601.

[69] DONG J L, FATEMI B, PEROZZI B, et al. Don't forget to connect! improving RAG with graph-based reranking[J]. arXiv:2405.18414, 2024.

[70] YU Y, PING W, LIU Z, et al. RankRAG: unifying context ranking with retrieval-augmented generation in LLMs[J]. arXiv:2407.02485, 2024.

[71] ANDERSON N, WILSON C, RICHARDSON S D. Lingua: addressing scenarios for live interpretation and automatic dubbing[C]//Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track), 2022: 202-209.

[72] JIANG H Q, WU Q H, LUO X F, et al. LongLLMLingua: accelerating and enhancing LLMs in long context scenarios via prompt compression[J]. arXiv:2310.06839, 2023.

[73] WANG Z R, ARAKI J, JIANG Z B, et al. Learning to filter context for retrieval-augmented generation[J]. arXiv:2311.08377, 2023.

[74] XU F Y, SHI W J, CHOI E. RECOMP: improving retrieval-augmented LMs with compression and selective augmentation[J]. arXiv:2310.04408, 2023.

[75] KIM Y, KIM H J, PARK C, et al. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg: ACL, 2024: 2421-2431.

[76] ZHU K, FENG X C, DU X Y, et al. An information bottleneck perspective for effective noise filtering on retrieval-augmented generation[J]. arXiv:2406.01549, 2024.

[77] CUI J, LI Z, YAN Y, et al. Chatlaw: open-source legal large language model with integrated external knowledge bases [J]. arXiv:2306.16092, 2023.

[78] LI W Y, LI J A, RAMOS R, et al. Understanding retrieval robustness for retrieval-augmented image captioning[J]. arXiv: 2406.02265, 2024.

[79] LI X Z, LIU Z H, XIONG C Y, et al. Structure-aware language model pretraining improves dense retrieval on structured data[C]//Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg: ACL, 2023: 11560-11574.

[80] SHI T Y, LI L Z, LIN Z J, et al. Dual-feedback knowledge retrieval for task-oriented dialogue systems[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2023: 6566-6580.

[81] LIN X V, CHEN X, CHEN M, et al. RA-DIT: retrieval-augmented dual instruction tuning[J]. arXiv:2310.01352, 2023.

[82] ROSSET C, CHUNG H L, QIN G H, et al. Researchy questions: a dataset of multi-perspective, decompositional questions for LLM web agents[J]. arXiv:2402.17896, 2024.

[83] FENG J Z, TAO C Y, GENG X B, et al. Synergistic interplay between search and large language models for information retrieval[J]. arXiv:2305.07402, 2023.

[84] SHAO Z H, GONG Y Y, SHEN Y L, et al. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: ACL, 2023: 9248-9274.

[85] LI M F, MIAO S Q, LI P. Simple is effective: the roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation[J]. arXiv:2410.20724, 2024.

[86] TAN J J, DOU Z C, ZHU Y T, et al. Small models, big insights: leveraging slim proxy models to decide when and what to retrieve for LLMs[J]. arXiv:2402.12052, 2024.

[87] YUE Z R, ZENG H M, SHANG L Y, et al. Retrieval augmented fact verification by synthesizing contrastive arguments[J]. arXiv:2406.09815, 2024.

[88] WANG Z, LIU A, LIN H, et al. RAT: retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation[J]. arXiv:2403.05313, 2024.

[89] TRIVEDI H, BALASUBRAMANIAN N, KHOT T, et al. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 10014-10037.

[90] KIM G, KIM S, JEON B, et al. Tree of clarifications: answering ambiguous questions with retrieval-augmented large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2023: 996-1009.

[91] LI X X, ZHAO R C, CHIA Y K, et al. Chain-of-knowledge: grounding large language models via dynamic knowledge

adapting over heterogeneous sources[J]. arXiv:2305.13269, 2023.

[92] ZHANG J W. Graph-ToolFormer: to empower LLMs with graph reasoning ability via prompt augmented by ChatGPT [J]. arXiv:2304.11116, 2023.

[93] NAKANO R, HILTON J, BALAJI S, et al. WebGPT: browser-assisted question-answering with human feedback [J]. arXiv:2112.09332, 2021.

[94] JIANG Z B, XU F, GAO L Y, et al. Active retrieval augmented generation[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2023: 7969-7992.

[95] ASAI A, WU Z, WANG Y, et al. Self-RAG: learning to retrieve, generate, and critique through self-reflection[J]. arXiv:2310.11511, 2023.

[96] LU H Z, LIU Z X. Improving retrieval-augmented code comment generation by retrieving for generation[J]. arXiv:2408.03623, 2024.

[97] XIA Y, ZHOU J B, SHI Z H, et al. Improving retrieval augmented language model with self-reasoning[J]. arXiv:2407.19813, 2024.

[98] YANG D J, RAO J M, CHEN K Z, 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. New York: ACM, 2024: 730-740.

[99] WANG C R, LONG Q Q, XIAO M, et al. BioRAG: a RAG-LLM framework for biological question reasoning[J]. arXiv:2408.01107, 2024.

[100] LIN X Y, WANG W J, LI Y Q, 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. New York: ACM, 2024: 365-374.

[101] OVADIA O, BRIEF M, MISHAELI M, et al. Fine-tuning or retrieval? comparing knowledge injection in LLMs[J]. arXiv:2312.05934, 2023.

[102] SOUDANI H, KANOULAS E, HASIBI F. Fine tuning vs. retrieval augmented generation for less popular knowledge [C]//Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. New York: ACM, 2024: 12-22.

[103] LEE J, CHEN A, DAI Z Y, et al. Can long-context language models subsume retrieval, RAG, SQL, and more? [J]. arXiv:2406.13121, 2024.

[104] JIANG X K, FANG Y, QIU R H, et al. TC-RAG: turing-complete RAG's case study on medical LLM systems[J]. arXiv:2408.09199, 2024.

[105] BARNETT S, KURNIAWAN S, THUDUMU S, et al. Seven failure points when engineering a retrieval augmented generation system[C]//Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI. New York: ACM, 2024: 194-199.

[106] ZHAO X, LU J, DENG C, et al. Beyond one-model-fits-all: a survey of domain specialization for large language models[J]. arXiv:2305.18703, 2023.

[107] BLAGOJEVI V. Enhancing RAG pipelines in haystack: introducing DiversityRanker and LostInTheMiddleRanker[EB/OL]. (2023-08-09)[2024-10-07]. https://towardsdatascience.com/enhancing-rag-pipelines-in-haystack-45f14e2bc9f5.

[108] SINGAL R, PATWA P, PATWA P, et al. Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs[C]//Proceedings of the 7th Fact Extraction and Verification Workshop. Stroudsburg: ACL, 2024: 91-98.

[109] LEE J S, HSIANG J. Patent claim generation by fine-tuning OpenAI GPT-2[J]. World Patent Information, 2020, 62: 101983.

[110] PARK J S, O'BRIEN J, CAI C J, et al. Generative agents: interactive simulacra of human behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. New York: ACM, 2023: 1-22.

[111] WU J D, ZHU J Y, QI Y L, et al. Medical graph RAG: towards safe medical large language model via graph retrieval-augmented generation[J]. arXiv:2408.04187, 2024.

[112] DONG Y, MU R H, ZHANG Y H, et al. Safeguarding large language models: a survey[J]. arXiv:2406.02622, 2024.

[113] ROFFO G. Exploring advanced large language models with LLMsuite[J]. arXiv:2407.12036, 2024.

[114] LENG Q, UHLENHUTH K, POLYZOTIS A. Best practices for LLM evaluation of RAG applications[EB/OL]. (2023-09-12)[2024-10-07]. https://www.databricks.com/blog/LLM-auto-eval-best-practices-RAG.

[115] ES S, JAMES J, ANKE L E, et al. RAGAs: automated evaluation of retrieval augmented generation[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2024: 150-158.

[116] LIU Y, HUANG L Z, LI S C, et al. RECALL: a benchmark for LLMs robustness against external counterfactual knowledge[J]. arXiv:2311.08147, 2023.

[117] SAAD-FALCON J, KHATTAB O, POTTS C, et al. ARES: an automated evaluation framework for retrieval-augmented generation systems[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2024: 338-354.

[118] TANG Y X, YANG Y. MultiHop-RAG: benchmarking retrieval-augmented generation for multi-hop queries[J]. arXiv:2401.15391, 2024.

[119] LYU Y J, LI Z Y, NIU S M, et al. CRUD-RAG: a comprehensive Chinese benchmark for retrieval-augmented generation of large language models[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-32.

[120] XIONG G Z, JIN Q, LU Z Y, et al. Benchmarking retrieval-augmented generation for medicine[J]. arXiv:2402.13178, 2024.

[121] WANG S, KHRAMTSOVA E, ZHUANG S Y, et al. FeB4RAG: evaluating federated search in the context of retrieval augmented generation[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2024: 763-773.

[122] XU Z K, LI Y H, DING R X, et al. Let LLMs take on the latest challenges! A Chinese dynamic question answering benchmark[J]. arXiv:2402.19248, 2024.

[123] WANG S T, LIU J N, SONG S R, et al. DomainRAG: a Chinese benchmark for evaluating domain-specific retrieval-augmented generation[J]. arXiv:2406.05654, 2024.

[124] YU X D, CHENG H, LIU X D, et al. ReEval: automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks[C]//Findings of the Association for Computational Linguistics: NAACL 2024. Stroudsburg: ACL, 2024: 1333-1351.

[125] HOFSTÄTTER S, CHEN J C, RAMAN K, et al. FiD-Light: efficient and effective retrieval-augmented text generation [C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2023: 1437-1447.

[126] CUCONASU F, TRAPPOLINI G, SICILIANO F, et al. The power of noise: redefining retrieval for RAG systems[C]// Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2024: 719-729.

[127] SALEMI A, ZAMANI H. Evaluating retrieval quality in retrieval- augmented generation[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2024: 2395-2400.

[128] ZHU K L, LUO Y F, XU D L, et al. RAGEval: scenario specific RAG evaluation dataset generation framework[J]. arXiv:2408.01262, 2024.

[129] RU D, QIU L, HU X, et al. RAGChecker: a fine-grained framework for diagnosing retrieval-augmented generation [J]. arXiv:2408.08067, 2024.

[130] TU S Q, WANG Y C, YU J F, et al. R-Eval: a unified toolkit for evaluating domain knowledge of retrieval augmented large language models[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2024: 5813-5824.

[131] WANG A, PRUKSACHATKUN Y, NANGIA N, et al. SuperGLUE: a stickier benchmark for general- purpose language understanding systems[C]//Advances in Neural Information Processing Systems, 2019.

[132] PETRONI F, PIKTUS A, FAN A, et al. KILT: a benchmark for knowledge intensive language tasks[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2021: 2523-2544.

[133] YANG Z L, QI P, ZHANG S Z, et al. HotpotQA: a dataset for diverse, explainable multi-hop question answering[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 2369-2380.

[134] THORNE J, VLACHOS A, CHRISTODOULOPOULOS C, et al. FEVER: a large-scale dataset for fact extraction and verification[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Stroudsburg: ACL, 2018: 809-819.

[135] DINAN E, ROLLER S, SHUSTER K, et al. Wizard of Wikipedia: knowledge-powered conversational agents[J]. arXiv: 1811.01241, 2018.

[136] DEYOUNG J, JAIN S, RAJANI N F, et al. ERASER: a benchmark to evaluate rationalized NLP models[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 4443-4458.

[137] ZHANG S, LIU X D, LIU J J, et al. ReCoRD: bridging the gap between human and machine commonsense reading comprehension[J]. arXiv:1810.12885, 2018.

[138] GOTTSCHALK S, DEMIDOVA E. EventKG: a multilingual event-centric temporal knowledge graph[C]//Proceedings of the 15th International Conference on the Semantic Web. Cham: Springer, 2018: 272-287.

[139] HUANG J, SHAO H Y, CHANG K C, et al. Understanding jargon: combining extraction and generation for definition modeling[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2022: 3994-4004.

[140] 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.

[141] LIANG X, SONG S C, NIU S M, et al. UHGEval: benchmarking the hallucination of Chinese large language models via unconstrained generation[J]. arXiv:2311.15296, 2023.

[142] KAMALLOO E, THAKUR N, LASSANCE C, et al. Resources for brewing BEIR: reproducible reference models and an official leaderboard[J]. arXiv:2306.07471, 2023.

[143] KASAI J, SAKAGUCHI K, LE B R, et al. RealTime QA:

what's the answer right now?[C]//Advances in Neural Information Processing Systems, 2024 .

[144] FISCH A, TALMOR A, JIA R, et al. MRQA 2019 shared task: evaluating generalization in reading comprehension [C]//Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Stroudsburg: ACL, 2019: 1-13.

[145] ZHENG L, CHIANG W L, SHENG Y, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023: 46595-46623.

[146] GIENAPP L, SCELLS H, DECKERS N, et al. Evaluating generative ad hoc information retrieval[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2024: 1916-1929.

[147] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.

[148] FEI Z, SHEN X, ZHU D, et al. LawBench: benchmarking legal knowledge of large language models[J]. arXiv:2309.16289, 2023.

[149] MULUDI K, FITRIA K M, TRILOKA J, et al. Retrieval-augmented generation approach: document question answering using large language model[J]. International Journal of Advanced Computer Science and Applications, 2024, 15(3): 776-785.

[150] KURATOV Y, BULATOV A, ANOKHIN P, et al. In search of needles in a 11M haystack: recurrent memory finds what LLMs miss[J]. arXiv:2402.10790, 2024.

[151] EDGE D, TRINH H, CHENG N, et al. From local to global: a graph RAG approach to query-focused summarization [J]. arXiv:2404.16130, 2024.

[152] YASUUNAGA M, AGHAJANYAN A, SHI W, et al. Retrieval-augmented multimodal language modeling[C]//Proceedings of the International Conference on Machine Learning, 2023: 39755-39769.

[153] LI J, LI D, SAVARESE S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of the International Conference on Machine Learning, 2023: 19730-19742.

[154] ZHU W R, YAN A, LU Y J, et al. Visualize before you write: imagination-guided open-ended text generation[C]//Findings of the Association for Computational Linguistics: EACL 2023. Stroudsburg: ACL, 2023: 78-92.

[155] ZHAO J M, HAFFARI G, SHAREGHI E. Generating synthetic speech from SpokenVocab for speech translation[C]//Findings of the Association for Computational Linguistics: EACL 2023. Stroudsburg: ACL, 2023: 1975-1981.

[156] CHAN D M, GHOSH S, RASTROW A, et al. Using external off-policy speech-to-text mappings in contextual end-to-end automated speech recognition[J]. arXiv:2301.02736, 2023.

[157] YANG A, NAGRANI A, SEO P H, et al. Vid2Seq: large-scale pretraining of a visual language model for dense video captioning[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 10714-10726.

[158] NASHID N, SINTAHA M, MESBAH A. Retrieval-based prompt selection for code-related few-shot learning[C]//Proceedings of the 2023 IEEE/ACM 45th International Conference on Software Engineering. Piscataway: IEEE, 2023: 2450-2462.

[159] DU Y, LI S, TORRALBA A, et al. Improving factuality and reasoning in language models through multiagent debate[J]. arXiv:2305.14325, 2023.

[160] LIANG T, HE Z W, JIAO W X, et al. Encouraging divergent thinking in large language models through multi-agent debate[J]. arXiv:2305.19118, 2023.

[161] CHEN J C, SAHA S, BANSAL M. ReConcile: round-table conference improves reasoning via consensus among diverse LLMs[J]. arXiv:2309.13007, 2023.

[162] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Advances in Neural Information Processing Systems, 2022: 24824-24837.