大语言模型检索增强生成优化技术研究综述

袁 乐[1] 刘绍华[1),2)] 王 禹[1] 朱尚威[1] 王 焘[3] 毛天露[4]

[1)](北京邮电大学智能信息物理融合系统研究实验室 北京 100876)
[2)](北京邮电大学安全生产智能监控北京市重点实验室 北京 100876)
[3)](中国科学院软件研究所基础软件与系统重点实验室 北京 100190)
[4)](中国科学院计算技术研究所移动计算与新型终端北京市重点实验室 北京 100190)

摘 要 大语言模型(Large Language Models,LLMs)凭借出色的生成能力,已成为自然语言处理领域的核心技术。然而,其面临幻觉(Hallucination)、知识过时以及推理过程不透明且难以追溯等挑战,限制了其在实际应用中的可靠性。检索增强生成(Retrieval Augmented Generation,RAG)技术作为一种有效解决方案,通过整合外部数据库的实时信息,显著提升了模型的准确性和可信度,尤其在知识密集型任务中表现优异。RAG巧妙融合了LLM的内置知识与外部动态信息资源,实现了知识的持续更新和领域特定信息的无缝衔接。本文深入剖析RAG核心流程,探讨各环节的优化潜力及其相互关联,并提出具体的增强优化策略。通过对比检索增强与语言模型增强的不同方法,分析其对RAG系统性能的影响及适用场景。本文进一步阐述了RAG技术的必要性与局限性,并展望了未来的研究方向和潜在技术突破,旨在为相关领域提供清晰的理论洞察与实践参考。

关键词 大语言模型;检索增强生成;知识更新;幻觉;知识密集型任务
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2026.00383

A Comprehensive Review of Optimization Techniques in Retrieval-Augmented Generation for Large Language Models

YUAN Le[1] LIU Shao-Hua[1),2)] WANG Yu[1] ZHU Shang-Wei[1] WANG Tao[3] MAO Tian-Lu[4]

[1)](Intelligent Cyber-Physical Systems Research Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876)
[2)](Beijing Key Laboratory of Work Safety Intelligent Monitoring, Beijing University of Posts and Telecommunications, Beijing 100876)
[3)](Key Laboratory of Foundational Software and Systems, Institute of Software, Chinese Academy of Sciences, Beijing 100190)
[4)](Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

Abstract Large Language Models (LLMs), with their remarkable generative capabilities, have emerged as a cornerstone of natural language processing and have achieved strong performance across tasks such as machine translation, information extraction, summarization, and dialogue. Nevertheless, LLMs still encounter persistent challenges including hallucination, knowledge staleness, insufficient factual grounding, and the opacity and untraceability of their reasoning processes, which weaken reliability and interpretability in practical deployments—especially in knowledge-intensive or time-sensitive settings where correctness and transparency are essential. Retrieval-Augmented Generation (RAG) has therefore been introduced as an effective paradigm that couples LLMs with external knowledge retrieval. By integrating real-time and verifiable

收稿日期:2025-03-25;在线发布日期:2025-11-06。本课题得到国家自然科学基金(91938301)资助。 **袁 乐**,博士研究生,主要研究领域为RAG和Agent以及自动化程序缺陷定位与修复。E-mail:yuanle@bupt.edu.cn。**刘绍华**(通信作者),博士,副教授,主要研究领域为人工智能与网络分布式计算。E-mail:liushaohua@bupt.edu.cn。**王 禹**,博士研究生,主要研究领域为RAG与Agent。**朱尚威**,博士研究生,主要研究领域为自然语言处理和图像光电容积脉搏波。**王 焘**,博士,副研究员,主要研究领域为生成式人工智能模型加速技术。**毛天露**,博士,副研究员,中国计算机学会(CCF)会员,主要研究领域为人工智能、建模与仿真。

information from external databases into the generation process, RAG enhances accuracy, factual consistency, and controllability while alleviating the inherent incompleteness of purely parametric knowledge. Through this hybrid approach, models can continuously update knowledge, adapt to domain-specific requirements, and deliver more trustworthy outputs in complex tasks. This paper presents a comprehensive and systematic review of RAG optimization from a process-oriented perspective. Using the core workflow of indexing, retrieval, and generation as the structural backbone, it analyzes the optimization potential of each stage and establishes an integrated six-category enhancement framework that includes pre-retrieval, retriever, retrieval-strategy, index, post-retrieval, and LLM enhancement. Each category is examined in terms of motivation, underlying mechanism, and application characteristics, forming a unified and logically coherent view of the RAG pipeline and clarifying the interfaces and dependencies among components. Furthermore, the paper conducts a comparative study of retrieval-oriented versus LLM-oriented enhancement routes, articulating their respective goals, advantages, limitations, and complementarities, and providing a structured lens for understanding how retrieval and generation can be coordinated to improve end-to-end behavior. In addition, the survey summarizes mainstream datasets, benchmark tasks, and evaluation metrics commonly used in RAG research, offering quantitative baselines and methodological references that support reproducible assessment and fair comparison across systems. Building on these findings, the paper emphasizes the necessity of RAG as a complement to the expanding context window of LLMs: while longer contexts improve long-document handling, RAG remains crucial for precision, adaptability, and cost-effectiveness by selectively injecting relevant external evidence. At the same time, the study consolidates key limitations reported in current work—covering retrieval quality, generation consistency, long-term memory and context management, computational efficiency, and cross-modal extension—and highlights their implications for system design and deployment. The survey also clarifies the relationship between RAG and context-window expansion, delineating their complementary roles and boundaries to guide practical configuration choices. Finally, the review outlines forward-looking research directions derived from these observations, including dynamic and multi-hop retrieval mechanisms for complex reasoning, efficient organization and reuse of long-term context, consistency enhancement under noisy retrieval, efficiency optimization through lightweight architectures and distributed computation, structured knowledge fusion to support domain-specific reasoning, and multimodal RAG to broaden application scope. By systematically synthesizing the RAG workflow, contrasting enhancement strategies, and summarizing research progress, datasets, evaluation indicators, limitations, and trends, this paper provides clear theoretical insights and practical references for the ongoing optimization and sustainable development of retrieval-augmented large language models, offering a comprehensive perspective that supports continued exploration and refinement in real-world, knowledge-intensive scenarios.

Keywords large language models; retrieval-augmented generation; knowledge updating; hallucination; knowledge-intensive tasks

1 引 言

近年来，以GPT系列[1-2]、Claude系列[3-4]、Gemini系列[5]以及开源的Llama系列[6]、Qwen系列[7]、DeepSeek系列[8-9]为代表的大语言模型在自然语言处理领域取得了显著进展。这些模型在MMLU[10]、MMMU[11]、GPQA[12]、HumanEval[13]、

LiveCodeBench[14]、MATH[15]等基准测试中展现出卓越性能。然而，尽管其生成能力突出，这些模型在处理特定领域任务时仍暴露出若干局限性[16]。尤其当查询涉及超出训练数据范围的内容或需依赖最新信息时，模型易生成错误信息或出现“幻觉”[17]，从而影响其可靠性和实用性。

为应对上述挑战，RAG技术[18]应运而生，通过引入外部知识检索模块，动态补充模型缺失或遗忘的知识，实现生成内容的精准性与时效性的双重提升。相较于仅依赖模型内部参数存储知识，RAG在可控性、可解释性和轻量化部署等方面具有明显优势，已成为近年来强化语言模型的重要研究方向。

如图1所示，RAG技术起源于开放域问答（Open-Domain Question Answering，Open-QA）的复兴。当时，预训练语言模型尚未成熟，知识容量有限，难以应对广泛事实性问题。研究者提出“检索＋生成/抽取”范式，以外部知识库弥补不足[19-20]。这一时期的范式转变源于当时模型知识存储的局限性，检索机制不仅提升了事实准确性，还奠定了从外部动态获取信息的思想基础。

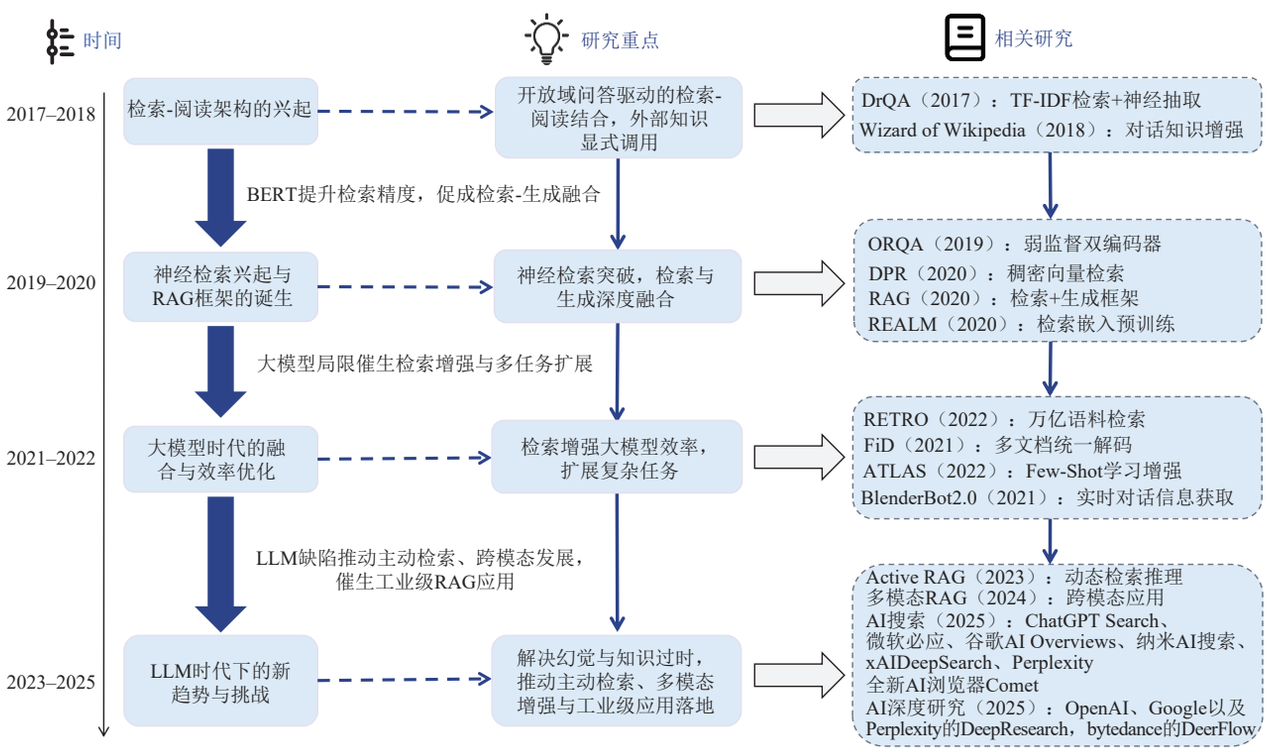

图1　RAG技术发展及研究进展

随着BERT[21]等预训练模型的兴起，RAG技术进入快速发展期。神经检索技术（如ORQA[22]和DPR[23]）通过对比学习提升检索精度。在此背景下，Lewis等人[18]于2020年正式提出RAG框架，结合BART[24]与神经检索器，在知识密集型任务中取得优异表现。同期，REALM[25]将检索嵌入预训练，实现知识与生成的无缝衔接。这一阶段RAG技术的提出标志着参数化知识与非参数化知识结合范式的成型，旨在解决模型内部知识难以精确检索和更新的难题，同时满足对可解释性与时效性的需求。

随着GPT-3[26]等超大规模语言模型的问世，RAG技术在大模型背景下迎来新一轮演进。尽管超大模型记忆能力强大，但知识更新困难与长尾事实遗忘问题日益凸显。为此，研究者探索通过检索增强较小模型以匹敌超大模型性能[27-28]。同时，RAG扩展至少样本学习（如Atlas[29]）和对话系统（如BlenderBot 2. 0[30]），并通过KILT[31]基准标准化评估。这一阶段的范式转变源于大模型封闭知识的局限性，检索机制不仅降低了参数规模依赖，还满足了复杂任务中整合多源信息的需求，研究重心逐步转向知识获取效率与模型可控性。

2023年之后，ChatGPT[32]的成功将RAG推向前沿。Shi等人提出的RePlug[33]方法通过优化检索文档引导“黑盒”大模型生成准确答案，展示了RAG对封闭API模型的适用性。与此同时，主动检索增强（Active RAG[34]）范式兴起，模型可在生成过程中

动态决定检索时机与内容，并结合思维链[35]（Chain of Thought，CoT）提示提升复杂推理能力。多模态RAG[36]的探索亦在拓展，例如通过检索图像丰富文本生成，或通过检索文本辅助图像问答。在工业界，RAG的实用价值得到广泛验证并不断深化。例如，微软新必应与谷歌AI Overviews通过实时网页检索和语义理解，提升搜索的时效性与可靠性；纳米AI搜索整合DeepSeek等先进模型，实现跨模态搜索与高效信息处理；Perplexity的AI浏览器Comet[37]通过“代理搜索”实现自主任务执行，优化用户交互体验；xAI的DeepSearch[38]基于Grok 3，利用X平台数据提供精准摘要与深度推理；OpenAI、Google及Perplexity的DeepResearch工具[39-41]通过自动化研究流程，能在数分钟内生成结构化专业报告，加速金融、医疗及科研领域的文献综述与数据分析；字节跳动的DeerFlow[42]通过多代理架构整合搜索、代码执行与多模态输出，开源框架支持快速生成报告与播客，优化复杂研究任务。这一阶段的趋势聚焦于解决大模型的幻觉与时效性瓶颈，RAG通过低成本、高效更新知识库，结合AI搜索与深度研究的协同创新，奠定了其在大模型演进中的核心地位。

随着RAG技术在LLM时代的迅猛发展，其应用场景与研究方向持续拓展，相关综述性文献已屡见不鲜。例如，Gao等人[43]提出面向LLM的系统性分类框架，将RAG划分为“朴素RAG”、“高级RAG”和“模块化RAG”三大范式。Gupta等人[44]则回顾了RAG的演进历程，涵盖问答、文本摘要及知识推理等应用场景。Ding等人[45]聚焦检索增强大语言模型，按照体系结构、训练策略和应用场景分类。Zhao等人[46]关注面向AI生成内容的RAG应用。这些综述各具特色，为研究者提供了不同维度的参考。

然而，现有综述在梳理RAG技术脉络方面虽取得进展，但局限性不容忽视，包括覆盖面碎片化、技术细节深度不足及前瞻性分析欠缺等问题，限制了其在理论指导与实践应用中的效用。有别于上述工作，本综述从优化视角出发，以RAG工作流程为抓手，条分缕析各环节的增强策略、理论基础及实现方法，旨在构建全局的理论框架与实践指南。尽管Zhao等人[46]的工作涉及RAG增强环节的优化方法，但其分析局限于增强技术的浅层罗列，缺乏系统性框架串联与深入剖析。相较之下，本文通过细致的工作流分析与跨技术对比，揭示RAG优化的关键节点及其理论依据，对每项技术的实现细节、适用场景及改进方向进行全面探讨，以快速把握优化潜力与未来路径，兼具系统性与实践导向。

本综述以RAG工作流程为主线，串联各种潜在增强技术，深入对比检索增强与LLM增强的策略，探讨其目标、优势、劣势及应用场景的异同与联系，系统整理常用数据集与评估指标，评述RAG的必要性、缺陷与挑战，并展望未来发展方向。其特色在于以流程化视角剖析增强优化技术，通过技术对比与理论洞察指导实践，为研究者与开发者提供优化方向与开发指引，同时为学术界进一步探索RAG提供启示建议。

本文结构组织如下：第1节回顾RAG技术的发展脉络与研究现状；第2节整体勾勒RAG的工作流程顶层设计；第3节以RAG工作流为主线，详细探讨各环节的增强优化技术；第4节聚焦检索增强与LLM增强的协同性与差异性；第5节系系统总结RAG领域常用数据集与评估指标；第6节深入探讨RAG技术的必要性、局限性及挑战；第7节总结全文并对RAG的未来发展方向进行前瞻性展望。

2 检索增强生成框架

RAG框架作为一种“检索-阅读”框架[47]，如图2所示，其核心优势在于融合了检索器的能力，相较于传统的语言模型，可以生成更为精确和可靠的内容。下面详细阐述RAG框架的工作流程及关键组成，包括索引、检索和生成。

2.1 索引创建

索引是RAG框架中的基石，旨在高效组织和准备数据，以便于后续的检索和内容生成。此过程涵盖多个关键步骤，具体包括数据准备、向量化表示及索引创建。

（1）数据准备。作为索引流程的起始点，数据准备包括原始数据的清洗和提取。数据源可能呈现多种形式，如PDF、Word、Excel及HTML文件。这些原始数据需转化为标准化的纯文本格式，以便后续处理。在转换过程中，文本会被划分为更小的、易于管理的单元，称为“块”。块的划分依据语言模型的处理能力和上下文限制。

（2）向量化表示。数据转化为块后，接下来的步骤是将这些文本块转换成向量形式。选择适当的嵌入模型对于提升检索阶段的相似性比较能力极为关键。通过向量化，系统能够在检索阶段迅速且准确地评估查询与文本块间的相似度。

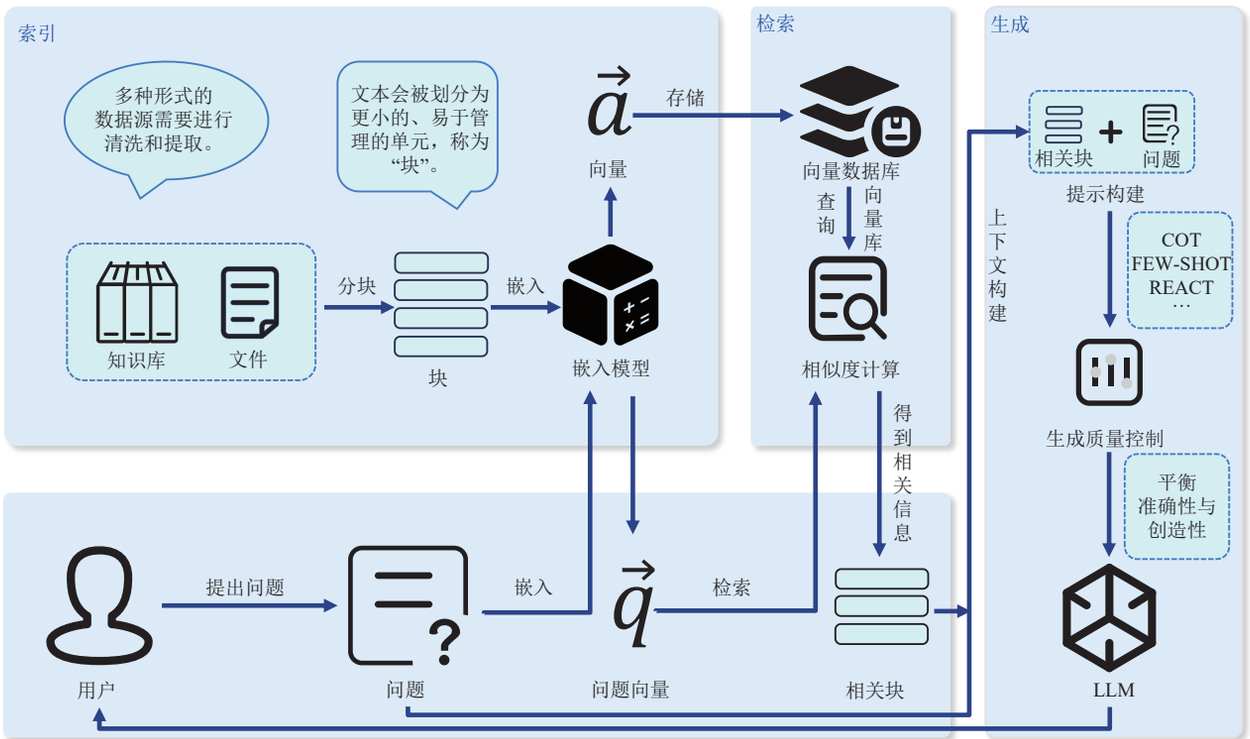

图2　RAG框架

（3）创建索引。索引创建是流程的最后一步，涉及将文本块及其相应的向量表示以键值对的形式进行存储。这一索引结构使系统能够在接收到用户查询时，快速定位并检索到最相关的文本块，实现高效的搜索功能。

2.2　检索

检索阶段的主要任务是在接收到用户的查询后，将其转换为向量表示，并计算与索引中存储的向量块的相似度，以此选择最相关的信息块。

（1）查询处理。此环节包括将用户查询转化为向量表示，使用的嵌入模型一般与索引阶段相同，以确保向量表示的一致性和准确性。

（2）相似度计算。系统对查询向量与索引中的向量化块进行相似度评分，此过程基于向量空间模型，通过比较向量间的距离来确定相似度。

（3）块选择。系统根据相似度评分，选择与查询最接近的前k个块。这些检索到的块将作为扩展上下文，用于生成阶段产生最终的响应。

2.3　生成

生成阶段的主要任务是将用户查询和检索到的信息块综合成连贯的输入提示，进而利用大语言模型生成精确且相关的响应。

（1）提示构建。系统首先构建一个包含用户查询和检索到的文档内容的提示。模型将这些信息结合，形成完整的输入序列，确保生成模型能访问所有必要的上下文信息，以生成准确相关的响应。

（2）响应生成。在提示构建完成后，LLM根据综合的输入序列生成响应。响应的生成既可以依赖于模型的内在知识库，也可以严格基于提供的文档信息，这取决于模型的配置和预设的应用目标。

（3）生成质量控制。确保生成内容的高质量是此阶段的核心。生成的响应应避免与事实不符的“幻觉”信息。此外，内容应保持高度相关性，且避免包含有害或带有偏见的信息。为了达到这一目标，模型需要在保持内容准确性的基础上引入适度的创造性，以增强回答的丰富性和适用性。

3　增强优化环节

RAG系统通过将外部知识库与LLM相结合，显著提升了文本生成的准确性与相关性。然而，其工作流程仍面临多重挑战，涉及从输入数据的准确性到生成内容的可靠性和关联性，以及增强过程的有效性等多个维度。例如，输入中的噪声或不准确信息可能导致检索结果偏误；检索器的低精确度或低召回率可能引发信息幻觉或遗漏；生成内容则可能

引入幻觉、上下文无关性，甚至有害或偏见信息。如何高效整合检索到的文本片段，生成连贯、一致且具创新价值的输出，成为RAG系统的核心挑战之一。

本节系统梳理RAG工作流程中各环节的增强优化策略，涵盖预检索增强、检索器增强、检索策略增强、索引增强、检索后增强及LLM增强六个关键方面。如图3所示，RAG技术流程包括从输入处理到最终输出的完整链条，各环节的优化技术旨在协同提升系统整体性能。首先，预检索增强阶段聚焦输入数据的初步处理，通过关键词识别与意图提取优化查询质量，为后续检索奠定基础。其次，检索器增强阶段通过算法优化实现检索器与LLM的对齐，提升检索精确性并间接改善生成质量。接着，检索策略增强阶段优化检索逻辑与方法，以增强检索的灵活性与深度，确保覆盖多样化信息需求。索引增强阶段则针对检索数据库的索引结构进行优化，通过提升查询匹配的速度与准确性强化系统效率。检索后增强阶段对检索结果进行后处理，通过筛选与重排序确保输出内容的关联性与可靠性。最后，LLM增强阶段利用大语言模型的语义分析能力，进一步精炼生成文本的连贯性与质量。各环节紧密衔接，构成一套系统化的技术路线图，为RAG性能的全面提升提供了理论支持与实践指南。

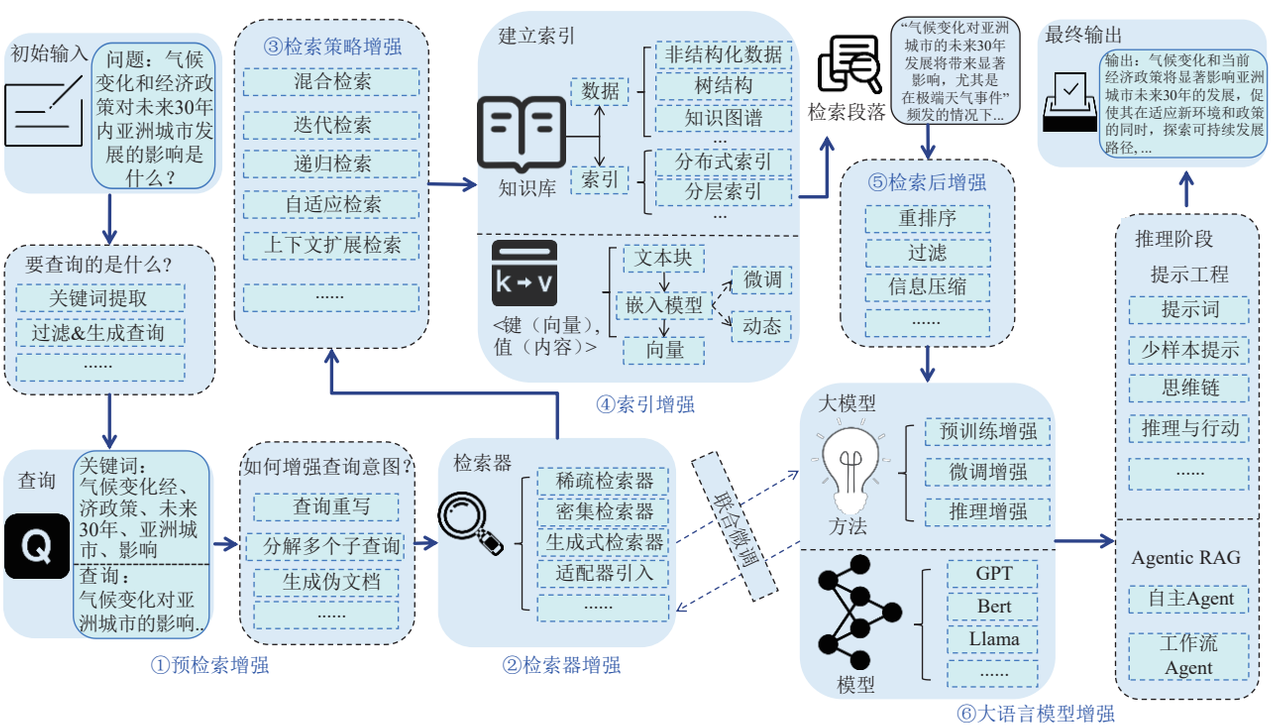

图3 RAG流程中的增强技术

3.1 预检索增强

在RAG系统中，预检索增强作为优化性能的首要环节，旨在通过处理原始输入或优化查询表达，提升查询意图的清晰度与检索结果的精确性，为后续检索与生成奠定基础。原始用户输入常直接用作检索查询，虽简便却因含无关信息、模糊措辞或多义词而难以准确捕捉用户意图，进而影响检索相关性与生成质量。QE-RAG[48]实验验证了这一问题：在面对拼写错误或结构混乱的查询时，传统RAG系统检索准确率下降约10%～20%，生成结果的ROUGE-L[49]得分平均降低15%，显著影响输出可靠性和一致性。预检索增强聚焦两大核心问题——用户查询意图理解与检索相关性提升，并据此分为输入增强与查询增强两大技术路径：输入增强通过提取关键信息（如实体或语义）精炼输入内容，减少噪声；查询增强通过改写或扩展查询优化语义表达，增强多样性与贴近度。两者分别作用于输入内容与查询形式，前者注重信息提炼，后者强调语义优化。本节系统梳理两类技术的实现路径，如表1所示，根据意图理解与相关性提升维度的分野与融合，分析当前挑战、解决方案及未来趋势，为研究与实践提供理论洞察与指南。

表1　预检索增强技术分析对比

类别	子类别	维度/问题	代表方法	核心原理	优势	挑战	适用场景
输入增强	元数据提取	意图理解（快速定位关键信息）	Pakhale[50]	基于NER的关键实体识别	快速定位、精准度高	难以处理复杂查询、多义词	结构化数据、领域特定任务
	语义提取	相关性提升（深入理解语义）	Zhang等人[51]	上下文分析与语义关系挖掘	深入理解语义、适配复杂输入	计算复杂度高、依赖上下文质量	长文本、复杂查询
查询增强	查询重写	意图理解（消除模糊性、优化表达）	RaFe[52], DMQR-RAG[53], RQ-RAG[54]	改写优化（反馈、LLM、多策略）	意图清晰、多样性强	依赖外部模型性能、计算成本高	开放域问答、多步推理任务
	语义扩展	相关性提升（覆盖潜在意图）	Rajaei等人[56], QOQA[57]	语义多样化生成（回译、对齐）	覆盖性强、精度高	质量依赖外部工具、资源需求大	长查询、跨语言或模糊意图
	主题过滤与迭代	意图理解＋相关性提升（动态聚焦意图）	AT-RAG[58]	主题建模与多轮推理优化	动态调整、领域适配性强	依赖主题准确性、成本高	多文档综合、领域特定问答

3.1.1　输入增强

输入增强通过提取关键信息优化原始输入，旨在提升RAG系统对用户意图的初步理解并聚焦检索范围，特别是在复杂查询或领域特定任务中，因用户输入常含冗余或模糊信息而需精炼。其技术路径包括两类：元数据提取利用命名实体识别（NER，Named Entity Recognition）[50]提取结构化线索（如人名、地名），快速定位需求并减少无关结果，从而增强意图理解，尤其适用于结构化数据场景（如金融、法律）；语义提取通过上下文分析（如Zhang等人[51]的语义层次图）挖掘语义关系，超越词汇匹配以深入理解复杂查询，进一步优化检索相关性，适用于长文本与多义词场景，映射至检索相关性优化。理论上，输入增强依托信息抽取理论，通过结构化（元数据）与非结构化（语义）分析降低信息熵，为检索提供低噪声、高相关性的起点。

3.1.2　查询增强

查询增强技术的核心目标是通过多样化策略优化查询表达，以提升检索精度，并准确捕捉用户意图，确保检索结果与用户真实需求高度对齐。本节将系统探讨查询重写、语义扩展和主题过滤等关键技术，结合对比分析揭示其在RAG系统中的应用优势与局限，并展望未来发展方向。

（1）查询重写。查询重写是查询增强的核心方法，通过改写原始查询，使其更贴近目标文档的语义表达，消除模糊性以明确用户意图。RaFe[52]利用排名器的反馈信号训练查询重写模型，依赖排名器自动优化重写效果。DMQR-RAG[53]则通过生成多版本查询，结合信息平衡、关键词、伪答案和核心内容提取四种策略，增强检索的多样性与相关性。而RQ-RAG[54]利用LLM自动改写、分解和消歧查询，提升单跳与多跳问答的表现。EVO-RAG[55]则采用课程指导的强化学习优化查询重写，平衡探索性与精确性，减少冗余子查询，提升多跳问答的检索效率与答案准确率。

（2）语义扩展。语义扩展通过生成语义相近但表达不同的查询版本，增强多样性，更全面覆盖用户潜在意图。Rajaei等人[56]提出基于回译的查询多样化方法，将原始查询翻译成多语言后回译，生成语义一致但表达多样的版本。QOQA[57]则利用LLM重写查询，通过查询-文档对齐分数筛选最优版本。

（3）主题过滤与迭代。主题建模和多轮推理可动态调整查询，提高精度并逐步逼近用户意图核心。AT-RAG[58]将查询增强与主题过滤和迭代推理相结合，利用BERTopic进行主题建模，为查询分配主题以过滤无关信息，并通过链式思维推理迭代优化查询。

查询增强技术在RAG系统中展现出多样化的优势与挑战。现有方法如RaFe依赖排名器反馈优化查询，适于低资源场景但受限于排名器精度；DMQR-RAG与RQ-RAG通过多版本查询生成或LLM改写提升复杂查询处理能力，但计算成本较高；回译与QOQA分别通过跨语言扩展和对齐评分增强检索效果，但计算开销及评分依赖性不容忽视；AT-RAG则凭借主题过滤与推理在特定领域表现优异，资源需求却较大。理论上，查询增强通过语义优化与动态调整弥合查询-文档语义差距，其挑战包括意图理解的多义性（如“苹果”指代水果还是公司？）、高计算复杂性及领域适配性不足。现有研究提出通过上下文感知模型（如BERT）消歧、轻量级模型（如DistilBERT[59]）降低延迟，或融入领域知识图谱提升适配性。未来方向可聚焦多模态增强、自适应策略优化及联合训练查询重写与检索器，以突破瓶颈，推动高效、精准的查询增强框架在复杂推理与多领域任务中的应用。

3.2 检索器增强

检索器负责从海量文档中提取与查询相关的信息，其性能直接影响RAG系统效果。提升检索命中率固然是核心目标，但确保检索结果与LLM的语义需求、推理能力及生成偏好高度对齐同样至关重要。若检索文档与LLM需求不匹配，可能导致生成内容在语境连贯性、语义准确性或输出风格上出现偏差。例如，当检索文档缺乏查询所需的背景信息时，LLM可能生成偏离意图的回答。检索器增强旨在通过优化检索机制及其与LLM的对齐策略，弥合查询与文档间的语义鸿沟，提升系统的精度与实用性。本节系统梳理检索器的主要类型及其增强技术，如表2所示，分析其核心原理与应用价值，探讨当前挑战及未来趋势。

表2 检索器增强技术分析对比

类别	子类别	维度/问题	代表方法	核心原理	优势	挑战	适用场景
检索器	稀疏检索	计算效率（高效检索）	BM25[60]	基于关键词匹配与统计度量	计算高效、实现简便	语义捕获弱、泛化性有限	大规模检索、实时性任务（网页搜索、新闻推荐）
	密集检索	语义相关性（提升语义精度）	DPR[23]，ColBERT[64]	密集向量表示与语义匹配	语义理解强、精度高	资源消耗大，训练复杂	开放域问答、文档分类
	生成式检索	语义相关性＋对齐适配性（灵活匹配）	ListGR[73]，PAG[75]	生成文档标识符与语义排序	灵活性高、潜力大	训练成本高、扩展性待提升	知识密集排序任务
检索器与LLM对齐	微调检索器	对齐适配性（定制化匹配LLM需求）	AAR[77]，ARL2[78]，ToolLLM[79]，DocReLM[80]	利用反馈或领域数据优化检索参数	定制化强、资源节省	依赖数据质量、泛化性受限	零样本问答、领域任务
	适配器引入	对齐适配性（灵活调整输出）	PRCA[81]，Oreo[82]	模块化增强检索-生成匹配	灵活性高、无需改动核心	设计复杂、训练周期长	系统集成与多任务场景
	偏好对齐	语义相关性＋对齐适配性（深度对齐）	DPA-RAG[83]，LarPO[84]	知识偏好一致性优化	推理能力强、质量提升显著	策略复杂、依赖外部机制	复杂推理与知识密集型任务

3.2.1 检索器

检索器的基本任务是从文档集合中检索与查询相关的结果，现有方法主要包括稀疏检索、密集检索和生成式检索。以下从原理、性能对比、优劣势及适用性等维度对三者进行系统分析。

（1）稀疏检索。稀疏检索以BM25[60]为代表，依赖关键词匹配和统计指标（如词频-逆文档频率（Term Frequency-Inverse Document Frequency，TF-IDF[61]））评估查询与文档的相关性。其核心优势在于计算效率高、实现简便，适合大规模文本检索场景，如网页搜索和新闻推荐。在MS MARCO[62] Passage Ranking数据集上，BM25的MRR@10（Mean Reciprocal Rank，MRR）为0.187，此外在BEIR[63]基准的零样本评估中表现稳健，部分任务中甚至优于深度模型，其NDCG@10（Normalized Discounted Cumulative Gain，NDCG）超过DPR。然而，由于仅基于词面匹配，BM25无法有效建模深层语义关系，限制了其在同义表达、隐含意图等语义复杂任务中的适应性。尽管如此，BM25作为基础召回组件，在实际系统中仍广泛用于混合检索与多阶段排序流程，具备工程部署价值。

（2）密集检索。密集检索通过深度神经网络将查询与文档编码为稠密向量，并利用向量空间中的近似最近邻算法进行高维语义匹配。代表方法包括DPR、ColBERT[64]、Contriever[65]、BGE[66]及text-embedding-ada-002[67]等。其中，DPR是首个在Open-QA任务中验证有效性的双塔结构方法。在MS MARCO数据集上，DPR的MRR@10达到0.338，较BM25提升15.1%。在NQ[68]、TriviaQA[69]与WebQuestions[70]等任务上，DPR的Top-20命中率分别达到78.4%、79.4%与73.2%，显著优于BM25（59.1%、66.9%、55.0%）。这表明密集检索在语义建模和复杂问答任务中具有显著优势。然而，在SQuAD[71]数据集上，BM25的Top-20命中率（68.8%）略高于DPR（63.2%），显示方法间对任务结构和语料分布的敏感性差异。此外，在BEIR等零样本检索基准中，DPR的表现有所下降（如MS MARCO子集上NDCG@10为0.177，低于BM25的0.228），表明其泛化能力仍受训练语料的限制。

(3) 生成式检索。生成式检索(Generative Retrieval, GR)是近年来兴起的一种范式,通过序列到序列模型直接生成与查询相关的文档标识符,在大规模语义检索中展现出显著潜力。Pradeep等人[72]研究了GR在MS MARCO 8.8百万文档上的扩展性,强调合成查询对提升检索效果的关键作用。ListGR[73]通过列表级优化显著改进文档排序精度,在MS MARCO 100K数据集上取得MRR@3为0.4656(较NCI[74]提升2.97%),适用于等级相关性任务。PAG[75]在MS MARCO数据集(880万段落)上实现MRR@10为0.385(较NCI提升23.2%),通过前瞻式规划机制将束搜索宽度缩减至1/10,查询延迟降低22倍,有效兼顾效率与准确性。MVDR-GR[76]将生成式检索建模为多向量稠密检索,在MS MARCO(323,569文档)上实现MRR@10为0.589(较DSI提升25.0%),展现多向量框架的性能优势。尽管GR在语义理解和灵活性上表现优异,其训练复杂性和高计算资源需求仍需进一步优化以推动更广泛应用。

三类方法各具优势:稀疏检索以高效著称,适用于资源受限和初级召回场景;密集检索在语义精度上表现显著,适合知识密集型任务如问答与检索增强生成;GR则探索更深层的语义表达潜力,适合构建统一式的端到端系统。在实际应用中,随着任务复杂度的上升与计算资源的丰富,密集与生成式方法日益成为主流。但稀疏方法仍在部署效率和鲁棒性方面保持独特优势,混合检索(如BM25初检+DPR精排)已成为当前系统构建中的主流选择,未来趋势也将聚焦于集成效率、泛化能力与语义对齐的平衡优化。

3.2.2　检索器与大语言模型对齐

对齐旨在提升检索内容与LLM生成需求的相关性与适配性,从而保障生成结果的准确性与一致性。本小节将探讨多种对齐策略,包括微调检索器、引入适配器以及基于偏好的优化方法。

(1)微调检索器。微调通过调整检索器参数增强其与LLM的适配性。监督信号微调(如AAR[77],ARL2[78])利用LLM的输出作为反馈信号优化检索器权重,适用于零样本任务如开放域问答。然而,其效果依赖高质量监督信号,且可能受训练数据偏差影响。

领域知识微调(如ToolLLM[79],DocReLM[80])使用特定领域数据集微调检索器,在特定领域场景(如工具使用、学术问答)中检索精度较高,但需大量领域数据且维护成本高。微调的理论价值在于通过监督学习桥接语义差距,但其泛化性受限于训练数据的覆盖度。

(2)引入适配器。适配器通过附加模块实现灵活对齐。PRCA[81]采用强化学习优化检索与生成之间的匹配,通过奖励机制提升生成质量,适用于问答和文本生成任务,但其训练过程较慢且需设计合适的奖励函数。Oreo[82]作为上下文重构器,从检索文档中提取关键信息并重组为简洁上下文,通过三阶段训练(监督微调、对比多任务学习、强化学习)提升生成准确性。适配器方法的核心优势在于无需改动原有模型,理论上通过任务分解提升系统可扩展性,但效果依赖适配器设计的精细度。

(3)偏好对齐。偏好对齐方法关注检索器与LLM之间知识偏好的匹配。DPA-RAG[83]通过外部(检索器-重排序器)和内部(LLM预对齐)双重机制弥合偏好差距,在知识密集型问答任务中表现优异。LarPO[84]通过引入硬负样本优化生成偏好,提升检索鲁棒性,使LLaMA-7B在AlpacaEval[85]的胜率从21.3%升至37.8%,收敛速度提高3倍,兼顾准确性与效率。偏好对齐通过引入偏好建模,突破传统相关性匹配的局限,但实现复杂度较高。

微调直接优化检索器,适合资源有限场景;适配器提供模块化灵活性,便于系统集成;偏好对齐从根本解决知识不一致,在复杂任务中更具潜力。三者分别从参数调整、模块扩展和偏好匹配角度实现对齐,实践选择应基于任务复杂度与资源条件。

检索器增强的核心在于提升检索精度与LLM的对齐有效性,关注维度涵盖语义相关性、对齐适配性及计算效率。稀疏检索以高效著称但语义相关性有限,适合简单任务;密集检索提升语义精度但效率待优化;生成式检索在语义与适配性上潜力突出,却受效率瓶颈制约;微调检索器增强适配性且效率较高,但语义效果依赖数据质量;适配器方法兼顾语义与适配性,效率因模块化设计而异;偏好对齐优化检索与生成间的协同,但复杂的优化过程可能带来效率瓶颈。当前难点包括语义对齐深度不足、效率-效果权衡及泛化能力瓶颈,为此可通过开发轻量级语义建模(如高效向量压缩)、优化训练流程(如自适应数据生成)及引入多任务学习提升泛化性等步骤推进。未来趋势指向多模态增强以提升语义全面性、自适应对齐框架动态应对任务复杂性、高效生成技术(如轻量级GR框架)降低计算成本,以及端到端优化实现检索器与LLM的语义与偏好无缝衔接,从而推动RAG系统的理论深化与应用拓展。

3.3 检索策略增强

在RAG系统中，传统的单次检索方法在处理复杂任务时常常暴露出局限性，例如信息支持不足、检索结果相关性低或信息遗漏等问题。为应对这些挑战，研究者们提出了多种高级检索策略，通过多轮迭代、动态调整和上下文增强等方式优化检索过程，显著提升模型在复杂任务中的表现。如表3、表4和表5所示，本节深入探讨三类主流检索策略：多步优化策略、动态调整策略和上下文增强策略，系统分析其核心机制、优势与局限性以及适用场景，并探讨当前难点与未来趋势。

3.3.1 多步优化策略

多步优化策略通过多轮迭代或递归的检索过程，逐步细化查询或答案，增强信息整合的深度与广度。这类方法在需要复杂推理或信息逐步逼近的任务中尤为重要，例如多跳问答、事实核验和长篇文档生成。

(1)迭代检索。迭代检索通过闭环反馈机制，利用生成结果优化后续检索，逐步提升答案质量。代表方法如Auto-RAG[86]利用LLM的推理与决策能力，实现多轮迭代检索，自动识别信息需求、生成并优化查询，每次检索后分析结果，无需人工干预或手动规则。在2WikiMultiHopQA[87]和HotpotQA[88]多跳问答任务上，Auto-RAG的F1比分别较基础RAG提升约28%和10%，展现显著优势。IUM[89]通过迭代优化搜索引擎参数，结合期望最大化算法调整参数，每轮基于反馈改进效果。AT-RAG[58]则通过结合BERTopic主题引导的初轮检索与链式推理精炼查询，动态触发新检索以提升复杂多跳查询的效率与准确性，在HotpotQA和2WikiMultiHopQA任务中整体分数分别较基线提升约1.0和1.6分，正确性和完整性均表现最佳。三者的共性在于构建生成-检索反馈循环，类似于强化学习中的“试错优化”，理论上能够逼近全局最优解。然而，迭代次数的增加显著提高了计算开销，且每轮生成的质量直接影响后续检索的效果，因此需要设计高效的收敛条件。

表3 多步优化策略分析对比

类别	维度/问题	代表方法	核心原理	优势	挑战	适用场景
迭代检索	信息完备性（逐步提升答案质量）	AT-RAG[58], Auto-RAG[86], IUM[89]	闭环反馈优化后续检索	答案质量高、适应复杂任务	计算开销大、需确保连贯性	多跳推理、知识库查询
递归检索	信息完备性＋动态适配性（深化精度）	AirRAG[90], HIRO[92], SiReRAG[93], RAG-Star[95]	基于前阶段结果层次化聚焦	精度高、处理歧义能力强	复杂度随深度增加、可能过度细化	歧义澄清、多层次推理

表4 动态调整策略分析对比

类别	维度/问题	代表方法	核心原理	优势	挑战	适用场景
置信度判断	效率平衡（减少冗余计算）	CtrlA[96], WeKnow-RAG[97]	监控置信度动态触发检索	高效应对不确定性、计算少	阈值敏感、易误判	实时问答、高不确定性
生成内容评估	信息完备性＋动态适配性（精确整合）	Self-RAG[100], Open-RAG[102], DeepRAG[103]	自我评估生成内容判断检索需求	知识整合强、精确性高	评估复杂、依赖模型能力	多步骤推理、知识密集任务
任务复杂度	动态适配性（灵活应对需求）	Adaptive-RAG[105], MBA-RAG[106]	根据任务难度动态调整检索策略	效率与效果兼顾、灵活性强	复杂度评估难、需高效反馈	动态决策、复杂查询

表5 上下文增强策略分析对比

类别	维度/问题	代表方法	核心原理	优势	挑战	适用场景
混合检索	信息完备性（提升召回与精度）	Blended RAG[109], Hybrid RAG[110], Ask-EDA[111]	融合关键词与语义搜索综合优化	召回率高、兼顾语法语义	权重平衡难、融合复杂	复杂查询、多模态任务
上下文扩展	信息完备性＋动态适配性（增强语境）	ConvRAG[113], CAG[116], ADACQR[117]	扩展局部/全局上下文提供多层次背景	语境理解强、减少信息遗漏	计算成本高、易引入噪声	会话问答、多义词处理

(2)递归检索。递归检索基于前一阶段检索结果优化后续查询，逐步收窄搜索范围，聚焦最相关信息，适用于嵌套问题或歧义澄清。AirRAG[90]利用蒙特卡罗树搜索(Monte Carlo Tree Search，MCTS[91])驱动递归检索，从问题出发构建树形结构，通过五种推理动作生成子查询，依据前次结果选择最佳路径，适合多跳推理等复杂问答。HIRO[92]采用层次化递归方法，通过计算查询与节点的相似度修剪无关分

支，避免信息过载，为LLM提供精准上下文，适用于长文本检索。SiReRAG[93]通过相似性树和相关性树递归整合语义与实体信息，借鉴RAPTOR[94]分层索引技术并加入实体提取，适合挖掘语义与实体关系的复杂任务。RAG-Star[95]则利用MCTS和检索验证递归优化推理路径，从问题扩展子查询，每轮进行评估并修正结果，适合复杂推理。这些方法通过树形或层次化分解，类似认知科学中的“逐步分解”过程，能够有效处理嵌套问题，显著提升精度。然而，递归深度过大可能导致信息过度细化或计算复杂度激增，因此需要在深度与效率间权衡。迭代检索通过闭环优化动态调整答案，适合开放域问答；递归检索通过层次聚焦深化理解，更适于多层次推理。迭代检索的计算负担随轮次线性增长，而递归检索的复杂性呈指数级，需根据任务需求选择。

3.3.2　动态调整策略

动态调整策略通过实时反馈自适应优化检索过程，使模型根据任务需求灵活调整策略。其核心是从被动检索转向主动适应，提升生成内容的相关性和准确性。关键在于判断生成过程中是否需触发检索以优化结果。本节从检索触发机制角度分析自适应检索方法，分为三类：基于置信度判断的检索、基于生成内容评估的检索和基于任务复杂度的检索。

（1）置信度判断。基于置信度判断的检索方法通过监控生成过程中的置信度动态触发检索，在模型信心不足时补充外部信息。CtrlA[96]结合语言模型置信度与诚实性约束优化RAG检索时机，当置信度低或内容不符时触发外部检索并生成验证查询，以防虚假内容，特别适用于长篇和多跳任务。WeKnow-RAG[97]通过置信度自评提升生成可靠性，若置信度低（如事实存疑），返回“我不知道”并启动检索。在CRAG[98]基准上，引入置信度机制使准确率和综合分数（如综合分数从0.10升至0.12）均优于基础的RAG。Vendi-RAG[99]则在生成后引入判别器对答案质量进行评分，若置信度不足，则通过调整检索多样性动态优化文档集合，迭代提升答案可信度。这类方法共性在于以置信度为信号减少冗余计算，理论上类似统计推断中的阈值检验。其优势在于高效应对高不确定性任务，但阈值设置敏感性可能导致误判。

（2）生成内容评估。基于生成内容评估的检索方法通过模型对已生成内容的自我评估，判断是否需进一步检索完善输出。Self-RAG[100]引入反思机制，动态评估答案质量并按需触发检索，适用于知识密集型与多步推理任务，在PopQA[101]上准确率达55.8%，较同规模RAG基线提升近4%，检索次数更少，推理效率更高。Open-RAG[102]通过反思性评估优化LLM在复杂推理任务中的表现，在生成后用嵌入的反思标记（如“是否需要检索”“内容相关性”）检查输出，若不足则触发检索。DeepRAG[103]通过逐步推理和动态检索提高效率，将查询分解为子问题，评估生成内容是否支持推理，若不足则检索。Sufficient Context RAG[104]从上下文充分性出发，利用中间推理评估生成所依赖的信息是否完备，若发现关键推理缺失，则动态补充检索，保障生成质量与一致性。这类方法借鉴元认知反思原理，确保生成质量。其优势在于精确性高，但评估复杂性依赖模型能力。

（3）任务复杂度。基于任务复杂度的检索通过评估任务复杂度决定是否触发检索。Adaptive-RAG[105]根据查询复杂度动态调整策略，用小型分类器判断复杂性，简单问题直接生成答案，复杂问题则启动多轮检索和推理，提升准确性并避免简单任务中的冗余计算。MBA-RAG[106]基于多臂赌博机算法，依据查询复杂度动态选择最佳检索策略，通过平衡“探索”和“利用”优化决策，简单任务直接生成答案，复杂任务采用多轮检索，并以动态奖励机制提升效率。两者共性在于通过动态规划平衡效率与效果，适合多变需求，但需高效反馈机制保障检索准确性与时效性。

动态调整策略的理论基础源于决策理论与控制论，强调通过反馈循环实现系统自适应优化。置信度判断类似统计推断中的显著性检验，适用于效率优先场景；生成内容评估借鉴元认知反思机制，适合高精度任务；任务复杂度调整则接近强化学习的动态规划，适应动态环境并优化资源分配。这种策略将检索从静态映射转为主动决策，理论上能更高效分配计算资源并提升生成质量，但效果依赖判断精度与反馈实时性。实践需根据任务实时性与精度需求优化判断机制，降低误判风险。

3.3.3　上下文增强策略

上下文增强策略通过提供更多背景信息优化检索，帮助模型更好理解复杂问题。该策略通过扩展检索范围，覆盖更广上下文，使结果更全面，特别适合需要深层语境支持的任务，如会话问答和多义词处理。本节从混合检索和上下文扩展两方面进行分析。

（1）混合检索。混合检索[107]通过整合关键字搜

索与语义搜索提升检索效果：首先生成关键字查询检索特定词汇文档，随后利用向量空间模型识别相关内容，并采用互惠等级融合[108]（Reciprocal Rank Fusion，RRF）算法综合文本匹配度和语义相关性，提高检索的准确性。Blended RAG[109]融合BM25、DPR与稀疏编码器检索，在NQ零样本设置下EM（Exact Match，EM）达42.63%，显著优于单一检索与大模型基线，且稀疏编码器在超大文档集下具备更低延迟与更高效率。Hybrid RAG[110]整合稀疏检索与密集检索，提升检索与推理效果，特别适用于法律和医学等复杂任务。Ask-EDA[111]通过结合BM25、Sentence Transformer和RRF算法，提升设计领域问答系统的检索与排序效果。Rayo等人[112]采用混合检索从法规文本生成答案，结合BM25和Sentence Transformer，克服同义词与语言差异挑战。该类方法优势在于提升召回率与精度，理论上类似多通道信号增强，但需优化权重平衡与融合复杂度。

（2）上下文扩展检索。上下文扩展检索通过扩展局部（如句子窗口）和全局（如会话历史）上下文，提供多层次背景信息。会话历史检索通过整合多轮对话内容，增强语义连贯性与上下文一致性，适用于开放问答与多轮交互，有助于提升模型对用户意图的理解与响应质量。句子窗口检索则扩展局部上下文，补充背景信息，在多义词解析与复杂推理中同样有效。ConvRAG[113]通过问题精炼器和细粒度检索优化会话问答中的上下文依赖，结合会话历史提升查询与文档匹配度，确保答案相关性，在QReCC[114]数据集上BLEU-4[115]达18.58，ROUGE-L达32.40，较Perplexity.ai等主流RAG系统显著提升，且推理效率更高无需多轮检索。CAG[116]则通过动态判断上下文检索需求，结合向量候选方法计算查询与会话历史的相似度，优化历史使用策略，在SQuAD数据集上实现上下文相关性得分0.338、答案相关性0.709，显著优于经典RAG和多轮LLM判定方案，且推理速度提升百倍以上，适合高并发场景。ADACQR[117]通过对话上下文驱动的查询重写机制，将多轮语义融入精炼查询，并对齐稀疏与稠密检索以提升检索一致性与覆盖广度，在QReCC上实现NDCG分数52.5、Recall@10分数76.5、Recall@100分数93.7，全面优于同类方法。Ms. WoW[118]引入多源知识融合机制，通过结构化信息与语义帧扩展对话背景，模拟知识“即插即用”的场景，提升上下文完整性与适应能力，在Wizard of Wikipedia[20]上，Ms. WoW-Gold通过精确选取金标准知识，将F1提升至0.321，显著优化对话生成效果。Alonso等人[119]提出结合时间和事件的长时记忆系统，通过链表检索和语义检索处理复杂会话历史。这类方法的共性在于借鉴语境效应增强推理，优势在于减少信息遗漏，但计算成本高且易引入噪声。

上下文增强策略的理论基础源于语义学的语境依赖性和信息论的信息冗余优化。混合检索通过多元信息融合突破单一检索局限，类似信号处理中的多通道增强；上下文扩展检索借鉴认知心理学的“语境效应”，通过扩展背景提升推理深度。理论上，这类策略能显著降低信息遗漏风险并增强生成一致性，但需解决冗余信息带来的噪声问题。混合检索平衡召回率与精度，关键在于权重优化，可通过任务特定加权或自适应融合算法提升效果。上下文扩展检索深化推理，适合长篇生成和会话问答，但计算成本高且易引入噪声，可用注意力机制过滤无关上下文或限制窗口大小提高效率。总体而言，混合检索适于多样化查询，上下文扩展检索更适合深层语境任务，实践需根据任务需求与资源限制优化信息过滤。

检索策略增强聚焦于提升检索过程的灵活性与适配性，其关注维度包括信息完备性、动态适配性及效率平衡。技术分析表明，迭代检索通过多轮查询优化信息获取，提升完备性与适配性，但整体效率偏低。递归检索进一步加深检索层级，增强语义关联与上下文理解，但计算代价随深度增加。置信度判断机制以判别可信度引导检索流程，效率较高，但易受阈值设计影响，适配性有限。生成内容评估机制则通过分析初步生成结果反向优化检索指令，在提升语义适配性的同时牺牲了一定计算效率。任务复杂度驱动策略根据问题结构动态调整检索深度与内容，适配性强，但在效率与完备性之间需精细权衡。此外，混合检索方法整合关键词与语义检索，兼顾相关性与语义覆盖，性能表现依赖融合算法的优化质量；上下文扩展策略通过引入历史或局部上下文增强查询理解，在提升信息完备性的同时，带来上下文处理负担。

当前检索策略增强仍面临三大挑战：一是计算效率瓶颈，多轮或深度检索显著增加处理开销；二是多策略之间协同不足，策略组合缺乏动态调度机制；三是泛化能力有限，在跨领域或复杂任务中适配性仍不稳定。为缓解上述问题，已有研究提出多种优化路径：通过引入高效索引机制（如知识图谱缓存）与压缩技术减少冗余计算；设计混合框架实现多策

略协同与动态分配；结合多任务学习与元学习机制提升策略泛化能力与任务适配性。未来的发展趋势主要包括：多模态信息融合以增强检索全面性与语义表达力；智能化决策机制（如强化学习）引导策略选择；高效计算技术（如模型蒸馏、量化）降低成本；以及动态语境管理（如自适应窗口）提升上下文处理效率。这些方向有望推动检索策略增强在跨领域、复杂场景中的深度应用。

3.4　索引增强

在RAG系统中，索引增强是提升检索效率与生成质量的关键环节。索引优化不仅提升系统的响应速度和检索精度，还能提供更高质量的上下文输入，从而保障生成内容的准确性与一致性。如表6和表7所示，本节围绕数据源增强和索引结构优化两大维度，系统分析提升数据质量、优化数据组织和改进索引架构的核心机制，结合技术特性与应用价值，探讨当前挑战及未来趋势。

表6　数据源增强技术分析对比

类别		维度/问题	代表方法	核心原理	优势	挑战	适用场景
增强数据质量		数据质量（提升准确性与时效性）	LlamaIndex[120]	文块优化与元数据集成	检索速度快、语义精度高	更新频率高、标准化不足	问答系统、动态信息分析
优化数据结构	非结构化数据优化	数据质量＋结构适配性（灵活处理杂乱数据）	RAGtrans[121], Zhang等人[122], ISAG[123]	跨语言检索与噪音抑制	灵活性强、时效性好	处理效率低、一致性难保障	跨语言任务、新闻生成
	结构化数据优化	结构适配性（增强推理深度）	CABINET[124]，DoTTeR[128]，Graph RAG[132], KET-RAG[134], ToG[137], ToG-2[140]	知识图谱与表格编码	推理深入、信息整合强	构建成本高、跨领域一致性差	多跳推理、法律分析

表7　上下文增强策略分析对比

类别	维度/问题	代表方法	核心原理	优势	挑战	适用场景
分布式索引	检索效率（高吞吐量低延迟）	Milvus[142]	分片并行处理与负载均衡	查询速度快、可扩展性强	资源消耗大、硬件依赖性强	高并发检索、云服务
分层索引	结构适配性＋检索效率（优化长文本精度）	SiReRAG[93], RAPTOR[94], KG-Retriever[144]	递归聚类与层次化索引	精度高、计算熵低	结构复杂、多步开销大	长文本检索、多跳推理

3.4.1　数据源增强

数据源增强通过提升数据质量和优化数据结构，确保输入数据的准确性、时效性和语义丰富性，为检索与生成奠定基础。

（1）数据质量。高质量数据是保障检索相关性与生成准确性的前提，核心在于语义理解和内容更新的及时性。高质量文本输入依赖噪声清理、歧义消除和事实验证，例如去除冗余字符、修正拼写错误、核实数据来源，以保障语义连贯性与可信度。LlamaIndex[120]通过文块大小调整提升语义处理灵活性：问答任务采用小文本块加快检索，内容生成任务使用大文本块保留上下文，从而兼顾响应速度与结果准确性。时效性是数据质量的重要因素，依赖文档内容的定期更新以避免信息滞后。例如，新闻检索需每日更新以捕捉最新事件。集成元数据（如发布时间、作者）有助于优化过滤与排序，特别在市场分析等动态场景中，筛选最新文档可显著提升检索效果。这类方法的优势在于提升检索速度和语义精度，其原理类似于通信中的信噪比优化。然而，频繁更新的需求和元数据标准化不足，仍在一定程度上限制了其应用效果。

数据结构优化通过改进数据的组织形式提升检索效率和生成质量，根据数据类型的特性可分为非结构化数据优化和结构化数据优化两大方向。

（2）非结构化数据优化。非结构化数据优化旨在通过信息抽取与噪音抑制技术，提高从杂乱数据中提取高质量信息的能力。RAGtrans[121]结合维基百科等非结构化文档与翻译任务，通过多任务训练提升翻译准确性与跨语言知识获取能力，在英中翻译中基于Qwen2.5-7B实现BLEU分数56.82（提升6.71），显著优于zero-shot基线，支持多语言辅助与高效大规模检索。Zhang等人[122]聚焦于RAG系统中OCR处理非结构化数据所引入的噪音问题，并在OHRBench框架下对系统表现进行评估，提出引入视觉语言模型以实现降噪和提升数据质量。此外，ISAG[123]依托实时搜索引擎动态获取非结构化信

息，结合PARSER-LLM与EXTRACTOR-LLM提取要点，并通过混合排序优化结果，在真实API请求和多跳推理任务中F1达0.77、EM达0.80，较传统RAG显著提升，且推理token消耗降低约21%。此类方法的优势在于灵活性与时效性强，但大规模处理效率低且一致性难保障。

(3) 结构化数据优化。结构化数据优化通过精准建模和高效索引提升模型对表格和知识图谱等有序数据的处理效率与准确性。CABINET[124]通过为表格单元格分配相关性评分并生成自然语言描述，提升LLM处理复杂表格的聚焦能力与鲁棒性，在WikiTQ[125]和FeTaQA[126]上准确率较DATER[127]提升3.2%和9.6%，560M参数模型实现高效推理，在大表格和高噪场景下表现优异。DoTTeR[128]通过引入基于排名的表格编码方式，结合表格排名信息有效提升表格-文本检索和多跳推理性能，在OTT-QA[129]数据集上相较OTTeR[130] F1值提升1.3%，且推理速度更优，支持高效大规模块级检索与跨块推理。RET-LLMs[131]采用读写存储机制动态更新知识图谱，提高实时性与精度，但跨领域融合仍具挑战。Graph RAG[132]通过社区检测算法提升全局摘要效果，在Podcast[133]数据集上全面性与多样性胜率达83%和82%，较基础RAG提升超30%，C0层摘要token成本仅为全文的2.5%，实现显著性能提升与资源节省，但图索引与信息合并策略仍有优化空间。KET-RAG[134]则在Graph RAG基础之上结合知识图谱与文本-关键词二部图，利用PageRank算法精简信息，提升了索引效率与检索质量。HippoRAG[135]引入类脑记忆机制，将文档表示为语义图并结合个性化PageRank提升结构化检索的联想性与长期记忆保持。HippoRAG 2[136]在此基础上优化图结构构建与上下文融合策略，进一步增强非参数持续学习能力和多跳推理稳定性。ToG[137]将LLM与知识图谱结合，通过束搜索动态选择最优推理路径，提升多跳推理能力与可解释性，在CWQ[138]和WebQSP[139]任务中基于GPT-4实现EM值分别为69.5%和82.6%，较CoT提升超20%，且推理调用次数少、过程可追溯。ToG-2[140]进一步交替检索图谱实体与文本背景信息，强化结构化数据利用，提升推理准确性与深度，有效缓解幻觉问题。这类方法的优势在于推理深度与信息整合能力强，但跨领域一致性差与构建成本高是主要挑战。

数据质量增强通过降低噪声优化“信源质量”，类似通信理论中的信噪比提升，而结构优化通过语义建模和知识表示强化信息组织的逻辑性，二者协同提升系统语义理解能力，为生成模型提供可靠输入分布并减少预测偏差。非结构化优化适用于跨语言和时效性任务，通过动态语义增强提升灵活性，但在大规模数据处理中效率较低；结构化优化则借助知识图谱增强推理深度，适合多跳推理任务，但跨领域一致性较弱。实践中，时效性任务如新闻生成宜优先采用非结构化优化并结合互联网搜索保持信息新鲜度，而知识密集型任务如法律分析则应依赖结构化优化以确保推理严谨性。

3.4.2 索引结构优化

索引结构优化通过改进数据组织与访问方式提升检索效率与准确性，主要策略包括分布式索引和分层索引。

(1)分布式索引。分布式检索策略[141]在RAG系统中至关重要，尤其在大规模数据处理中，通过分片与并行处理提升效率与可扩展性。其核心原理是将知识库拆分为多个“分片”，分配至不同节点并行查询：接收到查询请求后，各节点同步检索对应分片，返回相关文档并合并排序，最终输出至生成模型。分布式索引策略主要包括集中式、完全分布式和混合式三种，各具效率与扩展性特点。这一方法显著加快查询速度，增强系统扩展能力，确保RAG在海量数据场景下高效响应。Milvus[142]采用分布式架构优化高维向量的存储与检索，具备高可用性与可伸缩性。其灵活的索引策略支持多种类型(如IVF_FLAT、HNSW)，可根据应用动态选择最优方案，显著提升查询效率，并适配多GPU环境下的资源管理需求。TELERAG[143]通过前瞻检索优化RAG推理，在预取IVF簇时利用CPU-GPU协作，隐藏数据传输延迟。实验表明，TELERAG在HotpotQA、NQ和TriviaQA数据集上，端到端推理延迟平均降低约41.9%(加速比1.72倍)，且在单GPU(RTX 4090 24 GB内存)上支持61 GB数据存储和Llama-3-8B模型，展现了高效、低内存占用的优势。

(2) 分层索引。分层索引策略通常通过多层次结构化组织数据，优化检索效率与准确性。RAPTOR[94]采用递归聚类与树状索引加速长文本检索，提升大规模查询的精度，适用于复杂任务。SiReRAG[93]构建相似性树与相关性树双层索引，前者按语义聚合同类知识，后者基于实体关系组织命题信息，协同提升多跳推理的覆盖性与准确性。KG-Retriever[144]引入层次化索引图，结合文档层与知识图谱层两轮匹配，提升检索效果并降低计算成本。

分布式索引基于并行计算的“空间换时间”原则提升系统吞吐量，而分层索引借鉴分治法，通过层次分解优化复杂度。二者通过优化检索路径降低计算熵，体现效率与深度的架构创新。在高并发场景（如云服务）中，分布式索引因吞吐量优势更具适用性，但资源消耗较高，可借助云原生架构或边缘计算优化资源管理。分层索引在长文本与推理任务（如学术研究）中表现更优，但结构复杂，可利用图压缩与稀疏检索减少推理时间。针对动态数据适应性问题，自适应调整机制可提升更新效率。未来，分布式与分层索引的优化将聚焦动态自适应框架、多模态索引与智能管理，以进一步提升系统性能与灵活性。

索引增强聚焦于提升数据可访问性与检索效率，其核心关注维度包括数据质量、结构适配性及检索效率。技术分析表明，数据质量优化通过提升输入信息的准确性与时效性改善整体性能，但更新频率高可能影响检索效率。非结构化数据优化有助于增强语义表达与适配性，尤其适用于多样化知识源，但处理复杂度高、实时性受限。结构化数据优化在适配性与准确性方面表现优异，便于构建实体关系图谱，但索引创建与维护成本较高。分布式索引强调高并发与可扩展性，适配大规模检索需求，其性能依赖于合理的分片与路由策略。分层索引则通过多级结构在检索精度与上下文关联性之间取得平衡，但其效率受到索引层级设计的制约。

尽管索引增强技术已取得阶段性进展，仍面临诸多挑战，如计算效率瓶颈、跨领域适配能力不足以及动态更新机制不完善等。为应对这些问题，已有研究尝试引入高效嵌入表示与并行处理机制提升索引响应速度，结合多模态知识图谱增强结构表达能力，并通过增量更新与异步维护机制缓解知识滞后问题。未来发展方向将聚焦于三大趋势：一是多模态融合索引以增强语义理解与表达能力，丰富数据质量维度；二是自适应索引结构，根据任务需求动态重构索引形态，提升适配性与系统鲁棒性；三是借助GPU/TPU等硬件加速资源降低计算开销，推动索引在高实时性场景中的应用落地。这些路径为索引增强技术在复杂、多样化任务中的持续演进提供了坚实支撑。

3.5　检索后增强

在RAG系统中，检索后增强通过优化检索内容的质量与结构，提升生成结果的准确性与效率。由于LLM的上下文窗口受限，且检索阶段可能引入噪声或冗余信息，直接将未经处理的检索结果输入模型可能导致关键信息遗漏、幻觉或计算负担加重。研究表明，LLM在处理多文档时存在显著的“位置偏置”现象，Cuconasu等人[145]发现模型更倾向于使用前位文档，忽视后部有效证据，影响生成准确性。检索后增强通过重排序与过滤以及信息压缩两大技术方向，精细化处理检索输出，确保输入LLM的内容高度相关且精简。如表8所示，本节系统分析其理论基础、核心机制与应用价值，探讨当前挑战与未来趋势。

表8　检索后增强技术分析对比

类别	子类别	维度/问题	代表方法	核心原理	优势	挑战	适用场景
重排序与过滤	基于规则	相关性优化（提升排序效率）	Yan等人[146]，Zheng等人[147]，Hagen等人[148]	通用性计算与特异性评分	高效、可解释性强	语义捕捉有限	资源受限、时效性任务
	基于模型	相关性优化＋语义完整性（深度语义匹配）	FiD-Light[149]，W-RAG[150]，RADIO[151]，CFT-RAG[152]，G-RAG[153]，RankCoT[156]	LLM评分与推理优化	语义捕捉强、适应性高	训练成本高、推理延迟	复杂查询、多跳推理
信息压缩		输入精简性（降低计算负担）	MAC[158]，Refiner[160]，COCOM[161]，Shi等人[162]	元学习与嵌入压缩	效率高、计算负担低	语义损失、泛化性不足	资源受限、长文本推理

3.5.1　重排序与过滤

重排序与过滤旨在优化检索结果，筛选并优先呈现最相关、最有价值的内容，以提升LLM对关键信息的利用效率。其核心在于优化信息流，减少冗余，提升内容质量。根据实现方式不同，重排序可分为基于规则与基于模型两类，各具理论根基与适用性。

（1）基于规则的重排序。基于规则的重排序通过预定义的启发式算法或排序规则优化检索结果，强调高效性与可解释性，适用于资源受限或需高透明度的场景。Yan等人[146]针对生物医学检索任务，利用MeSH本体评估文档与查询的通用性，并通过规则调整排名，使其概括程度更契合查询需求，适用于综述类查询。Zheng等人[147]通过归一化逆文档频率与熵计算文档特异性，优先返回焦点窄、内容具体的文档，并采用硬剪裁或软剪裁策略优化排序。Hagen等人[148]引入排序公理（如词频、文档长度、术语

接近性),构建偏好矩阵,并通过投票融合算法优化排名。这类方法的理论源于传统信息检索的相关性量化假设,优势在于计算高效且透明,适合资源受限场景,但语义捕捉能力有限,难以处理复杂上下文。

(2) 基于模型的重排序。基于模型的重排序利用深度学习技术捕捉文档与查询的语义关系,提升排序精度与任务适应性。FiD-Light[149]采用列举式自回归方法优化文档排序,并使用文本引用作为信息指针,适用于企业知识管理、技术支持文档搜索。W-RAG[150]利用LLM对BM25初检结果打分重排,生成弱监督数据优化密集检索器,在NQ数据集上使DPR检索的F1值达0.1528,优于BM25的0.1374,Top-1生成延迟仅1.04秒。RADIO[151]通过推理蒸馏和基于推理的对齐,筛选支持生成器推理的文档。CFT-RAG[152]结合布谷鸟过滤器与层次化实体树,高效筛选重排高价值文档,在复杂查询中比传统方法快138倍,在大规模实体检索中准确率达66.5%、与基线持平。G-RAG[153]利用图神经网络和抽象意义表示构建文档图,增强跨文档推理能力,在NQ数据集上将MRR提升至19.8%,较无Reranker和BERT的基线分别提升7.7和1.9个百分点,显著优于PaLM 2[154]等大模型排序方案。RankRAG[155]则通过将上下文排序信号融入生成器训练,统一排序与生成过程,在HotpotQA上EM提升4.3%、F1提升3.9%,显著优化答案相关性与一致性。RankCoT[156]结合CoT和重排序信号,通过自反思优化知识提炼,提升多跳推理任务的答案质量。ASRank[157]提出基于“答案气味”的零样本文档重排序方法,利用LLM评估文档回答潜力并重排,无需训练数据,在NQ上EM提升3.4%、*F*1提升2.7%,展现出出色的跨任务泛化能力。这类方法的理论根基为排序学习与语义表示,优势在于语义理解强、适应性高,但训练与推理成本较高。

基于规则的重排序因计算复杂度低、效率高,适用于资源受限环境(如边缘设备)或需高透明度的任务(如医疗RAG的来源过滤),但其语义理解能力有限,难以应对复杂查询。相比之下,基于模型的重排序凭借强大的语义捕捉能力,更适用于复杂查询与多跳推理任务(如开放域问答),但训练与推理成本较高。在实践中,资源受限任务可优先采用基于规则的方法,并结合领域自适应规则提升效率;高精度任务则更适用于基于模型的方法,可借助弱监督或自监督降低标注需求。

3.5.2 信息压缩

信息压缩旨在削减冗余内容、优化输入长度,同时保留关键语义以提升生成效率与质量。MAC[158]通过元学习和无梯度优化将新知识压缩为调制参数存储于记忆库,避免直接更新LLM权重,在StreamingQA[159]等任务中*F*1达21.14%,较基线提升2%~4%,适应速度快10倍,显存占用降低约2/3,知识保留率高达96%。Refiner[160]采用“提取-重构”策略,通过层次化组织相关信息,将输入token减少80.5%,多跳问答准确率提升1.6%至7.0%。COCOM[161]利用嵌入式压缩将输入压缩至1/128,推理速度提高5.69倍,计算量减少22倍,同时保持90%以上准确率。Shi等人[162]基于抽象意义表示(Abstract Meaning Representation,AMR)的概念蒸馏方式,通过层次化压缩和注意力筛选提取核心语义,优化长文本处理,超越传统TF-IDF和截断方法。Jin等人[163]则结合检索重排序与显式/隐式微调,压缩无关信息,提升RAG在长输入下的稳健性和准确性。此类方法的理论基于信息论与降维,通过降维或提炼降低冗余,优势在于降低计算负担与提升长文本处理能力,挑战在于语义损失与泛化性不足。

信息压缩的核心在于最小化信息熵损失,其优势在于提升长文本处理能力,劣势则在于压缩率与语义完整性的权衡。嵌入压缩具备高压缩率与低计算成本,适用于资源受限或长文本场景;语义提炼侧重语义保留,更适合知识密集型与多跳推理任务。实践中,嵌入压缩提升处理效率,语义提炼确保信息完整性,可根据任务需求动态调整压缩率。

检索后增强致力于提升检索内容的质量与效率,其关注维度涵盖相关性优化、输入精简性及语义完整性。技术分析表明,基于规则的重排序方法在相关性提升方面表现稳定,操作简单、效率高,但对深层语义的捕捉能力有限,且对输入精简支持不足。基于模型的重排序方法依托语言模型强大的理解能力,可同时优化相关性与语义完整性,但对计算资源依赖较高,输入控制能力有待提升。信息压缩技术则在输入精简方面优势明显,能有效缓解上下文窗口限制,但其在保证语义完整性与内容相关性方面表现受限,效果依赖于压缩策略的精细设计。

当前检索后增强仍面临三方面挑战:一是语义捕捉与信息保留之间的权衡难度大,规则方法难以覆盖深层语义,压缩机制可能引入信息丢失;二是计算资源开销较高,尤其在模型重排序与压缩解析环节;三是跨任务、跨领域的泛化能力不足,难以适应

多样化的应用需求。为应对上述挑战，已有研究提出若干优化方向：如采用混合策略，通过规则方法进行初步筛选，再结合模型重排提升精度，并在压缩前引入语义权重评估机制，以降低信息损失；利用轻量化模型技术（如蒸馏、量化）降低延迟与成本；引入自适应优化机制（如元学习与多任务训练）提升方法的泛化能力与鲁棒性。未来发展趋势主要集中在三个方向：其一，混合重排序机制将融合规则与模型方法，实现效率与精度的协同优化；其二，自适应压缩机制将根据输入动态调整压缩策略，平衡上下文精简与语义保留；其三，多模态增强路径将整合文本与图像等异构信息源，进一步提升内容质量与任务表现。这些路径将推动检索后增强在复杂问答、文档生成、多轮对话等任务中的深入应用，具有广阔的理论价值与实践潜力。

3.6　大语言模型增强

在LLM与RAG系统深度融合的背景下，增强LLM的预训练、微调和推理能力已成为提升系统性能的核心路径。这些技术旨在解决RAG系统中生成质量、任务适应性和推理深度的关键问题：如何通过检索扩展知识边界、优化任务特异性并提升复杂推理能力？如表9和表10所示，本节围绕预训练增强、微调增强和推理增强三大方向，系统分析其理论基础、技术机制与应用价值，探讨当前挑战与未来趋势。

表9　预训练增强与微调增强技术分析对比

类别	维度/问题	代表方法	核心原理	优势	挑战	适用场景
预训练增强	知识边界（扩展外部知识覆盖）	REALM[25], Atlas[29], UniGen[164]	知识检索与共享编码	知识丰富、语义表示强	计算效率低、语义对接难	开放领域问答、少样本任务
微调增强	任务特异性＋推理深度（优化特定任务）	Yoran等人[165], RoG[166], RAFT[167], RA-DIT[168], Reward-RAG[169], JMLR[170]	领域微调与奖励驱动	任务适配强、推理能力高	过拟合、数据稀缺	多跳推理、医学法律分析

表10　推理增强技术分析对比

类别		维度/问题	代表方法	核心原理	优势	挑战	适用场景
提示工程		推理深度（优化推理路径）	FLARE[177], RAT[178]	CoT与动态检索提示	推理质量高、无需调参	提示复杂性、检索噪声	代码生成、数学推理
Agentic RAG	自主智能体	推理深度＋知识边界（动态规划）	RTLFixer[181], PlanRAG[182], AutoAgent[183], Search-o1[186], RAG-Gym[187]	ReAct与自主决策	动态调整、自主性强	资源消耗大、协作复杂	代码调试、多跳推理
	工作流智能体	任务特异性＋推理深度（协同优化）	DepsRAG[188], MAIN-RAG[189], MMOA-RAG[190], CoA[191]	多智能体协作与任务分解	模块化高效、协作性强	调度复杂、泛化性不足	软件管理、长文本任务

3.6.1　预训练增强

预训练增强通过在LLM预训练阶段融入检索机制，使模型在构建语义表示时动态获取外部知识，从而扩展知识边界并提升任务适应性。其核心在于减少对大规模参数的依赖，增强知识密集型任务的表现。REALM[25]通过在预训练阶段引入知识检索机制，利用交叉注意力机制整合外部检索库的知识，显著提升了开放领域问答和科学研究任务中的推理效率，同时降低计算开销。Atlas[29]则聚焦于少样本学习，通过优化检索器与生成器的协同作用，在法律和医疗等领域的少量样本场景下表现出色。而UniGen[164]通过预训练增强进一步优化RAG任务，采用共享编码器同时训练生成式检索和基于文档的答案生成，并借助Q-Connector和D-Connector弥合检索与生成的语义差距，有效增强LLM对查询和文档的理解并减少噪声干扰。

预训练增强可视为一种“外部记忆增强”机制，其核心在于通过动态知识库扩展，使模型在预训练阶段学习更广泛的语义表示。这一方法不仅提升了LLM对外部知识的理解能力，还通过检索机制弥补了静态预训练的局限性，强化了知识边界扩展和语义丰富性。

3.6.2　微调增强

微调增强通过针对特定任务或领域调整LLM参数，将外部知识内化至模型中，优化任务特异性和推理能力。Yoran等人[165]通过融合Self-ask和自然语言推理（Natural Language Inference，NLI），增强了LLM在多跳推理任务中的鲁棒性，有效过滤噪声

信息，适用于金融预测等场景。RoG[166]结合知识图谱与规划-检索-推理框架，通过微调提升复杂任务的准确性与可解释性，在WebQSP和CWQ多跳问答中Hits@1达85.7%和62.6%，较当时SOTA最高提升11个百分点，显著优于ChatGPT等主流大模型。RAFT[167]针对医学、法律等领域任务引入干扰文档微调，引导模型忽略无关内容，显著提升开卷问答表现，在HotpotQA上准确率达35.3%，较传统微调和RAG基线提升近7倍，在NQ等复杂问答中亦优于常规方法，接近GPT-3.5+RAG水平。RA-DIT[168]通过联合微调检索器和生成器，优化少样本学习和跨领域任务（如医学和法律）的协同效率。Reward-RAG[169]采用奖励驱动监督和强化学习，借助CriticGPT奖励模型优化检索质量，显著提升问答和事实验证任务的生成效果。JMLR[170]通过联合训练检索器与LLM，引入LLM-Rank Loss优化检索策略，在PubMedQA[171]等医学任务中准确率达70.5%，较基础RAG提升2.8个百分点，显著降低训练资源消耗，解释性与事实性均优于主流大模型。IM-RAG[172]引入多轮“内在独白”机制，结合检索与推理的中间过程训练，显著增强多跳推理与信息整合能力，在HotpotQA和MuSiQue[173]中相较原始RAG分别提升EM分数5.3%和F1分数3.9%，有效促进复杂任务中的推理路径建构与结果准确性。RAG-DDR[174]引入可微数据奖励机制，以生成质量反向优化检索器与生成器，实现端到端联合微调。相较原始RAG在NQ、TriviaQA与HotpotQA上分别提升EM得分3.4%、2.6%、3.8%，显著增强了检索-生成一致性与跨任务泛化能力。

微调增强基于任务导向的知识迁移理论，通过任务特定数据调整模型参数，使LLM内化外部知识并优化推理路径。该方法在知识密集型任务中实现高效适配，弥补通用预训练的局限性。

3.6.3 推理增强

在RAG系统中，推理增强旨在通过优化LLM的推理过程，提升其在复杂任务中的准确性、可控性与任务适应性。其核心目标在于克服LLM在多步推理、动态上下文及知识密集型任务中的局限性，例如上下文窗口限制导致的信息丢失、检索噪音干扰生成质量，或自主规划能力的不足。通过外部引导与智能代理机制，推理增强技术优化LLM的推理路径，并赋予其更高的自主决策能力。本节细分为提示工程[175]与Agentic RAG[176]两大技术路径，分别探讨其技术机制、理论基础、应用场景、现阶段挑战及未来发展方向。

（1）提示工程。提示工程通过设计策略性提示，利用LLM的上下文学习能力引导生成过程，从而提升推理能力与输出质量。该方法无需调整模型参数，仅通过输入端优化即可显著改善复杂任务表现。例如，FLARE[177]融合CoT与Few-shot[26]提示，构建基于提示的多轮推理增强机制，动态获取知识并迭代修正答案，在2WikiMultihopQA任务中EM和$F1$值达到51.0%和59.7%，较单轮RAG提升约11%，显著优于多轮检索和问题分解方法。RAT[178]则通过动态检索与CoT推理链逐步修正，实现边推理、边检索、边优化的多步推理机制，有效减少幻觉，在代码生成、数学推理和规划任务中显著提升准确性与稳定性，在HumanEval中pass@1达69.3%，较经典RAG提升约7个百分点，GSMHard[179]准确率提升超31%，全面超越CoT与传统RAG。

提示工程基于LLM的上下文学习与注意力机制，通过外部引导优化内部推理路径。从认知科学角度看，CoT提示类似人类的逐步推理，Few-shot提示则模拟示例驱动的学习模式。动态检索（如RAT）进一步强化检索-推理协同，提升LLM在复杂任务中的可控性与逻辑一致性。

Agentic RAG通过引入智能代理机制赋予LLM自主规划、动态检索和多步推理的能力，突破传统RAG的静态检索与推理分离的局限性。其核心在于通过代理的协作与决策优化“检索-推理-生成”全流程，适用于动态任务、长文本处理和知识密集型推理场景。根据代理协作模式，Agentic RAG可分为自主智能体和工作流智能体两大类。

（2）自主智能体。自主智能体通过“思考-执行-观察”循环实现动态调整，独立完成任务分解、检索触发与推理优化，强调单代理的决策能力。其代表方法包括ReAct[180]框架及其衍生方法。RTLFixer[181]通过ReAct代理的“思考-执行-观察”循环，融合RAG与编译器反馈，构建Verilog语法错误自动修复框架，在修复任务中实现98.5%的成功率。PlanRAG[182]利用有向无环图进行外部规划优化，在HotpotQA等任务中超越ReAct 6%-10%，计算效率提升31%，展现了多跳推理的高效性。AutoAgent[183]提供零代码模块化多代理框架，结合动态任务规划与Web检索，在MultiHop-RAG任务[184]中比LangChain[185]提升10.68%准确率，适用于复杂推理场景。Search-o1[186]通过智能搜索触发与文档推理，动态决策检索时机与策略，在多跳推

理和高难度科学任务中提升6.4%～27.7%的准确率，突破了ReAct的静态模式局限。RAG-Gym[187]引入过程监督优化ReSearch代理，深度融合推理与搜索，在知识密集型任务中提升25.6%成功率，其精准的搜索调整能力优于Search-o1。自主智能体基于强化学习与规划理论，通过“试错优化”循环赋予LLM动态决策能力。其核心优势在于深度融合检索与推理，提升复杂任务的自主性与适应性。

(3)工作流智能体。工作流智能体通过多智能体协作优化RAG全流程，强调模块化分工与协同效率。多个代理分别负责检索、推理与生成，采用流水线或并行方式提升系统性能。DepsRAG[188]将RAG与工作流驱动的多智能体系统相结合，动态规划查询流程并优化软件依赖管理，将查询准确性提升3倍，在软件供应链安全评估中表现优越。MAIN-RAG[189]引入预测、评估和生成等多智能体协作，采用自适应过滤机制筛选高质量文档，在TriviaQA等任务中提升2%-11%准确率。MMOA-RAG[190]通过多智能体强化学习算法，协同优化检索-推理-生成全流程，利用共享奖励机制在HotpotQA等多跳任务中超越ReAct和Self-RAG[100]，实现了动态适应的推理增强。CoA[191]则通过Worker智能体和Manager智能体的流水线协作，优化长文本任务中的上下文利用，在HotpotQA等任务中提升2%-10%准确率，并在8K tokens内超越200K tokens的Claude-3，解决了信息丢失问题。工作流智能体基于分布式计算与协作优化理论，通过任务分解与模块化协作提升RAG的推理效率。其核心在于多智能体分工实现全局优化，弥补单代理在复杂任务中的局限性。

自主智能体适用于动态任务(如代码调试)，通过灵活决策优化推理深度；工作流智能体适用于结构化任务(如长文本分析)，依靠多代理协作提升处理效率。自主智能体决策灵活但执行效率受限，而工作流智能体支持并行但依赖模块协同。

LLM增强聚焦于知识扩展、任务适配与推理能力的提升，其核心维度包括知识边界、任务特异性及推理深度。技术分析显示，预训练增强通过大规模语料扩展模型的知识覆盖范围，在通用任务中表现稳定，但推理能力较依赖外部检索机制。微调增强通过在特定任务上的有监督训练，显著提升模型的任务适配性与推理深度，但知识边界受限于训练数据质量与范围。提示工程利用精设计的输入提示引导模型生成，优化推理路径与响应质量，在适配性与知识利用方面效果可控性强，但依赖提示设计经验。自主智能体方法通过自主规划与反思机制提升推理深度与知识联想能力，在开放式任务中展现潜力，但通用性与适配性尚需提升。工作流智能体通过多智能体协同完成复杂任务，强化任务分解与信息整合能力，适配性与推理能力较强，但整体知识覆盖仍受限于协作流程设计。

当前LLM增强面临三大挑战：一是资源消耗高，如大型模型预训练与智能体规划任务计算开销大；二是数据稀缺问题显著，尤其微调依赖高质量标注数据，限制模型泛化能力；三是推理一致性与跨任务泛化性不足，易出现逻辑断裂或幻觉。为应对这些挑战，已有研究提出多项应对策略。包括通过模型优化(如模型蒸馏、结构稀疏化)降低计算成本；利用数据增强方法(如合成数据生成、迁移学习)缓解标注依赖；以及探索协同设计机制，结合NLI与元学习，提升推理稳定性与多任务泛化能力。未来的发展趋势将集中于三个方向：其一，多模态融合机制将整合文本、图像等异构信息源，进一步拓展知识边界与上下文理解能力；其二，自适应增强框架根据任务需求动态切换增强方式，提升系统适配性与稳健性；其三，高效计算支持(如轻量化架构与硬件加速)将降低模型运行成本，推动LLM增强技术在复杂、高频任务中的落地应用。这些路径有望为LLM增强在多样化、高要求场景中的深入应用提供理论支撑与实践基础。

4　增强环节对比分析

第3节以RAG的工作流程为导向，系统剖析了各环节的增强技术，着眼于开发者优化性能并提升应用效果。为此，明确RAG技术的核心目标是理解其优化方向的关键。RAG本质上旨在弥补LLM的固有局限性。尽管LLM在自然语言处理任务中表现出色，但其基于概率预测的生成机制易产生“幻觉”，即生成错误或缺乏依据的内容。此外，LLM的知识依赖训练数据，难以实时更新，且在特定领域或专有数据场景中存在知识盲区。为解决这些问题，RAG依托两大核心策略：外部信息检索与LLM能力优化。前者通过获取高质量外部知识，弥补知识滞后与专业领域覆盖不足；后者提升LLM对检索信息的理解与利用能力，增强生成内容的准确性与时效性。两策略相辅相成，共同构筑RAG的优化路径。基于此，本节将增强技术分为“检索增强”和

“LLM增强”两大类，清晰阐述其作用与优化方向。这种分类不仅便于开发者针对性地提升RAG系统性能，也深化了对优化外部检索机制与强化语言模型能力协同作用的理解。通过此双管齐下的方式，RAG有效克服LLM的局限，提升AI生成系统的准确性、智能化水平及应用价值。

4.1 检索增强与LLM增强

在RAG检索系统中，检索增强技术依据其在信息处理流程中的阶段与作用，可分为“输入端增强”、“检索端增强”和“输出端增强”三个层次。这一分类框架源于各阶段在检索流程中承担的独特目标与需求。具体而言，输入端增强通过优化用户查询的语义表达与初步信息提取，提升系统对查询意图的理解精度，为后续检索奠定基础；检索端增强聚焦于优化检索器设计、索引结构及检索策略，确保检索过程兼顾高效性与精确性，为生成任务提供信息支持；输出端增强则通过筛选、重排序及信息压缩等后处理技术，进一步提升检索结果的质量与适用性，确保检索内容在相关性与准确性上达到最优。这一分层视角清晰揭示了各层次技术在检索流程中的功能定位及其对系统整体性能的贡献，为RAG系统的优化设计提供了实践指南。

输入端增强技术通过优化用户查询的语义表达与信息提取，为RAG系统提供高质量输入，表11展示了输入端各技术之间的指标对比。其核心包括输入增强与查询增强：前者通过元数据提取和语义提取优化信息定位；后者通过重写与扩展优化查询深度，确保系统精准捕捉用户需求。两者互补，结合高效信息抽取与动态查询优化，构建完整的输入端增强体系。

表11 输入端增强（指标数值为%，Acc表示准确率）

类别	研究	年份	关键指标		补充说明
			HotpotQA（EM）	FreshQA[192]（Acc）	
预检索增强	RaFe[52]	2024	46.79（Acc）	63.01	基于Qwen1.5-32B和Google搜索，结合bge-reranker重排序，查询重写训练仅耗时0.67小时，成本低、易部署，实际部署中使用2-3个重写实现性能与效率最优平衡。
	DMQR-RAG[53]	2024	41.12	**76.67**	基于GPT-4和Bing搜索，结合bge-reranker重排序，通过并行检索将平均重写数减少40%，在提升检索质量的同时有效控制计算开销。
	RQ-RAG[54]	2024	40.32	69.67	基于LLaMA3-8B和text-embedding-3-large，通过控制树搜索深度（≤4）、均衡查询操作分布，实现仅用40 K数据在小模型上逼近GPT-4的多跳问答性能。
	EVO-RAG[55]	2025	**57.6**	-	基于Qwen-3-8B，采用融合BM25与BGE embedding的RRF-BGE检索方法，在动态奖励调度作用下，平均减少约15%检索深度，10.4步内达到最优EM表现。

检索端增强技术通过优化检索器设计、检索策略及索引结构，确保从文档库提取的信息高效、精准并与LLM生成需求深度对齐。其技术路径主要涵盖三个维度：检索器优化，通过稀疏、稠密和生成式检索提升性能，并结合微调、适配器和偏好对齐策略进一步增强效果；检索策略增强，采用多步检索、动态调整和上下文扩展等方法提升灵活性与相关性，表12展示了检索策略增强各技术之间的指标对比；索引增强，聚焦数据源质量与索引结构优化，提升信息组织与访问效率，表13展示了索引增强各技术之间的指标对比。这些路径相辅相成，检索器优化奠基，策略增强提升适配性，索引优化保障高效访问，共同构建检索端增强体系。

输出端增强技术通过精细化后处理优化检索内容的质量与结构，旨在克服LLM上下文窗口限制及检索噪声干扰，表14展示了输出端增强各技术之间的指标对比。其核心技术路径分为重排序与过滤和信息压缩两大维度。重排序与过滤采用基于规则和基于模型的方法，提取最具针对性信息，削减冗余并提升相关性；信息压缩通过嵌入压缩和语义提炼，精炼检索内容并保持关键信息完整性，优化长文本处理效率与生成质量。两者协同作用，重排序与过滤保障信息针对性，信息压缩提升输入精炼度，共同构建输出端增强体系。

LLM增强技术通过优化预训练、微调及推理三大环节，全面提升模型的生成质量、任务适配性与推理能力，表15展示了LLM增强各技术之间的指标对比。其技术路径包括三大维度：预训练增强通

表12　检索策略增强（指标数值为%，Acc表示准确率，2Wiki为2WikiMultihopQA的缩写）

研究	年份	关键指标				补充说明
		HotpotQA		2Wiki		
		EM	F1	EM	F1	
Auto-RAG[86]	2024	-	44.9	-	48.9	基于LLaMA3-8B和E5-base-v2，平均迭代2-3次、推理时间约8秒，性能显著优于FLARE和Self-RAG等同类方法。NQ(EM)：37.9%；TriviaQA(EM)：60.9%
AirRAG[90]	2025	**75.2**(Acc)	79.6	**74.2**(Acc)	**76.0**	基于Qwen2.5-14B和multilingual-e5-base，结合MCTS递归检索与树状推理，自适应控制检索深度，在受控推理代价下平均F1值提升超17%
SiReRAG[93]	2025	61.70	76.48	59.60	67.94	基于GPT-4o和text-embedding-3-small，平均推理时间约2.3秒，时间-池效率比约0.5，表示检索池扩展3倍时推理仅增1.5倍
RAG-Star[95]	2024	48.0	68.6	48.0	61.7	基于GPT-4o，结合BGE-large-en-v1.5和FAISS[193]，采用可控模拟次数与深度的MCTS，在50次模拟后趋于收敛，推理效率优于传统树搜索方法
CtrlA[96]	2024	34.7	44.9	36.9	43.7	基于Mistral-7B[194]和BM25，平均检索频率仅3.28次，显著低于同类方法如FLARE，生成更精确且无明显额外计算成本
Vendi-RAG[99]	2025	42.2	57.0	47.2	58.9	基于GPT-3.5 Turbo和all-mpnet-base-v2，采用多轮自适应迭代策略，embedding与召回高效，整体推理延迟与传统RAG相当
Open-RAG[102]	2024	66.2	**80.1**	60.7	70.9	基于LLaMA2-13B和Beam Retrieval[195]，结合自适应检索频率与MoE部分专家激活机制，实现优于同类方法的推理速度
DeepRAG[103]	2025	40.7	51.54	48.10	53.25	基于LLaMA3-8B和BM25，通过步骤模拟选择最优路径，将平均检索次数降至0.92，显著低于Auto-RAG等方法，同时提升准确率
Sufficient Context[104]	2025	67.5(Acc)	-	-	-	基于Gemini 1.5 Pro和e5-base-v2，支持高效全流程API推理与大批量处理，6000 token级拼接实现高吞吐低延迟
Adaptive-RAG[105]	2024	42.0	53.82	40.60	49.75	基于FLAN-T5-XL-3B[196]和BM25，平均推理时间小于1.5秒（复杂查询约27秒），分类器仅需60M小模型部署，推理速度较多步方法快2-3倍
MBA-RAG[106]	2025	40.60	52.44	49.40	58.33	基于FLAN-T5-XL-3B和BM25，平均检索步数仅1.80，较Adaptive-RAG节省20%以上开销，依靠精准策略匹配，在性能与成本上实现优越平衡

表13　索引增强（指标数值为%，2Wiki为2WikiMultihopQA的缩写）

研究	年份	关键指标				补充说明
		HotpotQA		2Wiki		
		EM	F1	EM	F1	
KET-RAG[134]	2025	35.2	47.7	-	-	基于GPT-4o-mini和text-embedding-3-small，采用KG Skeleton与Keyword Graph的双通道复合检索，索引成本较Graph RAG降低90%，推理性能与高成本方法持平或更优
HippoRAG[135]	2024	52.6	63.5	**65.0**	**71.8**	基于Llama-3.3-70B和Contriever，单步检索即可完成多跳推理，较多轮方法快6-13倍、成本低10-30倍，具备高效性与可扩展性
HippoRAG-2[136]	2025	**62.7**	**75.5**	**65.0**	71.0	基于Llama-3.3-70B和Contriever，采用单步图检索与一次性过滤，在保持与传统RAG相当推理时延的同时显著提升准确率，支持并行与大规模工业部署
ToG-2[140]	2025	42.9	-	-	-	基于GPT-3.5-Turbo，结合结构化KG多跳检索与文本上下文检索（BGE系列），平均推理时间27.3秒、5.4次调用，运行时间为ToG的39%，调用次数下降66%
KG-Retriever[144]	2024	32.8	-	35.0	-	基于Qwen1.5-7B，结合层级图索引（KG层+文档图）与多策略协同检索，平均推理延迟低于1.5秒，性能优于所有迭代式方法，推理速度提升6-15倍

过融入检索机制扩展知识边界；微调增强通过任务或领域特定参数调整，将外部知识内化至模型；推理增强则依托提示工程及Agentic RAG，优化推理路径并赋予自主决策能力。三者相辅相成，预训练增强奠定知识基础，微调增强确保任务精准适配，推理增强通过动态优化提升逻辑一致性与生成质量，共同构筑LLM增强体系。

4.2　检索增强与LLM增强对比分析

在RAG系统中，检索增强与LLM增强作为两大核心技术路径，分别从外部知识输入和内在模型

表14 输出端增强(指标数值为%,Acc表示准确率)

类别	研究	年份	关键指标		补充说明
			HotpotQA(EM)	NQ(EM)	
检索后增强	FiD-Light[149]	2023	29.2	51.1	基于FLAN-T5-XL和GTR-Base,结合生成来源定位与重排序机制,根据生成器输出的段落索引调整排序,实现解码延迟下降3倍、总延迟低于100 ms,性能优于基础RAG方法
	RADIO[151]	2024	-	36.65	基于Llama-3.1-8B和e5-base-v2,结合推理蒸馏、cosine重排序与InfoNCE[197]微调,兼容bge模型,训练阶段可并行生成推理理由,推理阶段仅需重排序器与生成器,无额外开销
	RankRAG[155]	2025	42.7	54.2	基于Llama3-RankRAG-70B(Zero-shot)和Dragon[198]检索器,采用指令微调的生成一体化重排序模型,无需独立reranker,数据需求低、效率高,实际部署中在Top-N=20~30时实现准确率与延迟的优异平衡
	RankCoT[156]	2025	33.91(Acc)	49.98(Acc)	基于Qwen2.5-14B和BGE-large Embedding,结合答案包含性进行排序与自我优化,生成结果更简洁,推理输入精简,速度更快,token成本更低
	ASRank[157]	2025	42.6	27.5	基于Llama-3-70B和DPR,结合零样本、高效可扩展的答案线索+T5重排序,ASRank实现一次大模型生成配合小模型批量重排,推理快速、成本仅约$15,显著优于纯大模型重排序,具备工业落地能力
	Refiner[160]	2024	**67.2**(Acc)	-	基于LLaMA-3-8B和Contriever,Top-10检索构建文档上下文,实现90.5%的token压缩率,输出仅为LongLLMLingua[199]的1/5,LLM利用率高达96%
	COCOM[161]	2025	43.0	**55.4**	基于Mistral-7B和SPLADE-v3[200],结合Top-5检索与DeBERTa-v3[201]重排序构建文档上下文,实现解码提速最高5.69倍,GFLOPs降低22倍,显存占用减少约1.27倍

能力两个维度优化系统性能,协同应对LLM的固有局限性。如表16所示,本节通过对比分析两者的技术路径、作用机制、优势与局限性以及适用场景,揭示其差异性与互补性。

4.2.1 技术路径与作用机制

检索增强通过外部知识的获取与整合优化生成过程,其技术路径覆盖输入端、检索端和输出端三个阶段。其作用机制依赖外部资源,通过弥补模型知识盲点增强生成内容的准确性与上下文相关性,体现"以外补内"的优化哲学。

与之相对,LLM增强通过提升模型内在能力优化性能,其技术路径包括预训练增强、微调增强和推理增强。其作用机制强调"以内强内",通过强化模型的学习、适应与推理能力减少外部依赖,体现从根本提升自主性能的优化哲学。

检索增强与LLM增强的技术路径体现出截然不同的优化哲学:检索增强通过外部知识补充克服模型知识局限,突出动态性与扩展性;LLM增强则通过内在能力提升强化模型独立性,强调长期性与稳定性。这一本质差异决定了其作用机制的侧重,但两者并非互斥,而是互补共存。在实际应用中,可融合检索增强的外部知识支持与LLM增强的内在推理能力,构建兼具知识广度与深度的智能系统,从而实现更优的生成效果。

4.2.2 优势与局限性分析

检索增强通过动态引入外部知识提升生成质量,其优势体现在三个方面。首先,其知识时效性与覆盖广度确保模型实时获取最新信息并扩展知识边界,如RET-LLMs[131]利用可读写记忆模块和三元组存储,在时效性问答中优于Alpaca-7B[204];基于迭代效用最大化(IUM[89])的搜索引擎优化方法,通过实时反馈与在线学习,有效增强模型的信息时效性与知识边界扩展能力,在NQ、TriviaQA等数据集上显著提升了EM得分。其次,减少"幻觉"通过外部知识约束模型输出,提升事实准确性,如CtrlA[96]引入内在控制机制,通过诚实性引导和置信度监测触发检索,在2WikiMultihopQA[87]上实现了EM得分36.9、F1得分43.7的优异表现;WeKnow-RAG[97]结合置信度自评、多阶段检索和知识图谱重排名,在CRAG[98]上将准确率从39.3%提升至40.9%;Self-RAG[100]通过自反思和反思标记评估输出,在PopQA[101]和PubHealth[205]任务中准确率分别达55.8%和73.5%,显著提升事实准确性与可解释性。最后,多领域扩展性得益于模块化设计,如

表15　LLM增强（指标数值为%，Acc表示准确率）

类别	研究	年份	关键指标		补充说明
			HotpotQA（EM）	NQ（EM）	
预训练增强	REALM[25]	2020	-	40.4	基于BERT-base，模型仅330M，较T5-11B小33倍却效果更优，适配单机GPU，推理仅需Top-5检索文档，效率高且部署轻量
	Atlas[29]	2022	-	42.4	基于FLAN-T5-XL和Contriever，采用解码融合机制与困惑度蒸馏策略，在仅11B参数下准确率超越540B模型，64-shot精度大幅领先，支持索引压缩与仅查询端更新，实现动态知识迁移
	UniGen[164]	2024	-	**57.83**	基于T5-base和生成式检索，模型仅367M，无需外部索引，迭代优化仅两轮，推理速度较快
微调增强	RA-DIT[168]	2024	40.7（Acc）	43.9（Acc）	基于LLaMA-65B和DRAGON检索器，支持多chunk并行推理，无需拼接长上下文，比REPLUG更快且更具鲁棒性
	Reward-RAG[169]	2024	-	50.9	基于GPT-4和E5-large-unsupervised[202]，无需修改LLM，仅训练小型检索器，并用GPT-4o替代人工标注，显著节省成本与资源
	IM-RAG[172]	2024	**68.4**	-	基于Vicuna-1.5-7B和DPR，结合RankVicuna-7B重排序模型与LoRA[203]高效推理，检索仅0.061秒/次，最多3轮，端到端效率高，适合复杂多跳任务但延迟非极低
	RAG-DDR[174]	2025	39.0	52.1	基于Llama3-8B和bge-large，采用“检索→精炼→生成”流程，LLM快速判别提升效率，整体时延与基础RAG相当，推理高效、泛化强，易于大规模落地
推理增强	PlanRAG[182]	2025	40.44（Acc）	-	基于LLaMA3.1-8B和Contriever，通过并行执行子查询并保持恒定上下文大小，平均token使用量较低
	Search-o1[186]	2025	45.2	34.0	基于QwQ-32B-Preview和Bing Web Search API，采用Agentic检索、上下文精炼插入与批处理，支持并行与模块化的高效推理流程
	RAG-Gym[187]	2025	44.1	-	基于LLaMA-3.1-8B，结合BM25与BGE-Base，采用答案驱动式查询与10～20次动作采样，实现精准检索，推理效率高且泛化性强
	MMOA-RAG[190]	2025	36.15	-	基于LLaMA-3-8B-Instruct和Contriever，采用模块共享与强化学习联合优化，训练快速收敛，推理阶段无额外开销

表16　增强技术对比

维度	检索增强	LLM增强
技术路径	输入端增强：优化查询表达与信息提取 检索端增强：改进检索器设计、检索策略及索引结构 输出端增强：重排序、过滤与信息压缩	预训练增强：融入外部知识扩展模型能力 微调增强：调整参数适配特定任务 推理增强：改进提示与智能代理增强推理能力
作用机制	通过实时检索外部知识库，为模型提供动态、最新的事实依据，弥补知识时效性和广度的不足	通过优化模型参数和推理策略，提升内在语义理解与逻辑生成能力，增强自主处理复杂任务的效率
优势	- 知识时效性：能够动态获取最新信息，支持实时更新需求 - 知识覆盖广度：通过外部知识扩展模型信息边界 - 减少幻觉：提供事实依据降低生成错误	- 推理能力提升：增强复杂任务中的逻辑一致性 - 任务特异性强：适配特定领域的高精度需求 - 动态适应性：通过推理优化灵活应对多变任务
挑战	- 依赖外部数据质量：受知识库准确性和完整性限制 - 计算开销高：大规模检索增加资源消耗 - 实现复杂性：需平衡检索精度与效率	- 数据依赖性：微调需高质量数据支持 - 计算负担重：复杂推理增加处理时间 - 优化难度：推理策略设计需针对性调整
适用场景	- 时效性任务：如新闻生成，需快速更新信息 - 知识密集型任务：如开放域问答，需广泛知识支持 - 多领域任务：如专业文献检索，需灵活适配	- 推理密集型任务：如多跳推理，需逻辑深度 - 领域特异性任务：如医疗问答，需高精度 - 动态决策任务：如代码调试，需自主规划

DocReLM[80]在量子物理和计算机视觉领域准确率远超Google Scholar，展现了跨领域的适配能力。然而，检索增强的局限性包括对外部数据质量的高度依赖，低质量数据可能引入噪声；大规模检索的计算开销增加部署难度，如SiReRAG[93]在多跳任务中$F1$值提升1.9%，但推理时间较长；以及实现复杂性提高优化门槛，如DPA-RAG[83]虽准确率提升三倍，却因调参复杂而不易推广。这些不足表明，检索增强需在数据质

量、计算效率和设计简洁性上进一步优化。

LLM增强通过提升模型内在能力优化性能，其优势体现在三个方面。首先，推理能力提升显著改善复杂任务表现，如RoG[166]在CWQ[138]上准确率提升22.3%，并超越GPT-4，展现卓越多跳推理能力。其次，任务特异性通过微调实现定制化优化，如Reward-RAG[169]在PubMedQA[171]和BioASQ[206]等专业场景表现突出，满足高精度需求。最后，自主性与动态适应性使模型灵活应对动态任务，如RTLFixer[181]在Verilog修复中成功率达98.5%，PlanRAG[182]在HotpotQA[88]上较ReAct提升6%-10%的准确率，并提高31%的执行效率，展现出智能代理在决策优化中的显著优势。然而，LLM增强的局限性包括数据依赖性强，如RoG依赖高质量知识图谱，数据稀缺时效果受限；计算负担沉重，如Search-o1[186]在GPQA[12]等任务上准确率提升4.7%，但多轮检索和精炼增加耗时，高频搜索进一步推高成本，限制大规模应用；以及优化难度高，如DepsRAG[188]通过多代理和Agent-Critic机制提升3倍准确率，却因任务拆解与反复调整导致成本高昂，过多交互甚至降低性能，需平衡效率与效果。

4.2.3 适用场景与互补性

在知识密集型、时效性强和多领域专业任务中，检索增强展现出显著优势。DMQR-RAG[53]在AmbigNQ[207]和HotpotQA上表现优异，适用于开放域问答；RAGtrans[121]在翻译任务中BLEU[115]得分提升3.09，增强了跨语言生成能力。Auto-RAG[86]在FreshQA[192]上通过迭代检索实现知识更新，适用于新闻生成与实时分析。DocReLM[80]在学术检索中达44.12%准确率，适配量子物理等专业领域，支持多知识源切换，满足复杂任务需求。

在推理密集型、少样本特定领域和动态决策任务中，LLM增强展现出广泛适应性。FLARE[177]在2WikiMultihopQA上获得EM得分51.0，适用于多跳问答；RTLFixer在代码调试中修复率达98.5%，提升推理深度，适合数学建模与代码生成。Reward-RAG在PubMedQA上表现优异，通过微调或提示适配医学问答与金融分析等数据稀缺场景。PlanRAG在HotpotQA上检索效率提升31%，借助智能代理优化复杂任务规划，适用于软件调试与动态环境。

检索增强与LLM增强的适用场景虽有交集，但在复杂任务中协同作用往往能实现最佳效果。例如，在法律分析中，检索增强提供最新法规数据，确保知识时效性与覆盖度，而LLM增强优化条文推理，保障逻辑严密性，二者结合提升分析的全面性与可靠性。同样，在医疗诊断中，检索增强获取最新文献，LLM增强分析症状与文献关联，共同提高诊断精度。这种互补性源于技术侧重：检索增强弥补知识边界与时效性不足，LLM增强强化推理深度与任务适配能力。因此，在知识密集且推理复杂的场景中，二者协同可最大化生成质量与任务效率。

理论上，检索增强从“输入优化”角度通过外部知识动态扩展弥补LLM静态训练缺陷，LLM增强从“处理优化”角度提升语义处理与推理能力，形成“知识-智能”的协同闭环。实践上，时效性任务优先检索增强，推理任务侧重LLM增强，高精度场景则整合二者，通过优化检索策略与推理机制实现性能最优。

5 数据集与评估指标

5.1 数据集

数据集是RAG技术发展的基石，为模型训练与评估提供核心资源，推动其在多场景任务中的应用。本节系统梳理RAG领域常用数据集，涵盖特定领域问答、开放领域问答、结构化数据问答、推理问答、事实验证问答、长篇问答、低资源任务及对话式搜索问答八大类别，如表17所示。

(1) 特定领域问答。特定领域问答数据集聚焦医疗、法律、金融、API检索及代码编程等专业领域，旨在通过RAG精准处理术语并支持复杂推理。MedMCQA[208]提供193,155道多选题，覆盖21个学科，基准准确率47%，挑战诊断与因果推理；LawBench[209]基于中国民法设计20项任务，评估认知层次表现；FinTextQA[210]含1,262组问答对，覆盖金融子领域；ToolBench[79]提供16,464个API指令，支持多工具场景；CodeSearchNet[211]约600万函数-语言配对，挑战语义匹配。

(2) 开放领域问答。开放领域问答数据集覆盖广泛主题，评估RAG在大规模文档中的检索与生成能力。HotPotQA[88]含113,000个多跳问答，强调跨文档推理；PopQA[101]聚焦长尾实体，验证自适应检索效率；NQ[68]提供307,373个真实查询，主题多样；TriviaQA[69]包括超650,000问题-证据对，40%需跨句推理。

(3) 结构化数据问答。结构化数据问答数据集旨在测试RAG系统从表格或知识图谱提取信息的能力。FeTaQA[126]含10 330个表格问答对，挑战语义提炼与答案流畅性；WEBQSP[139]含5 810个问

表17　数据集汇总

任务类型	数据集名称	时间	主要特点	SOTA方法
特定领域问答	MedMCQA[208]	2022	医疗领域多选题，193 155个问题，覆盖21个医学科目，挑战深度推理	Med-PaLM2[215]在MedMCQA上达到72.3%准确率
	LawBench[209]	2024	中国民法法律任务评估，20个任务，评估模型在法律知识记忆、理解与应用	InternLM-Law[216]在LawBench上平均得分达67.71，在20个任务中14项夺冠，超越GPT-4，表现最佳
	FinTextQA[210]	2024	金融领域的长文本问答数据集，支持超长文档推理	FinTextQA[210]中Baichuan2-7B的RAG系统的ROUGE-L得分0.204，GPT-4评分为4.51/5
	ToolBench[79]	2024	API检索任务，16 464个API，增强LLM工具使用能力	ToolLLaMA[79]配合专用检索器在ToolBench的API检索任务上NDCG@5达84.9
	CodeSearchNet[211]	2019	代码检索，600万函数，六种编程语言，支持语义代码搜索	Cpt-code[217]在代码检索任务中实现MRR达93.5%，相比GraphCodeBERT[218]提升超20%
开放领域问答	HotPotQA[88]	2018	多跳问答，113 000个问题，强调多文档推理和解释性答案	Beam Retrieval[195]方法的问答EM分数达72.69%，*F*1达85.04%，显著优于以往方法，兼顾性能与效率
	PopQA[101]	2023	长尾实体问答，评估模型在不常见实体上的表现	MAIN-RAG[189]在PopQA上，准确率达64.0%，显著优于其他微调基线，表现最佳
	NQ[68]	2019	307 373个真实用户查询，提供长短答案，评估开放域问答	在NQ上，Atlas[29] (11B, Full-data)方法的EM达64.0%，显著优于现有方法
	TriviaQA[69]	2017	650 000个问题-答案-证据三元组，40%问题需跨句推理	在TriviaQA上，COCOM[161]方法在压缩上下文条件下实现EM达85.9%，接近完整上下文RAG性能，且推理速度提升约5.7倍
结构化数据问答	FeTaQA[126]	2022	表格问答，10 330个问题-答案对，生成自由文本答案	在FeTaQA上，CABINET[124]生成自由文本答案的SacreBLEU[219]得分达40.5，相比最佳基线提升超5.6分，表现最优
	WEBQSP[139]	2016	知识图谱问答，5810个问题，附带SPARQL查询	在WebQSP上，PoG[220]方法的EM(Hits@1)达96.7%，显著优于所有已有方法，表现最优
推理问答	2WikiMultihopQA[87]	2020	多跳问答，192 606个问题，结合维基百科和Wikidata	在2WikiMultiHopQA上，AirRAG[90]方法实现准确率74.2%、F1 76.0%，显著优于Auto-RAG等方法，表现最优
	MuSiQue[173]	2022	多跳推理，25 000个问题，确保连贯推理	在MuSiQue上，Online Finetuned Flow[221]在测试集上实现*F*1得分69.3%，为当前最佳公开结果
事实验证问答	Feverous[213]	2021	事实验证，87 026个声明，结合文本和表格证据	在FEVEROUS上，ClaimPKG[222]准确率达到83.8%，表现最优
长篇问答	ASQA[214]	2022	歧义问题长篇问答，6316个问题，生成详细回答	在ASQA上，GR-RAG[223]准确率达72.8%，全面优于传统RAG方法，表现最优
低资源任务	BEIR[63]	2021	Zero-shot信息检索，9个任务，18个数据集，评估跨领域能力	在BEIR上，CRISP[224]在主任务上实现NDCG@10＝54.5，相较原始ColBERT提升效果同时大幅减少计算成本，表现最优
对话式搜索问答	QReCC[114]	2021	开放域对话问答，80 000个问题，支持问题重写和多轮问答	在QReCC上，ADACQR[117]方法实现NDCG 52.5、Recall@10 76.5、Recall@100 93.7，在稀疏和密集检索设置下全面优于所有同类方法
	Wizard of Wikipedia[20]	2019	知识驱动对话，22 311个对话，基于Wikipedia生成回应	在Wizard of Wikipedia上，Ms. WoW-Gold[118]方法实现*F*1得分0.321，通过精确选择金标准知识显著提升对话生成质量，表现最优

题，81.5%可通过Freebase[212]解析，推动结构化查询应用。

(4)推理问答。推理问答数据集聚焦于测试RAG模型的多跳推理能力。2WikiMultihopQA[87]含192 606个问题，融合多源数据，挑战整合难度；MuSiQue[173]含25 000个可回答问题，确保连贯推理，推动逻辑严密性发展。

(5)事实验证问答。事实验证问答数据集要求RAG模型验证声明并检索支持证据，强调事实核查能力。Feverous[213]含87 026个声明，跨结构化与非

结构化数据推理，推动事实核查能力研究。

（6）长篇问答。长篇问答数据集旨在检验RAG系统生成详细且连贯回答的能力，强调完整性与语义连贯性。ASQA[214]含6316个歧义问题，注重流畅性与可追溯性，推动复杂问答研究。

（7）低资源任务。低资源任务数据集测试RAG模型在数据稀疏场景下的表现，聚焦跨领域泛化能力。BEIR[63]含19个子数据集，覆盖9类任务，评估检索与适应性，推动低资源优化。

（8）对话式搜索问答。对话式搜索问答数据集支持多轮交互与知识驱动的回答，旨在提升RAG在动态对话中的应用。QReCC[114]含80 000个问题，验证重写效果；Wizard of Wikipedia[20]含22 311个对话，推动知识选择与生成能力。

这些数据集为RAG提供多样化支持，推动其在检索与生成任务中的进步，并为未来数据集开发与应用拓展奠基。随着技术演进，RAG在更多领域的潜力将进一步释放。

5.2 评估指标

RAG系统性能评估采用多维度指标体系，如表18所示，涵盖分类正确性、检索与生成平衡、生成质量、检索排序、证据支持及文档覆盖等方面。分类与正确性指标如准确率（Accuracy）和精确匹配（Exact Match，EM）分别衡量总体预测精度和答案一致性，适用于二分类及短答案任务；检索与生成平衡指标（如$F1$值、精确度、召回率）综合评估准确性与覆盖度，适合不平衡数据集或双目标场景；生成质量指标（如BLEU[115]、SacreBLEU[219]、ROUGE-L[49]）通过n-gram重叠或最长公共子序列分析文本简洁性、相似性及连贯性，应用于短文本或长文本生成；检索排序指标（如NDCG、MRR、MAP（Mean Average Precision，MAP））聚焦相关性排序与定位效率，适用于多候选或信息检索任务；证据支持指标[213]（如Evidence Coverage Score）和文档覆盖指标（如Document and Passage Coverage）分别评估证据完整性与多源信息覆盖度，支撑复杂推理及多文档

表18 评估指标汇总

指标类型	指标名称	指标定义	适用场景
分类/正确性	准确率（Accuracy）	正确预测样本占总样本比例，评估答案或分类正确性	二分类或多分类任务，如判断答案正确性
	精确匹配（EM）	检查生成答案与参考答案是否完全一致	短答案或实体类问答任务
检索与生成平衡	$F1$值	综合精确度与召回率，平衡答案正确性与完整性	不平衡数据集，兼顾精确度与召回率任务
	精确度（Precision）	预测正确答案中实际正确的比例，关注结果准确性	多分类或检索任务，确保答案准确
	召回率（Recall）	实际正确答案中被预测正确的比例，关注信息覆盖度	需覆盖所有相关信息的任务
生成质量	BLEU[115]	比较生成文本与参考文本的n-gram重叠，评估简洁性	短答案生成任务，衡量生成质量
	SacreBLEU[219]	BLEU改进版，标准化评估生成文本与参考文本相似性	短答案生成任务，提供一致性评估
	ROUGE-L[49]	基于最长公共子序列，评估生成文本的连贯性与覆盖度。	文本生成或摘要任务，注重流畅性
检索排序	NDCG	根据相关性与位置重要性评估检索结果排序质量	多候选排序任务，需按相关性排序
	MRR	首个相关结果排名倒数的平均值，衡量快速定位能力	排序任务，需快速找到相关答案
	MAP	每个查询的平均精度均值，综合评估检索排序质量	信息检索任务，关注整体精度
证据支持	Evidence Coverage Score[213]	评估检索证据覆盖必要信息的比例，关注完整性	复杂推理或多跳问答，需整合多源证据
文档覆盖	Document and Passage Coverage[213]	评估检索文档或段落覆盖任务所需信息的比例或计数	多文档检索任务，需从多源提取信息

场景。这些指标共同构建RAG评估框架，推动其在多样化应用中的优化与发展。

6 讨　论

本节首先重点探讨RAG技术与上下文窗口扩展技术的关系，并阐述RAG技术存在的必要性。随后，深入分析RAG技术的当前局限性。尽管RAG在众多应用场景中展现出独特优势，但其在实际操作中仍面临诸多挑战。鉴于技术的持续进步和应用领域的不断拓展，对RAG的深入研究显示出广阔的发展潜力。

6.1　RAG与上下文扩展

近年来，LLM的上下文窗口从早期的4K-8K token扩展至当前的128K甚至更高（如Claude3的200K[3-4]、Gemini 1.5的1M token[225]），显著提升了长文本处理能力。Gemini 1.5提出的“大海捞针”功能[225]（在海量文档中精准检索信息）引发了对外部检索机制必要性的讨论。然而，尽管上下文窗口扩展增强了模型能力，RAG仍因其独特价值在特定场景中不可替代。本节分析RAG与上下文扩展的关系，探讨其互补性与局限性。

长上下文窗口指LLM单次处理的最大文本长度，其提升改善了复杂对话、长文档理解及深层分析的表现。例如，Gemini 1.5 Pro在百万级token中表现优异，国内Kimi支持200万字上下文。但U-NIAH[226]研究表明，长达128K token的上下文仍受“中间遗忘”影响，信息提取效率下降（标准差2.01），且计算成本、推理时间随长度增加显著上升，尤其在实时应用中成本劣势明显。更重要的是，长上下文窗口并不能完全弥补模型训练数据的固有缺陷。奠定人工智能理论基石的哥德尔不完备性定理表明，任何人工智能系统都无法同时保持一致性与完备性，而大语言模型的“幻觉”问题可能正是不一致性的必然体现。

相比之下，RAG通过结合外部知识库与生成模型，提供了一种动态提升模型能力的方法。U-NIAH工作指出，RAG在长上下文复杂任务中的表现优于单纯依赖长上下文窗口的模型，例如在多针任务中平均得分达9.04（标准差1.48），胜率高达82.58%，而在64K-128K token的超长上下文中胜率进一步提升至92.7%。LaRA[227]基准测试也表明，RAG对较弱模型（如Llama-3.2-3B-Instruct）的性能提升尤为显著，在128K上下文时准确率比长上下文模型高出6.48%。这些数据凸显RAG通过提升上下文窗口内“有效信息密度”，显著增强生成内容的精确性与相关性。此外，RAG的灵活性与成本效益是其另一关键优势：相较于长上下文窗口需一次性处理全部输入，RAG仅检索任务相关的少量文档，大幅降低计算开销；在实时更新或专业领域应用中，RAG能动态引入最新外部信息，有效弥补模型知识不足。例如，在幻觉检测任务中，RAG胜率比长上下文模型高出22.36%[227]，展现其在生成质量控制上的潜力。同时，RAG允许应用层根据需求精准控制输入信息，这种灵活性使其在客户支持、法律文档分析等场景中更具实用价值[228]。

如表19所示，尽管长上下文窗口和RAG在功能上存在重叠，但适用场景和逻辑差异显著：前者适合简单任务或强模型（如GPT-4o），凭借一次性加载大量信息支持生成；后者在复杂任务及弱模型中通过检索提供高质量上下文，展现更强适应性。两者非零和博弈，而是具协同潜力，混合方法如ChatQA2[229]结合128K上下文与RAG缩小专有模型差距，检索技术从文本块级转向文档级进一步提升效率。未来，二者融合将成研究重点，长上下文可借助RAG优化“中间遗忘”和幻觉问题，RAG则需提升检索精度与召回率，混合系统需平衡成本、速度与质量。长上下文扩展未削弱RAG地位，反而凸显其互补性：RAG以精确性、灵活性和成本效益弥补前者不足，长上下文为其提供广阔信息基础，协同发展将提升生成效率，或成未来自动化系统核心架构，为AI理论与实践开辟新可能。

表19　RAG与长上下文窗口的性能对比（基于U-NIAH[226]和LaRA[227]研究）

特性	RAG	长上下文窗口
复杂任务表现	胜率82.58%，得分9.04	下降，受中间遗忘影响
成本与效率	低，仅处理相关信息	大，随长度增加显著上升
实时更新能力	强，动态检索最新数据	弱，依赖预训练数据
适用模型	弱模型提升显著	强模型表现更优

6.2　局限性

RAG技术通过融合检索与生成提升LLM在知识密集型任务中的表现，显著改善回答准确性与时效性，但仍面临多重技术瓶颈，限制其在复杂场景中的应用。以下基于最新研究系统阐述其主要局限性。

（1）检索模块局限性。检索召回的不完整性与精度不足表现为，当知识库中缺乏目标信息时，系统

可能生成看似合理但错误的回答，而非明确标识“未知”，这反映出其知识边界识别能力的不足[230]。即便相关信息存在于知识库，检索排序算法的局限可能导致高相关性文档未进入Top K结果，或因上下文窗口限制未被送入生成模块，从而影响输出质量。此外，语义搜索对细粒度语义差异的敏感度不足，在专业领域中易混淆相似概念，进一步加剧检索噪声问题[231]。

(2)生成质量与一致性问题。尽管检索机制旨在减少幻觉，研究表明，即使提供准确的检索内容，LLM仍可能生成与证据矛盾的陈述，原因在于其倾向于依赖参数中的固有知识而非严格遵循检索信息[232-233]。在复杂问答场景中，若检索结果包含噪声或歧义，模型可能无法准确提取关键信息，导致生成回答不完整、模糊或偏离用户所需的格式与细粒度要求，从而降低一致性与实用性。

(3)长期记忆与上下文管理不足。传统RAG基于单次查询的检索机制难以处理多轮对话中的长期依赖关系。虽然层次化索引(如层次聚合树)与长上下文LLM被提出以扩展记忆能力，但随着对话深度的增加，计算开销激增且可扩展性受限，其效果尚未完全验证[234-235]。此外，随着大语言模型上下文窗口的不断扩展，如何调整RAG系统以适应这些变化，成为了当前研究中的一个重点问题。

(4)计算成本与效率挑战。相比纯生成模型，RAG引入的检索步骤增加了时延与资源消耗，尤其在处理大型向量数据库或高并发场景时，存储查询成本与LLM调用费用显著上升。尽管缓存机制可缓解部分开销，但这进一步提升了系统复杂性，不利于实时应用。

(5)可解释性与可控性不足。虽然检索证据提高了答案的可追溯性，LLM的“黑箱”特性使其推理过程缺乏透明度，难以确保生成内容严格依据检索信息。当前系统在实现答案完全归因性方面仍存挑战，限制了其在高可信度场景中的应用[236]。

(6)跨模态对齐不足。跨模态对齐不足是多模态RAG系统性能提升的核心瓶颈。文本与图像在嵌入空间的语义表示差异显著，导致模型难以精准捕捉跨模态关联，进而在文本-图像检索中形成语义鸿沟。例如，基于CLIP[237]的对比学习虽能对齐模态，但在复杂场景下因语义细粒度不足，常召回低相关内容。在工业领域，视觉丰富文档的检索任务中，图像上下文相关性较弱，需依赖文本摘要增强对齐效果[238]。为此，多模态嵌入与生成式方法通过整合文本、图像和布局信号改善了对齐精度，但高并发或实时场景下，计算成本高企及模态不平衡问题仍限制其应用[239]。因此，跨模态对齐不足不仅降低检索精度与生成一致性，还在专业领域中加剧相似概念混淆的风险。

(7)长上下文窗口的中间遗忘问题。长上下文窗口的中间遗忘问题是RAG系统处理长序列输入时的核心挑战。模型倾向于优先利用上下文的开头与结尾信息，忽略中间部分的关键内容，导致信息召回与利用效率下降。例如，支持数十万token的超长上下文模型在中间位置的信息处理上呈现U形性能曲线，准确率显著降低。在多文档问答任务中，当相关信息位于中间时，模型准确率下降超过20%，这一问题源于注意力机制的首尾偏见，使生成模块难以充分利用检索到的中间信息[240]。在多轮对话中，中间遗忘问题随上下文长度增加而加剧，同时计算开销与内存需求显著上升[241]。尽管多尺度位置编码通过优化注意力分配缓解了部分问题，但在复杂对话场景中的效果仍待验证，凸显RAG在长上下文管理中的局限性[242]。

7 总结与展望

RAG技术作为LLM时代的重要创新，通过融合外部动态知识资源与模型内部知识库，显著缓解了LLM在幻觉、知识过时及推理不透明性等方面的局限性，为知识密集型任务提供了更高的准确性与可信度。本综述以RAG的工作流程为主线，系统剖析了各环节的优化潜力，深入探讨了预检索增强、检索器增强、检索策略增强、索引增强、检索后增强及LLM增强等技术的实现细节、理论基础及其适用场景，并通过对比检索增强与LLM增强两大优化策略，揭示了其在目标、优势、劣势及应用场景上的异同与协同效应，构建了一个全面的优化框架。相较于现有文献综述，本文在流程化视角下的系统性分析与技术深度上更进一步，不仅弥补了碎片化与浅层描述的不足，还通过跨技术对比与理论洞察，为研究者与开发者提供了清晰的技术路径与实践向导，同时为RAG技术的持续演进前瞻了方向。此外，本文整理了RAG领域常用数据集与评估指标，反思了其与上下文窗口扩展的关系，并从必要性与局限性角度全面评估了当前技术现状，为后续研究奠定基石。

RAG技术通过结合检索与生成显著提升了

LLM的性能，但其在检索精度、生成质量、计算效率及应用范围等方面仍面临诸多限制。为解决这些问题并拓展RAG的潜力，学术界与工业界近年来提出了一系列创新研究方向。以下详细分析每个方向的具体问题与挑战。

(1)多跳检索与主动检索。传统RAG采用“一次检索后生成”的静态模式，面对需多步推理或长篇回答的复杂任务时，单次检索难以覆盖全部相关信息，导致答案不完整或不准确[230]。现有系统缺乏自适应调整检索范围与深度的能力，限制召回率与精度。主动检索(如FLARE[177])通过预测生成中信息需求触发多轮检索以提升覆盖率，但检索时机与内容的确定仍存挑战，且多次检索显著增加计算开销。查询优化迭代(如RQ-RAG[54])通过分解复杂问题为子查询增强动态性，但子查询生成与整合易引入噪声或逻辑断裂。因此，研究多跳检索与主动检索对提升RAG在多步推理任务(如多文档问答)中的表现至关重要，需开发智能检索触发机制与高效查询分解算法，以平衡高召回率、低延迟与结果整合，推动其在知识密集型应用(如法律分析、学术研究)中的实用化。

(2)长期记忆与上下文管理。RAG在多轮对话或长时交互中因缺乏长期记忆机制，难以维持上下文连贯性与历史信息利用。传统单次检索无法捕捉跨轮次依赖关系，而层次化索引(如层次聚合树)虽通过树状结构组织对话历史，但对话深度增加时计算复杂度激增、检索效率下降。检索-记忆结合(如Adaptive-Note[234])通过迭代存储中间笔记扩展上下文，但动态记忆更新需解决数据冗余、存储成本上升及检索准确性下降问题。长上下文LLM虽缓解窗口限制，信息稀释却可能削弱推理质量。因此，研究高效记忆组织与自适应更新策略对提升RAG在动态对话系统(如客服、智能助手)中的连贯性与准确性至关重要，需平衡记忆扩展与效率优化，推动其在超长对话、实时交互及持续学习场景中的应用。

(3)降低幻觉与提升一致性。RAG通过检索外部证据减少幻觉，但LLM仍可能因过度依赖参数知识或误解检索内容，生成与证据矛盾的输出。幻觉检测通过解析模型内部机制识别偏离，但计算成本高且泛化能力有限。生成架构改进(如ReRAG[243])通过强化检索约束提升一致性，但限制过严可能削弱生成灵活性与自然性。噪声或歧义检索结果进一步加剧信息提取与生成的对齐难度。因此，研究轻量化幻觉检测与灵活生成约束机制是提升RAG在检测精度、生成质量与效率间平衡的关键，增强其在高风险领域(如医疗、法律)的可信度与用户信任，推动可信AI系统发展。

(4)效率优化与可扩展性。RAG的检索-生成流程在面对大规模知识库或高并发场景时，受时延与资源消耗制约。知识缓存(如RAGCache[244])通过预存结果降低实时开销，但更新策略滞后与覆盖率不足。高效数据结构(如Cuckoo Filter[245])优化层次化检索效率，但数据规模增长时查询复杂度与存储需求双重挑战加剧(CFT-RAG[152])。分布式计算与硬件加速虽提升可扩展性，但成本高昂与硬件依赖性限制普适性[246]。因此，研究轻量化检索算法、高效缓存机制及分布式架构是平衡高并发与低延迟、降低部署成本的关键，推动RAG在企业级实时应用中的高效性与可扩展性。

(5)结构化知识融合。传统RAG以非结构化文档为检索单位，难以捕捉知识间结构化关系，导致专业领域检索针对性不足。本体引导的RAG(如OG-RAG[247])通过知识图谱与超图增强结构化支持，但高质量图谱构建依赖人工标注，复杂关系建模增加计算开销，且结构化与非结构化数据融合易引入对齐误差[248]。因此，研究自动化图谱构建与跨模态对齐方法是克服知识表示、融合效率及泛化性瓶颈的关键，提升RAG在领域特定任务(如医疗诊断、法律咨询)中的精准性与智能化。

(6)多模态RAG。当前RAG主要处理文本数据，限制了其在多模态场景(如图像、音频、视频)中的潜力[36]。多模态RAG需整合异构数据，但受跨模态嵌入质量不足、检索效率瓶颈及生成整合复杂性制约：嵌入模型(如CLIP[237])对细粒度语义捕捉有限，降低检索相关性；大规模多模态数据库索引与搜索(如FAISS[193])随数据规模增长计算复杂度激增，实时应用延迟显著[238]；多模态模型(如BLIP-2[249]、GPT-4V[250])易受单一模态噪声干扰，输出一致性与事实性难保证，且高维数据处理成本高、可解释性不足[251]。因此，研究高效模态对齐算法、轻量化检索技术及改进生成架构是突破跨模态融合瓶颈、提升生成质量与相关性的关键，推动RAG在多模态问答、跨媒体检索及领域任务(如医疗影像分析)中的创新应用，拓展其智能化边界。

RAG技术正向更智能的检索、更持久的记忆、更真实生成及更高计算效率演进。通过创新设计优化检索与生成环节，RAG有望突破现有局限，在知识问答、对话系统及决策支持等场景中释放更大潜

力。尽管已取得显著进展，其全面实用化仍需解决检索准确性、多步推理、长期记忆、效率优化及生成一致性等核心挑战。上述六个研究方向针对检索精度、生成一致性、上下文管理、计算效率及多模态扩展等关键问题，系统提出优化路径。每个方向均面临技术实现与应用落地的独特难题，需依托算法创新、架构优化及资源整合实现突破。这些研究不仅旨在解决RAG当前瓶颈，也为推动其在知识密集型任务、动态对话、多模态应用及工业场景中的广泛应用奠定理论与实践基础。未来随着这些领域深入发展，RAG有望实现更高智能性、效率及跨领域适应性，充分释放其潜力。

参 考 文 献

[1] OpenAI, Hurst A, Lerer A, et al. GPT-4o System Card. arXiv preprint arXiv:2410.21276, 2024

[2] OpenAI. Introducing GPT-4.5. https://openai. com/index/introducing-gpt-4-5/

[3] Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

[4] Anthropic. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet

[5] Team G, Anil R, Borgeaud S, et al. Gemini: A family of highly capable multimodal models. arXiv preprint arXiv: 2312.11805, 2024

[6] Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024

[7] Yang A, Yang B, et al. Qwen2.5 Technical report. arXiv preprint arXiv:2505.09388, 2025

[8] Liu A, Feng B, Xue B, et al. DeepSeek-V3 technical report. arXiv preprint arXiv:2412.19437, 2024

[9] Guo D, Yang D, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025

[10] Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding//Proceedings of the International Conference on Learning Representations. Online, 2021:1-27

[11] Yue X, Ni Y, Zheng T, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 9556-9567

[12] Rein D, Hou B L, Stickland A C, et al. GPQA: A graduate-level google-proof Q&A benchmark//Proceedings of the First Conference on Language Modeling(COLM 2024. Philadelphia, USA, 2024:1-31

[13] Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021

[14] Jain N, Han K, Gu A, et al. LiveCodeBench: Holistic and contamination free evaluation of large language models for code//Proceedings of the The Thirteenth International Conference on Learning Representations. Singapore, 2025:1-41

[15] Hendrycks D, Burns C, Kadavath S, et al. Measuring mathematical problem solving with the math dataset//Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2. Online, 2021:1-11

[16] Kandpal N, Deng H, Roberts A, et al. Large language models struggle to learn long-tail knowledge//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA, 2023: 15696-15707

[17] Zhang Y, Li Y, Cui L, et al. Siren's Song in the AI Ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219, 2023

[18] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks//Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020. Vancouver, Canada, 2020: 9459-9474

[19] Chen D, Fisch A, Weston J, et al. Reading wikipedia to answer open-domain questions//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, 2017: 1870-1879

[20] Dinan E, Roller S, Shuster K, et al. Wizard of wikipedia: Knowledge-powered conversational agents//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019:1-18

[21] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North. Minneapolis, USA, 2019: 4171-4186

[22] Lee K, Chang M W, Toutanova K. Latent retrieval for weakly supervised open domain question answering//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 6086-6096

[23] Karpukhin V, Oguz B, Min S, et al. Dense passage retrieval for open-domain question answering//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020. Online, 2020: 6769-6781

[24] Lewis M, Liu Y, Goyal N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, Translation, and Comprehension//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 7871-7880

[25] Guu K, Lee K, Tung Z, et al. REALM: retrieval-augmented language model pre-training//Proceedings of the 37th International Conference on Machine Learning. Online, 2020: 3929-3938

[26] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners//Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Vancouver, Canada, 2020, 33: 1877-1901

[27] Borgeaud S, Mensch A, Hoffmann J, et al. Improving language

models by retrieving from trillions of tokens//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA, 2022: 2206-2240

[28] Izacard G, Grave E. Leveraging passage retrieval with generative models for open domain question answering// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online, 2021: 874-880

[29] Izacard G, Lewis P, Lomeli M, et al. Few-shot learning with retrieval augmented language models. arXiv preprint arXiv: 2208.03299, 2022, 1(2): 4

[30] Weston J, Shuster K. Blender Bot 2.0: An open source chatbot that builds long-term memory and searches the internet. https://ai. meta. com/blog/blender-bot-2-an-open-source-chatbot-that-builds-long-term-memory-and-searches-the-internet/

[31] Petroni F, Piktus A, Fan A, et al. KILT: A benchmark for knowledge intensive language tasks//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online, 2021: 2523-2544

[32] OpenAI. Introducing ChatGPT. https://openai. com/index/ChatGPT/

[33] Shi W, Min S, Yasunaga M, et al. REPLUG: Retrieval-augmented black-box language models//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico, 2024: 8371-8384

[34] Xu Z, Liu Z, Yan Y, et al. ActiveRAG: Autonomously knowledge assimilation and accommodation through retrieval-augmented agents. arXiv preprint arXiv:2402.13547, 2024

[35] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models// Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022. OrleansNew, USA, 2022: 24824-24837

[36] Zhao R, Chen H, Wang W, et al. Retrieving multimodal information for augmented generation: A survey//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023): Findings of the Association for Computational Linguistics. Singapore, 2023: 4736-4756

[37] perplexity. AI. Comet: A browser for agentic search by perplexity. https://www.perplexity.ai/comet

[38] xAI. Grok 3 beta — The age of reasoning agents. https://x.ai/blog/grok-3

[39] OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/

[40] google. Gemini deep research. https://gemini.google/overview/deep-research/

[41] perplexity. AI. Introducing perplexity deep research. https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research.

[42] bytedance. Deep Research at Your Fingertipsat Your Fingertips. https://deerflow.tech/

[43] Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2023, 2

[44] Gupta S, Ranjan R, Singh S N. A comprehensive survey of Retrieval-Augmented Generation (RAG): Evolution, current landscape and future directions. arXiv preprint arXiv: 2410.12837, 2024

[45] Fan W, Ding Y, Ning L, et al. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models// Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024. Barcelona, Spain, 2024: 6491-6501

[46] Zhao P, Zhang H, Yu Q, et al. Retrieval-augmented generation for ai-generated content: A survey. arXiv preprint arXiv: 2402.19473, 2024

[47] Ma X, Gong Y, He P, et al. Query rewriting in retrieval-augmented large language models//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023. Singapore, 2023: 5303-5315

[48] Zhang K, Sun Z, Yu W, et al. QE-RAG: A robust retrieval-augmented generation benchmark for query entry errors. arXiv preprint arXiv:2504.04062, 2025

[49] Lin C Y. Rouge: A package for automatic evaluation of summaries. Text summarization branches out. 2004: 74-81

[50] Pakhale K. Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges. arXiv preprint arXiv:2309.14084, 2023

[51] Zhang T, Lee B, Zhu Q, et al. Document keyword extraction based on semantic hierarchical graph model. Scientometrics, 2023, 128(5): 2623-2647

[52] Mao S, Jiang Y, Chen B, et al. RaFe: Ranking feedback improves query rewriting for RAG//Proceedings of the Findings of the Association for Computational Linguistics: The 2024 Conference on Empirical Methods in Natural Language Processing. Miami, USA, 2024: 884-901

[53] Li Z, Wang J, Jiang Z, et al. DMQR-RAG: Diverse multi-query rewriting for RAG. arXiv preprint arXiv: 2411.13154, 2024

[54] Chan C M, Xu C, Yuan R, et al. RQ-RAG: Learning to refine queries for retrieval Aug-mented generation//Proceedings of the First Conference on Language Modeling (COLM 2024. Philadelphia, USA, 2024: 1-20

[55] Ji Y, Meng R, Li Z, et al. Curriculum guided reinforcement learning for efficient multi hop retrieval augmented generation. arXiv preprint arXiv:2505.17391, 2025

[56] Rajaei D, Taheri Z, Fani H. Enhancing RAG's retrieval via query backtranslations//Proceedings of the International Conference on Web Information Systems Engineering (WISE 2024. Doha, Qatar, 2024: 270-285

[57] Koo H, Kim M, Hwang S J. Optimizing query generation for enhanced document retrieval in RAG. arXiv preprint arXiv: 2407.12325, 2024

[58] Rezaei M R, Hafezi M, Satpathy A, et al. AT-RAG: An adaptive rag model enhancing query efficiency with topic filtering and iterative reasoning. arXiv preprint arXiv:

2410.12886, 2024

[59] Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019

[60] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval, 2009, 3(4): 333-389

[61] Robertson S E, Walker S. On relevance weights with little relevance information//Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval. Philadelphia, USA, 1997: 16-24

[62] Bajaj P, Campos D, Craswell N, et al. Ms marco: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268, 2016

[63] Thakur N, Reimers N, Rücklé A, et al. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models//Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021. Online, 2021: 1-16

[64] Khattab O, Zaharia M. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020. Xi′an, China, 2020: 39-48

[65] Izacard G, Caron M, Hosseini L, et al. Unsupervised dense information retrieval with contrastive learning. Transactions on Machine Learning Research, 2022,2022:1-21

[66] Chen J, Xiao S, Zhang P, et al. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv: 2402.03216, 2024

[67] OpenAI. New and improved embedding model. https://openai.com/index/new-and-improved-embedding-model/

[68] Kwiatkowski T, Palomaki J, Redfield O, et al. Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 2019, 7: 452-466

[69] Joshi M, Choi E, Weld D S, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 1601-1611

[70] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1533-1544

[71] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100,000+ questions for machine comprehension of text//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016. Austin, USA, 2016: 2383-2392

[72] Pradeep R, Hui K, Gupta J, et al. How does generative retrieval scale to millions of passages//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing(EMNLP 2023. Singapore, 2023: 1305-1321

[73] Tang Y, Zhang R, Guo J, et al. Listwise generative retrieval models via a sequential learning process. ACM Transactions on Information Systems, 2024, 42(5): 1-31

[74] Wang Y, Hou Y, Wang H, et al. A neural corpus indexer for document retrieval//Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022. OrleansNew, USA, 2022: 25600-25614

[75] Zeng H, Luo C, Zamani H. Planning ahead in generative retrieval: Guiding autoregressive generation through simultaneous decoding//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024. Washington, USA, 2024: 469-480

[76] Wu S, Wei W, Zhang M, et al. Generative retrieval as multi-vector dense retrieval//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024. Washington, USA, 2024: 1828-1838

[77] Yu Z, Xiong C, Yu S, et al. Augmentation-adapted retriever improves generalization of language models as generic plug-in//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 2023: 2421-2436

[78] Zhang L, Yu Y, Wang K, et al. ARL2: Aligning retrievers with black-box large language models via self-guided adaptive relevance labeling//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, 2024: 3708-3719

[79] Qin Y, Liang S, Ye Y, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs//Proceedings of the The Twelfth International Conference on Learning Representations (ICLR 2024. Vienna, Austria, 2024: 1-23

[80] Wei G, Pang X, Zhang T, et al. DocReLM: Mastering document retrieval with language model. arXiv preprint arXiv: 2405.11461, 2024

[81] Yang H, Li Z, Zhang Y, et al. PRCA: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, 2023: 5364-5375

[82] Li S, Ramarkrishnan N. Oreo: A plug-in context reconstructor to enhance retrieval-augmented generation. arXiv preprint arXiv: 2502.13019, 2025

[83] Dong G, Zhu Y, Zhang C, et al. Understand what LLM needs: Dual preference alignment for retrieval-augmented generation//Proceedings of the 2025 ACM Web Conference. Sydney, Australia: ACM, 2025: 4206-4225

[84] Jin B, Yoon J, Qin Z, et al. LLM alignment as retriever optimization: An information retrieval perspective. arXiv preprint arXiv:2502.03699, 2025

[85] Dubois Y, Galambosi B, Liang P, et al. Length-controlled alpacaeval: A simple way to debias automatic evaluators. arXiv preprint arXiv:2404.04475, 2024

[86] Yu T, Zhang S, Feng Y. Auto-rag: Autonomous retrieval-augmented generation for large language models. arXiv preprint arXiv:2411.19443, 2024

[87] Ho X, Nguyen A K D, Sugawara S, et al. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps//Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020. Barcelona, Spain, 2020: 6609-6625

[88] Yang Z, Qi P, Zhang S, et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 2369-2380

[89] Salemi A, Zamani H. Learning to rank for multiple retrieval-augmented models through iterative utility maximization. arXiv preprint arXiv:2410.09942, 2024

[90] Feng W, Hao C, Zhang Y, et al. Airrag: Activating intrinsic reasoning for retrieval augmented generation via tree-based search. arXiv preprint arXiv:2501.10053, 2025

[91] Kocsis L, Szepesvári C. Bandit based monte-carlo planning//Proceedings of the Seventeenth European Conference on Machine Learning (ECML 2006. Berlin, Germany, 2006: 282-293

[92] Goel K, Chandak M. HIRO: Hierarchical information retrieval optimization. arXiv preprint arXiv:2406.09979, 2024

[93] Zhang N, Choubey P K, Fabbri A, et al. SiReRAG: Indexing similar and related information for multihop reasoning//Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025. Singapore, 2025: 1-20

[94] Sarthi P, Abdullah S, Tuli A, et al. Raptor: Recursive abstractive processing for tree-organized retrieval//Proceedings of the The Twelfth International Conference on Learning Representations. Vienna, Austria, 2024: 1-22

[95] Jiang J, Chen J, Li J, et al. RAG-Star: Enhancing deliberative reasoning with retrieval augmented verification and refinement. arXiv preprint arXiv:2412.12881, 2024

[96] Liu H, Zhang H, Guo Z, et al. CtrlA: Adaptive retrieval-augmented generation via inherent control. arXiv preprint arXiv: 2405.18727, 2024

[97] Xie W, Liang X, Liu Y, et al. WeKnow-RAG: An adaptive approach for retrieval-augmented generation integrating Web search and knowledge graphs. arXiv preprint arXiv:2408.07611, 2024

[98] Yang X, Sun K, Xin H, et al. Crag-comprehensive rag benchmark//Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024. Vancouver, Canada, 2024: 10470-10490

[99] Rezaei M R, Dieng A B. Vendi-rag: Adaptively trading-off diversity and quality significantly improves retrieval augmented generation with llms. arXiv preprint arXiv:2502.11228, 2025

[100] Asai A, Wu Z, Wang Y, et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection//Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024. Vienna, Austria, 2024: 1-30

[101] Mallen A, Asai A, Zhong V, et al. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023. Toronto, Canada, 2023: 9802-9822

[102] Islam S B, Rahman M A, Hossain K S M T, et al. Open-RAG: Enhanced retrieval augmented reasoning with open-source large language models//Proceedings of the Findings of the Association for Computational Linguistics: The 2024 Conference on Empirical Methods in Natural Language Processing. Miami, USA, 2024: 14231-14244

[103] Guan X, Zeng J, Meng F, et al. DeepRAG: Thinking to retrieval step by step for large language models. arXiv preprint arXiv:2502.01142, 2025

[104] Joren H, Zhang J, Ferng C S, et al. Sufficient context: a new lens on retrieval augmented generation systems. arXiv preprint arXiv:2411.06037, 2024

[105] Jeong S, Baek J, Cho S, et al. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico, 2024: 7036-7050

[106] Tang X, Gao Q, Li J, et al. MBA-RAG: A bandit approach for adaptive retrieval-augmented generation through question complexity//Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025. DhabiAbu, United Arab Emirates, 2025: 3248-3254

[107] Bhagdev R, Chapman S, Ciravegna F, et al. Hybrid search: Effectively combining keywords and semantic searches//Proceedings of the 5th European Semantic Web Conference (ESWC 2008): The Semantic Web-Research and Applications. Tenerife, Spain, 2008: 554-568

[108] Cormack G V, Clarke C L A, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009. Boston, USA, 2009: 758-759

[109] Sawarkar K, Mangal A, Solanki S R. Blended RAG: Improving RAG (Retriever-Augmented Generation) accuracy with semantic search and hybrid query-based retrievers//Proceedings of the 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR 2024. JoseSan, USA, 2024: 155-161

[110] Yuan Y, Chengwu L, Yuan J, et al. A hybrid RAG system with comprehensive enhancement on complex reasoning//Proceedings of the 2024 KDD Cup Workshop for Retrieval Augmented Generation. Barcelona, Spain, 2024: 1-13

[111] Shi L, Kazda M, Sears B, et al. Ask-EDA: A design assistant empowered by LLM, hybrid RAG and abbreviation de-hallucination//Proceedings of the 2024 IEEE LLM Aided Design Workshop (LAD 2024. JoseSan, USA, 2024: 1-5

[112] Rayo J, de la Rosa R, Garrido M. A Hybrid Approach to

Information Retrieval and Answer Generation for Regulatory Texts. arXiv preprint arXiv:2502.16767, 2025

[113] Ye L, Lei Z, Yin J, et al. Boosting conversational question answering with fine-grained retrieval-augmentation and self-check//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024. Washington, USA, 2024: 2301-2305

[114] Anantha R, Vakulenko S, Tu Z, et al. Open-domain question answering goes conversational via question rewriting//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2021. Online, 2021: 520-534

[115] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002. Philadelphia, Pennsylvania, USA, 2002: 311-318

[116] Heydari M H, Hemmat A, Naman E, et al. Context awareness gate for retrieval augmented generation. arXiv preprint arXiv: 2411.16133, 2024

[117] Lai Y, Wu J, Zhang C, et al. AdaCQR: Enhancing query reformulation for conversational search via sparse and dense retrieval alignment//Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025. DhabiAbu, United Arab Emirates: Association for Computational Linguistics, 2025: 7698-7720

[118] Li X, Song L, Jin L, et al. A Knowledge Plug-and-play test bed for open-domain dialogue generation//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italy, 2024: 666-676

[119] Alonso N, Figliolia T, Ndirango A, et al. Toward conversational agents with context and time sensitive long-term memory. arXiv preprint arXiv:2406.00057, 2024

[120] LlamaIndex. Building performant RAG applications for production. https://docs. llamaindex. ai/en/stable/optimizing/production_rag/

[121] Wang J, Meng F, Zhang Y, et al. Retrieval-augmented machine translation with unstructured knowledge. arXiv preprint arXiv:2412.04342, 2024

[122] Zhang J, Zhang Q, Wang B, et al. OCR Hinders RAG: Evaluating the cascading impact of OCR on retrieval-augmented generation. arXiv preprint arXiv:2412.02592, 2024

[123] He G, Dai Z, Zhu J, et al. Zero-indexing internet search augmented generation for large language models. arXiv preprint arXiv:2411.19478, 2024

[124] Patnaik S, Changwal H, Aggarwal M, et al. CABINET: Content relevance-based noise reduction for table question answering//Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024. Vienna, Austria, 2024: 1-24

[125] Pasupat P, Liang P. Compositional semantic parsing on semi-structured tables//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015. Beijing, China, 2015: 1470-1480

[126] Nan L, Hsieh C, Mao Z, et al. FeTaQA: Free-form table question answering. Transactions of the Association for Computational Linguistics, 2022, 10: 35-49

[127] Ye Y, Hui B, Yang M, et al. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023. Taipei, China, 2023: 174-184

[128] Kang D, Jung B, Kim Y, et al. Denoising table-text retrieval for open-domain question answering//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italy, 2024: 4634-4640

[129] Chen W, Chang M W, Schlinger E, et al. Open question answering over tables and text//Proceedings of the 8th International Conference on Learning Representations (ICLR 2020. AbabaAddis, Ethiopia, 2020: 1-18

[130] Huang J, Zhong W, Liu Q, et al. Mixed-modality representation learning and pre-training for joint table-and-text retrieval in OpenQA//Proceedings of the Findings of the Association for Computational Linguistics: The 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022: 4117-4129

[131] Modarressi A, Imani A, Fayyaz M, et al. RET-LLM: Towards a general read-write memory for large language models//Proceedings of the Workshop "How Far Are We From AGI?" at the Twelfth International Conference on Learning Representations (ICLR 2024. Vienna, Austria, 2024: 1-10

[132] Edge D, Trinh H, Cheng N, et al. From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130, 2024

[133] Scott, Behind the TechK.. https://www.microsoft.com/en-us/behind-the-tech

[134] Huang Y, Zhang S, Xiao X. KET-RAG: A cost-efficient multi-granular indexing framework for graph-RAG. arXiv preprint arXiv:2502.09304, 2025

[135] Jimenez Gutierrez B, Shu Y, Gu Y, et al. HippoRAG: Neurobiologically inspired long-term memory for large language models//Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024. Vancouver, Canada, 2024: 59532-59569

[136] Gutiérrez B J, Shu Y, Qi W, et al. From rag to memory: Non-parametric continual learning for large language models. arXiv preprint arXiv:2502.14802, 2025

[137] Sun J, Xu C, Tang L, et al. Think-on-graph: deep and responsible reasoning of large language model on knowledge graph//Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024. Vienna, Austria, 2024: 1-31

[138] Talmor A, Berant J. The Web as a knowledge-base for

answering complex questions//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). New Orleans, USA, 2018: 641-651

[139] Yih W, Richardson M, Meek C, et al. The value of semantic parse labeling for knowledge base question answering// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Berlin, Germany, 2016: 201-206

[140] Ma S, Xu C, Jiang X, et al. Think-on-Graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation. arXiv preprint arXiv: 2407.10805, 2024

[141] PalvelSubash. Unlocking distributed retrieval strategies in RAG: The key to scalability and precision. https://subashpalvel. medium. com/unlocking-distributed-retrieval-strategies-in-rag-the-key-to-scalability-and-precision-d1e9f5a62809

[142] Wang J, Yi X, Guo R, et al. Milvus: A purpose-built vector data management system//Proceedings of the 2021 ACM SIGMOD/PODS International Conference on Management of Data (SIGMOD 2021. Xi′an, China, 2021: 2614-2627

[143] Lin C Y, Kamahori K, Liu Y, et al. Telerag: Efficient retrieval-augmented generation inference with lookahead retrieval. arXiv preprint arXiv:2502.20969, 2025

[144] Chen W, Bai T, Su J, et al. Kg-retriever: Efficient knowledge indexing for retrieval-augmented large language models. arXiv preprint arXiv:2412.05547, 2024

[145] Cuconasu F, Filice S, Horowitz G, et al. Do RAG Systems Suffer From Positional Bias? arXiv preprint arXiv:2505.15561, 2025

[146] Yan X, Li X, Song D. Document re-ranking by generality in bio-medical information retrieval//Proceedings of the 6th International Conference on Web Information Systems Engineering (WISE 2005. YorkNew, USA, 2005: 376-389

[147] Zheng L, Cox I J. Re-ranking documents based on query-independent document specificity//Proceedings of the 8th International Conference on Flexible Query Answering Systems (FQAS 2009. Roskilde, Denmark, 2009: 201-214

[148] Hagen M, Völske M, Göring S, et al. Axiomatic result re-ranking//Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM 2016. Indianapolis, USA, 2016: 721-730

[149] Hofstätter S, Chen J, Raman K, et al. Fid-light: Efficient and effective retrieval-augmented text generation//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023. Taipei, China, 2023: 1437-1447

[150] Nian J, Peng Z, Wang Q, et al. W-RAG: Weakly supervised dense retrieval in RAG for open-domain question answering. arXiv preprint arXiv:2408.08444, 2024

[151] Jia P, Xu D, Li X, et al. Bridging relevance and reasoning: rationale distillation in retrieval-augmented generation. arXiv preprint arXiv:2412.08519, 2024

[152] Li Z, Ruan Y, Liu W, et al. CFT-RAG: An entity tree based retrieval augmented generation algorithm with cuckoo filter. arXiv preprint arXiv:2501.15098, 2025

[153] Dong J, Fatemi B, Perozzi B, et al. Don′t forget to connect! improving RAG with graph-based reranking. arXiv preprint arXiv:2405.18414, 2024

[154] Anil R, Dai A M, Firat O, et al. Palm 2 technical report. arXiv preprint arXiv:2305.10403, 2023

[155] Yu Y, Ping W, Liu Z, et al. Rankrag: Unifying context ranking with retrieval-augmented generation in llms//Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024. Vancouver, Canada, 2024: 121156-121184

[156] Wu M, Liu Z, Yan Y, et al. RankCoT: Refining knowledge for retrieval-augmented generation through ranking chain-of-thoughts. arXiv preprint arXiv:2502.17888, 2025

[157] Abdallah A, Mozafari J, Piryani B, et al. ASRank: Zero-shot re-ranking with answer scent for document retrieval. arXiv preprint arXiv:2501.15245, 2025

[158] Tack J, Kim J, Mitchell E, et al. Online adaptation of language models with a memory of amortized contexts. //Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024. Vancouver, Canada, 2024: 130109-130135

[159] Liska A, Kocisky T, Gribovskaya E, et al. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models//Proceedings of the 39th International Conference on Machine Learning (ICML 2022. Baltimore, USA, 2022: 13604-13622

[160] Li Z, Hu X, Liu A, et al. Refiner: Restructure retrieved content efficiently to advance question-answering capabilities//Proceedings of the Findings of the Association for Computational Linguistics: The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024. Miami, USA, 2024: 8548-8572

[161] Rau D, Wang S, Déjean H, et al. Context embeddings for efficient answer generation in retrieval-augmented generation// Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM 2025. Hannover, Germany, 2025: 493-502

[162] Shi K, Sun X, Li Q, et al. Compressing long context for enhancing rag with amr-based concept distillation. arXiv preprint arXiv:2405.03085, 2024

[163] Jin B, Yoon J, Han J, et al. Long-context llms meet rag: Overcoming challenges for long inputs in rag//Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024. Vienna, Austria, 2024: 1-34

[164] Li X, Zhou Y, Dou Z. Unigen: A unified generative framework for retrieval and question answering with large language models//Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI 2024. Vancouver, Canada, 2024, 38(8): 8688-8696

[165] Yoran O, Wolfson T, Ram O, et al. making retrieval-augmented language models robust to irrelevant context// Proceedings of the Twelfth International Conference on

Learning Representations (ICLR 2024. Vienna, Austria, 2024: 1-22

[166] LUO L, Li Y F, Haf R, et al. Reasoning on graphs: Faithful and interpretable large language model reasoning//Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024. Vienna, Austria, 2024:1-24

[167] Zhang T, Patil S G, Jain N, et al. Raft: Adapting language model to domain specific rag//Proceedings of the First Conference on Language Modeling (COLM 2024. Philadelphia, USA, 2024: 1-12

[168] Lin X V, Chen X, Chen M, et al. Ra-dit: Retrieval-augmented dual instruction tuning//Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024. Vienna, Austria, 2024:1-25

[169] Nguyen T, Chin P, Tai Y W. Reward-RAG: Enhancing RAG with reward driven supervision. arXiv preprint arXiv: 2410.03780, 2024

[170] Wang J, Yang Z, Yao Z, et al. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. arXiv preprint arXiv: 2402.17887, 2024

[171] Jin Q, Dhingra B, Liu Z, et al. PubMedQA: A dataset for biomedical research question answering//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019. HongKong, China, 2019: 2567-2577

[172] Yang D, Rao J, Chen K, et al. Im-rag: Multi-round retrieval-augmented generation through learning inner monologues// Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024. Washington, USA, 2024: 730-740

[173] Trivedi H, Balasubramanian N, Khot T, et al. MuSiQue: Multihop questions via single-hop question composition. Transactions of the Association for Computational Linguistics, 2022, 10: 539-554

[174] Li X, Mei S, Liu Z, et al. RAG-DDR: Optimizing retrieval-augmented generation using differentiable data rewards. arXiv preprint arXiv:2410.13509, 2024

[175] DAIR. AI. Prompt Engineering Guide. https://www.promptingguide.ai/

[176] Singh A, Ehtesham A, Kumar S, et al. Agentic retrieval-augmented generation: A survey on agentic RAG. arXiv preprint arXiv:2501.09136, 2025

[177] Jiang Z, Xu F F, Gao L, et al. Active retrieval augmented generation//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023. Resorts World Convention Centre, Singapore, 2023: 7969-7992

[178] Wang Z, Liu A, Lin H, et al. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. arXiv preprint arXiv:2403.05313, 2024

[179] Cobbe K, Kosaraju V, Bavarian M, et al. Training verifiers to solve math word problems. arXiv preprint arXiv: 2110.14168, 2021

[180] Yao S, Zhao J, Yu D, et al. React: Synergizing reasoning and acting in language models//Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023. Kigali, Rwanda, 2023: 1-33

[181] Tsai Y D, Liu M, Ren H. RTLFixer: Automatically fixing RTL syntax errors with large language model//Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC 2024. FranciscoSan, USA, 2024: 1-6

[182] Verma P, Midigeshi S P, Sinha G, et al. PlanRAG: Efficient test-time planning for retrieval augmented generation// Proceedings of the Workshop on Reasoning and Planning for Large Language Models at the Thirteenth International Conference on Learning Representations (ICLR 2025. Singapore, 2025: 1-23

[183] Tang J, Fan T, Huang C. AutoAgent: A fully-automated and zero-code framework for LLM agents. arXiv e-prints, 2025: arXiv: 2502.05957

[184] Tang Y, Yang Y. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries//Proceedings of the First Conference on Language Modeling (COLM 2024. Philadelphia, Pennsylvania, USA, 2024:1-16

[185] LangChain. Applications that can reason. Powered by LangChain. https://www.langchain.com/

[186] Li X, Dong G, Jin J, et al. Search-o1: Agentic search-enhanced large reasoning models. arXiv preprint arXiv: 2501.05366, 2025

[187] Xiong G, Jin Q, Wang X, et al. RAG-Gym: Optimizing reasoning and search agents with process supervision. arXiv preprint arXiv:2502.13957, 2025

[188] Alhanahnah M, Boshmaf Y. DepsRAG: Towards agentic reasoning and planning for software dependency management// Proceedings of the NeurIPS 2024 Workshop on Open-World Agents (at the 38th Annual Conference on Neural Information Processing Systems). Vancouver, Canada, 2024: 1-12

[189] Chang C Y, Jiang Z, Rakesh V, et al. MAIN-RAG: Multi-agent filtering retrieval-augmented generation. arXiv preprint arXiv:2501.00332, 2024

[190] Chen Y, Yan L, Sun W, et al. Improving retrieval-augmented generation through multi-agent reinforcement learning. arXiv preprint arXiv:2501.15228, 2025

[191] Zhang Y, Sun R, Chen Y, et al. Chain of agents: Large language models collaborating on long-context tasks//Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024. Vancouver, Canada, 2024: 132208-132237

[192] Vu T, Iyyer M, Wang X, et al. FreshLLMs: Refreshing large language models with search engine augmentation//Proceedings of Findings of the Association for Computational Linguistics (ACL 2024. Bangkok, Thailand, 2024: 13697-13720

[193] Douze M, Guzhva A, Deng C, et al. The faiss library. arXiv preprint arXiv:2401.08281, 2024

[194] Jiang A Q, Sablayrolles A, Mensch A, et al. Mistral 7B. arXiv preprint arXiv:2310.06825, 2023

[195] Zhang J, Zhang H, Zhang D, et al. End-to-end beam retrieval

for multi-hop question answering//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico, 2024: 1718-1731

[196] Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. Journal of Machine Learning Research, 2024, 25(70): 1-53

[197] Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018

[198] Lin S C, Asai A, Li M, et al. How to train your dragon: Diverse augmentation towards generalizable dense retrieval//Findings of the Association for Computational Linguistics: EMNLP 2023 (Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing). Singapore, 2023: 6385-6400

[199] Jiang H, Wu Q, Luo X, et al. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, 2024: 1658-1677

[200] Lassance C, Déjean H, Formal T, et al. SPLADE-v3: New baselines for SPLADE. arXiv preprint arXiv:2403.06789, 2024.

[201] He P, Gao J, Chen W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543, 2021

[202] Wang L, Yang N, Huang X, et al. Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv: 2212.03533, 2022

[203] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models//Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022. Online, 2022: 1-13

[204] Taori R, Gulrajani I, Zhang T, et al. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html, 2023, 3(6): 7

[205] Kotonya N, Toni F. Explainable automated fact-checking for public health claims//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020. Online, 2020: 7740-7754

[206] Krithara A, Nentidis A, Bougiatiotis K, et al. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. Scientific Data, 2023, 10(1): 170

[207] Min S, Michael J, Hajishirzi H, et al. AmbigQA: Answering ambiguous open-domain questions//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020. Online, 2020: 5783-5797

[208] Pal A, Umapathi L K, Sankarasubbu M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering//Proceedings of the 2022 Conference on Health, Inference, and Learning (CHIL 2022). Online, 2022: 248-260

[209] Fei Z, Shen X, Zhu D, et al. LawBench: Benchmarking legal knowledge of large language Models//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024. Miami, USA, 2024: 7933-7962

[210] Chen J, Zhou P, Hua Y, et al. FinTextQA: A dataset for long-form financial question answering//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, 2024: 6025-6047

[211] Husain H, Wu H H, Gazit T, et al. Codesearchnet challenge: Evaluating the state of semantic code search. arXiv preprint arXiv:1909.09436, 2019

[212] Bollacker K, Cook R, Tufts P. Freebase: a shared database of structured general human knowledge//Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI 2007), Volume 2. Vancouver, Canada, 2007: 1962-1963

[213] Aly R, Guo Z, Schlichtkrull M S, et al. FEVEROUS: Fact extraction and VERification over unstructured and structured information//Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (NeurIPS 2021. Online, 2021: 1-14

[214] Stelmakh I, Luan Y, Dhingra B, et al. ASQA: Factoid questions meet long-form answers//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022. DhabiAbu, United Arab Emirates, 2022: 8273-8288

[215] Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nature Medicine, 2025: 1-8

[216] Fei Z, Zhang S, Shen X, et al. InternLM-Law: An open-sourced Chinese legal large language model//Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025. DhabiAbu, United Arab Emirates, 2025: 9376-9392

[217] Neelakantan A, Xu T, Puri R, et al. Text and code embeddings by contrastive pre-training. arXiv preprint arXiv: 2201.10005, 2022

[218] Guo D, Ren S, Lu S, et al. GraphCodeBERT: Pre-training Code Representations with Data Flow//Proceedings of the Ninth International Conference on Learning Representations (ICLR 2021. Online, 2021: 1-18

[219] Post M. A call for clarity in reporting BLEU scores//Proceedings of the Third Conference on Machine Translation: Research Papers. Brussels, Belgium, 2018: 186-191

[220] Tan X, Wang X, Liu Q, et al. Paths-over-graph: Knowledge graph empowered large language model reasoning//Proceedings of The Web Conference 2025 (WWW 2025). Sydney, Australia, 2025: 3505-3522

[221] Mineiro P. Online joint fine-tuning of multi-agent flows. arXiv preprint arXiv:2406.04516, 2024

[222] Pham H, Nguyen T D, Bui K H N. ClaimPKG: Enhancing claim verification via pseudo-subgraph generation with lightweight specialized LLM. arXiv preprint arXiv:2505.22552, 2025

[223] Song M, Sim S H, Bhardwaj R, et al. Measuring and

enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse//Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025. Singapore, 2025:1-41

[224] Veneroso J, Jayaram R, Rao J, et al. CRISP: Clustering multi-vector representations for denoising and pruning. arXiv preprint arXiv:2505.11471, 2025

[225] Team G, Georgiev P, Lei V I, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024

[226] Gao Y, Xiong Y, Wu W, et al. U-NIAH: Unified RAG and LLM evaluation for long context needle-in-a-haystack. arXiv preprint arXiv:2503.00353, 2025

[227] Li K, Zhang L, Jiang Y, et al. LaRA: Benchmarking retrieval-augmented generation and long-context LLMs-no silver bullet for LC or RAG routing. arXiv preprint arXiv:2502.09977, 2025

[228] AlammarJay. RAG is here to stay: Four reasons why large context windows can't replace it. https://cohere.com/blog/rag-is-here-to-stay

[229] Xu P, Ping W, Wu X, et al. Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities//Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025. Singapore, 2025:1-15

[230] Barnett S, Kurniawan S, Thudumu S, et al. Seven failure points when engineering a retrieval augmented generation system//Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI (CAIN 2024. Lisbon, Portugal, 2024: 194-199

[231] EmanuilovSimeon. Retrieval Augmented Generation limitations. https://unfoldai.com/rag-limitations/#:~:text=RAG%20systems%20can%20introduce%20additional, quick%20response%20times%20are%20critical

[232] Sun Z, Zang X, Zheng K, et al. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability//Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025. Singapore, 2025:1-30

[233] Niu C, Wu Y, Zhu J, et al. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, 2024: 10862-10878

[234] Wang R, Zha D, Yu S, et al. Retriever-and-Memory: Towards adaptive note-enhanced retrieval-augmented generation. arXiv preprint arXiv:2410.08821, 2024

[235] Aadhithya A A. Enhancing long-term memory using hierarchical aggregate tree for retrieval augmented generation. arXiv e-prints, 2024: arXiv: 2406.06124

[236] Qi J, Sarti G, Fernández R, et al. Model internals-based answer attribution for trustworthy retrieval-augmented generation//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, USA, 2024: 6037-6053

[237] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning (ICML 2021. Online, 2021: 8748-8763

[238] Riedler M, Langer S. Beyond Text: Optimizing RAG with multimodal inputs for industrial applications. arXiv preprint arXiv:2410.21943, 2024

[239] Xu M, Wang Z, Cai H, et al. A Multi-granularity multimodal retrieval framework for multimodal document tasks. arXiv preprint arXiv:2505.01457, 2025

[240] Liu N F, Lin K, Hewitt J, et al. Lost in the Middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 2024, 11: 157-173

[241] ToreneSpencer. Understanding the impact of increasing LLM context windows. https://www.meibel.ai/post/understanding-the-impact-of-increasing-llm-context-windows

[242] Zhang Z, Chen R, Liu S, et al. Found in the middle: How language models use long contexts better via plug-and-play positional encoding//Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024). Vancouver, Canada, 2024, 37: 60755-60775

[243] Ko R, Gürkan M K, Vural F T Y. ReRag: A new architecture for reducing the hallucination by retrieval-augmented generation//Proceedings of the 9th International Conference on Computer Science and Engineering (UBMK 2024. Antalya, Türkiye: IEEE, 2024: 961-965

[244] Jin C, Zhang Z, Jiang X, et al. Ragcache: Efficient knowledge caching for retrieval-augmented generation. arXiv preprint arXiv:2404.12457, 2024

[245] Fan B, Andersen D G, Kaminsky M, et al. Cuckoo filter: Practically better than bloom//Proceedings of the 10th ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT 2014. Sydney, Australia, 2014: 75-88

[246] RichardsDavid. Scaling RAG for big data: Techniques and strategies for handling large datasets.. https://ragaboutit.com/scaling-rag-for-big-data-techniques-and-strategies-for-handling-large-datasets/

[247] Sharma K, Kumar P, Li Y. OG-RAG: Ontology-grounded retrieval-augmented generation for large language models. arXiv preprint arXiv:2412.15235, 2024

[248] Yepes A J, You Y, Milczek J, et al. Financial report chunking for effective retrieval augmented generation. arXiv preprint arXiv:2402.05131, 2024

[249] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning (ICML 2023. Honolulu, USA, 2023: 19730-19742

[250] OpenAI. GPT-4V(ision) system card. https://openai.com/index/gpt-4v-system-card/

[251] Sharifymoghaddam S, Upadhyay S, Chen W, et al. Unirag: Universal retrieval augmentation for multi-modal large language models. arXiv preprint arXiv:2405.10311, 2024

附录X.

表20　术语表

缩写/术语	英文全称	中文释义
GR	Generative Retrieval	生成式检索
CoT	Chain of Thought	思维链
LLMs	Large Language Models	大语言模型
RAG	Retrieval Augmented Generation	检索增强生成
Open-QA	Open-Domain Question Answering	开放域问答
NER	Named Entity Recognition	命名实体识别
MCTS	Monte Carlo Tree Search	蒙特卡罗树搜索
RRF	Reciprocal Rank Fusion	互惠等级融合
NDCG	Normalized Discounted Cumulative Gain	归一化折扣累积增益
MRR	Mean Reciprocal Rank	平均倒数排名
MAP	Mean Average Precision	平均准确率
TF-IDF	Term Frequency-Inverse Document Frequency	词频-逆文档频率
AMR	Abstract Meaning Representation	抽象语义表示
EM	Exact Match	精确匹配
NLI	Natural Language Inference	自然语言推理

YUAN Le, Ph. D. candidate. His research interests include RAG, agent, and automated program fault localization and repair.

LIU Shao-Hua, associate professor. His research interests include artificial intelligence and distributed computing.

WANG YU, Ph. D. candidate. His research interest covers RAG and agent.

ZHU Shang-Wei, Ph. D. candidate. His research areas are natural language processing and image photoelectric volume pulse wave.

WANG Tao, Ph. D., associate research professor. His research interests include acceleration technologies for generative artificial intelligence models.

MAO Tian-Lu, Ph. D., associate researcher. Her research interests include artificial intelligence, modeling and simulation.

Background

Retrieval-Augmented Generation (RAG) addresses a core problem in natural language processing (NLP): how to improve the factuality, timeliness, and reliability of Large Language Models (LLMs) by combining parametric knowledge with external evidence. Internationally, research on RAG has progressed from pipeline-level designs to workflow-aware optimization, covering indexing, retrieval, and generation, together with systematic enhancements such as dense and hybrid retrieval, multi-step retrieval strategies, post-retrieval reranking and compression, and model-side refinement. Benchmarking has also expanded across open-domain QA, reasoning QA, structured-data QA, and low-resource settings, enabling more comprehensive assessment of effectiveness and generalization. Nonetheless, open challenges remain widely recognized, including retrieval coverage and ranking errors, noise propagation into generation, efficiency under large-scale databases and high concurrency, long-context management, and cross-modal alignment. These trends and gaps motivate a process-oriented synthesis and a unified optimization perspective.

This paper positions RAG optimization within the broader international landscape and advances the topic by constructing an integrated six-category enhancement framework—covering pre-retrieval, retriever, retrieval-strategy, index, post-retrieval, and LLM enhancement—based on the standard three-stage RAG workflow of indexing, retrieval, and generation. It further conducts a comparative analysis of retrieval augmentation and

LLM enhancement, clarifying their respective goals, mechanisms, advantages, limitations, and application scenarios. In addition, the paper consolidates commonly used datasets and evaluation indicators to support quantitative assessment and reproducible benchmarking. Overall, the contribution lies in providing a structured, process-centric synthesis that integrates method taxonomy with evaluation resources and distills current limitations and future prospects in RAG optimization research.

The study is carried out by the Intelligent Cyber-Physical Systems Research Laboratory of Beijing University of Posts and Telecommunications (BUPT), together with collaborators from the State Key Laboratory of Basic Software and System, Institute of Software, Chinese Academy of Sciences, and the Beijing Key Laboratory of Mobile Computing and New-Type Terminals, Institute of Computing Technology, Chinese Academy of Sciences. The project is supported by the National Natural Science Foundation of China (Grant No. 91938301). Within this broader effort, the paper focuses on the RAG optimization sub-task—namely, the workflow abstraction, the six enhancement categories, and the comparative perspective between retrieval-oriented and LLM-oriented approaches—thereby addressing a key component of trustworthy and efficient LLM-based intelligent systems.

The significance of the project lies in providing a systematized foundation for improving the factual consistency, robustness, and adaptability of LLM applications in knowledge-intensive scenarios, while clarifying evaluation resources for fair comparison and reproducibility across tasks and domains. The research team's prior experience covers RAG and agent systems, AI and distributed computing, generative-model acceleration, and modeling and simulation, which informs the survey's process-oriented organization and the emphasis on efficiency, reliability, and applicability. Accordingly, the paper contributes a consolidated view of what has been achieved internationally, identifies persistent bottlenecks, and delineates where RAG optimization can be most impactful within the larger program of advancing LLM-centered intelligent systems.