小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0249
2026 年 2 月 第 2 期
Vol.47 No.2 2026

面向文档的检索增强生成技术综述

黄天金[1],朱兴动[1],刘 凯[1],汪时交[2],赵 鹏[1]

[1](海军航空大学,山东 烟台 264001)
[2](92925 部队,山西 长治 046000)
E-mail:tianarea@126.com

摘 要:大语言模型(LLMs)的迅速发展正在全球范围内引发深刻的技术变革.检索增强生成(RAG)作为一种通过融合外部知识以提升模型输出准确性、时效性与可靠性的技术范式,已成为增强 LLM 应用效能的关键手段.从"文档"的视角,提出"文档全谱系"概念,构建一个包含文档处理、嵌入与索引、检索、生成四大组件的 RAG 基础框架,并进行形式化描述.围绕"文档特性-优化策略"映射关系,系统梳理各项技术的最新进展,涵盖复杂文档处理、领域适应性嵌入、高级检索策略、可信生成等关键环节.最后,分析当前挑战,并展望未来发展方向.为相关领域的研究者和实践者提供一个以解决真实世界文档问题为导向的系统性参考.

关 键 词:检索增强生成;大语言模型;文档理解;知识检索;模型上下文协议(MCP)
中图分类号:TP391 **文献标识码**:A **文 章 编 号**:1000-1220(2026)02-0282-16

Survey on Document-oriented Retrieval-augmented Generation

HUANG Tianjin[1],ZHU Xingdong[1],LIU Kai[1],WANG Shijiao[2],ZHAO Peng[1]
[1](Naval Aeronautical University,Yantai 264001,China)
[2](92925,Changzhi 046000,China)

Abstract:The rapid advancement of large language models(LLMs) is driving profound technological transformations across various domains. As a technique that enhances LLM capabilities by incorporating external knowledge,retrieval-augmented generation(RAG) has demonstrated significant potential in improving output accuracy,timeliness and reliability. This paper presents a comprehensive survey of document-oriented RAG techniques. From a document-centric perspective,the study introduces the concept of "document genealogy" and establishes a fundamental RAG framework encompassing components of document processing,embedding & indexing,retrieval,and generation. The research reviews techniques through the lens of "document characteristics to optimization strategies" mapping,covering critical aspects including complex document processing,domain-adaptive embedding methods,advanced retrieval approaches,along with their corresponding generation control and verification mechanisms. The paper further identifies current challenges and outlines potential future directions in this field. By providing a thorough technical review grounded in practical document processing requirements,this survey aims to serve as a valuable reference for researchers and practitioners working on real-world RAG applications.
Keywords:retrieval-augmented generation;large language models;document understanding;knowledge retrieval;model context protocol

0 引 言

自 Transformer 架构[1]问世以来,大语言模型(Large Language Models,LLMs)的发展日新月异,其在自然语言理解与生成方面展现出的强大能力[2],正以前所未有的速度渗透并重塑着科研、金融、法律、医疗等众多领域.然而大语言模型的应用实践中,其固有的"三重挑战"日益凸显:知识静态性导致其无法知晓最新信息;内容不可靠性表现为易产生"幻觉"[3,4],缺乏事实依据;过程不透明性则使得用户决策难以溯源.检索增强生成(Retrieval-Augmented Generation,RAG)[5]通过整合外部知识库,为解决上述挑战提供了强大的框架.然而,现实世界中绝大部分知识以广义的"文档"形态存在,本文认为面向文档的 RAG(Document-Oriented RAG)是 RAG 技术在各领域应用的基本实现形式,而对作为知识源头的"文档处理"这一环节的深度把握,直接决定了 RAG 系统的性能上限.

本文旨在弥补当前研究在以"文档"为核心知识载体的 RAG 技术路径上的不足.第 1 节概述 LLMs 与 RAG 的背景、应用现状,并强调面向文档的 RAG 的核心挑战.第 2 节提出面向文档的 RAG 的四组件基础框架并进行形式化定义.第 3 节围绕该框架,并基于"文档特性-优化策略"的核心映射思想,深入剖析复杂文档处理、领域适应性嵌入、高级检索与推理、可信生成与溯源等关键技术环节的最新研究进展.第 4 节总结当前面临的主要挑战,并对未来的发展方向进行展望.第 5 节对全文进行总结.本文期望通过这种以文档为中心的视角,为相关领域的研究者和实践者提供一个结构清晰、论述深

收稿日期:2025-06-18 收修改稿日期:2025-07-17 基金项目:国家自然科学基金重大研究计划项目(91538201)资助. 作者简介:**黄天金**,男,1986 年生,硕士研究生,工程师,研究方向为人工智能与自然语言处理;**朱兴动**,男,1967 年生,博士,教授,研究方向为装备综合保障;**刘 凯**(通信作者),男,1986 年生,博士,副教授,研究方向为人工智能与自然语言处理;**汪时交**,男,1983 年生,工程师,研究方向为装备保障与机械工程;**赵 鹏**,男,1988 年生,硕士研究生,工程师,研究方向为信号识别与计算机视觉.

入、内容前沿的有价值参考.

1　背景与现状

1.1　大语言模型的崛起与 RAG 的诞生

近年来,基于 Transformer 架构[1]的大语言模型(LLMs)取得了突破性进展[2].这些模型主要分为闭源(如 GPT 系列)和开源(如 Llama 系列[6]、Qwen 系列[7]、DeepSeek 系列[8])两大阵营,极大地推动了自然语言处理领域的发展,并在摘要生成[9]、机器翻译[10]、问答系统[11]、文本分类[12]、信息检索[13]等多种任务中展现出卓越能力.其核心技术普遍基于 Transformer 架构[1]及其注意力机制.凭借其卓越能力,LLMs 迅速渗透到数据管理、文档分析[14]、代码智能等多个领域,不仅能高效处理非结构化数据,还能能够生成连贯的文本,并胜任需要综合、翻译或文本增强等复杂任务.

然而,LLMs 并非万能,其面临的固有挑战,如"幻觉"问题[4](生成与事实不符的内容)、在特定专业领域知识上的欠缺、以及处理复杂长文本或多步骤推理任务时的能力局限[2,5],都限制了其应用范围.特别是对于闭源模型,由于无法访问模型参数进行微调[15],对其进行优化和定制化增强更具挑战性.检索增强生成(RAG)[5]正是在这一背景下被提出,通过在生成前从外部知识库检索相关信息,来弥补 LLM 在知识时效性、事实准确性方面的不足.RAG 技术本身也在不断演进,从朴素的检索-生成流程(Naive RAG),发展到包含预处理、重排序等环节的高级 RAG(Advanced RAG),再到功能解耦、可灵活组合的模块化 RAG(Modular RAG)[2],体现了从流程固化到功能解耦、再到系统可控的演进路径.同时,也出现了如 Self-RAG[16]、DeepRAG[17]、RAFT[18]等探索模型自主检索和端到端优化的新范式.这些端到端的方法在推理效率和部署简化方面具有一定优势,但在适应垂直领域应用场景中文档和任务的复杂性方面可能面临挑战.

1.2　RAG 的应用现状

凭借其有效结合 LLM 推理能力与外部知识精确性的优势,RAG 技术已成为推动 LLM 在各行业落地应用的关键技术,催生了丰富的应用场景.在企业知识管理中,RAG 被用于构建智能问答系统,提升内部信息检索效率.在金融领域,自动化财报分析[19]、市场资讯解读和合规审查[20]是其典型应用.法律领域的研究如 Legal Query RAG[21]通过微调领域模型,辅助律师进行案例检索和文书起草.在医疗与生物科学领域,BioRAG[22]等框架整合专业知识库,支持辅助诊疗和科研探索.工业制造与运维场景下,RAG 可辅助工程师查询维修手册,生成故障处理指引[23].此外,在软件工程(代码生成与修复[20])和教育(智能辅导系统[20])等领域,RAG 也展现了巨大潜力.值得注意的是,出于数据安全和业务深度整合的需求,许多企业和机构倾向于采用私有化部署方案[2].这种对私有化部署的偏好,也从侧面反映了 RAG 技术在处理领域用户核心与敏感信息方面的能力日益成熟.关键领域用户因此逐渐建立信任,并开始将核心业务流程交由 AI 进行赋能.

1.3　面向文档的视角:RAG 的核心挑战来源

本文特意采用"面向文档"而非更宽泛的"面向数据"这一视角,其原因在于 RAG 技术在现实应用中的核心挑战,根源于知识载体的人本复杂性."文档"一词,天然地蕴含了格式、布局、多模态和上下文等多重维度,这些都是人类在记录和传递知识时留下的印记.从 RAG 的历史渊源看,它诞生于处理文本语料库[5];从应用挑战看,它必须攻克扫描件的噪声、PDF 的复杂版式、专业文献的术语壁垒.

RAG 技术在现实应用中的复杂性,从根本上源于其知识的源头——文档.本文将"文档"广义地定义为任何承载信息的实体或数据集合.世界上所有知识,在其流转与固化的过程中,均以广义的"文档"形态得以保存和呈现.其核心功能在于作为知识的载体,服务于信息的存储、检索、分析、传递与生成,并构成智能系统(如 RAG)认知与决策的信息基础.

真实世界的文档构成了一个复杂且动态的"文档全谱系".这一谱系在数字化演进维度上,涵盖了从传统的物理卷宗、手稿(阶段 1:物理文档,RAG 面临数字化转换如 OCR[24]),到经过扫描数字化的图像文档或直接以电子文件形式创建的文档(阶段 2:电子化文档,如 PDF、Word,RAG 需处理复杂布局解析),再到信息以预定义模式或标签化结构组织的数据存储(阶段 3:结构化与半结构化数据存储,如数据库、JSON,RAG 需关注数据到自然语言转换),乃至数字原生的多媒体与动态内容(阶段 4:数字原生多媒体与动态内容,如音视频、社交媒体信息流(Feed),RAG 需实现高效多模态特征提取与实时处理).本文提出的"文档全谱系"概念,将数据库等结构化信息视为该谱系中高度规整的一端,从而在统一的框架下,将焦点始终置于解决由文档形态多样性带来的根本性难题上.在信息形态维度上,它则囊括了从纯文本、结构化表格数据,到包含图表、公式、照片的图文混排文档,以及音频、视频等多模态富媒体内容.文本内容的结构化程度可分为非结构化、半结构化和结构化.正是这种全谱系内在的高度复杂性,直接导致了 RAG 在文档解析、信息提取和语义理解上面临的根本性挑战[25].例如,处理低质量扫描件时 OCR 引入的字符识别错误与版面结构误判、解析数字原生 PDF 中复杂的矢量图形与嵌入式字体、从网页中准确抓取动态加载内容、从电子表格中提取具有复杂嵌套结构的表格数据、对齐音视频中的语音文字与视觉场景信息、以及理解各领域特有的专业术语和隐晦表达等,这些难题无不根植于文档本身形态与内容的特性[25].模糊字符、动态布局或手写体文本的存在,对传统的手动数据提取和验证方法在效率、准确性和处理量方面构成了严峻的制约[25].

同时,尽管多模态大语言模型正逐步展现出直接处理多种信息模态的能力,但在当前及可预见的未来一段时间内,信息在被多数 LLM 深度利用前,往往仍需转化为高质量的文本形式(或其有效的文本化表示)[26].这个"文本出口"的特性,进一步凸显了高质量文本处理环节的关键性.因此,一个 RAG 系统的性能上限,其瓶颈往往在于文档处理组件能否有效应对这一"文档全谱系"所带来的纷繁挑战.

更关键的是,单一文档的复杂性往往并非孤立存在,而是多重维度的挑战相互叠加.以一份扫描的财务报告 PDF 为例,它可能同时面临低质量 OCR 引入的字符级噪声、复杂表格与图表带来的结构解析难题、以及专业领域术语造成的语义理解障碍.正是这种多维挑战的交织,决定了试图用一种通用的、端到端的方案处理所有文档类型几乎是不可能的幻想.

此外,不同类型的问题(例如,查找具体数字 vs. 分析整体趋势)也往往需要不同粒度的信息来支撑回答. 因此,本文认为,真正实现高效能的"面向文档的 RAG",必须基于具体任务需求,深刻把握应用领域的文档特性与多重挑战,并为其设计最优的策略组合. 这种基于"文档特性-优化策略"映射关系的深度适配能力,是决定 RAG 能否在复杂多变的垂直领域成功落地的关键所在.

2 面向文档的 RAG:框架与定义

为了系统性地分析和讨论面向文档的 RAG 技术,本节首先对其核心问题进行形式化定义,然后提出一个强调文档处理环节的 4 组件基础框架,并对各组件的职责与核心任务进行阐述.

2.1 RAG 的形式化问题定义

从概率建模的角度看,RAG 的目标是在给定用户输入(查询)x 的条件下,生成一个高质量的回答 y. 这可以表示为建模条件概率 $P(y|x)$. 标准的语言模型直接基于其内部参数 θ 来建模这个概率:$P(y|x;\theta)$.

RAG 框架的核心思想是,在生成答案 y 之前,先从一个由众多文档构成外部知识库 $D=\{d_1,d_2,\cdots,d_m\}$ 中检索出相关的文档(或文档片段)z. 然后,基于查询 x 和检索到的知识 z 来生成答案. 理想情况下,这相当于通过引入知识库 D 对原始概率进行边缘化:

$$P(y|x)=\sum_{z\in D_{sub}}P(y|x,z)\cdot P(z|x) \tag{1}$$

其中,D_{sub} 代表从整个知识库 D 中检索到的相关文档子集. $P(z|x)$ 表示文档 z 与查询 x 的相关性概率(由检索器估计),而 $P(y|x,z)$ 表示在给定查询和相关文档的条件下生成答案 y 的概率(由生成器建模).

由于遍历整个知识库 D 或精确计算这个求和在实践中不可行,RAG 系统通过一个检索-生成的两阶段过程来近似这个理想模型. 首先,检索器根据 $P(z|x)$ 找到一个小的、高度相关的文档子集.

$$Z_k=\{z_1,\cdots,z_k\}=\arg\operatorname*{top\text{-}k}_{z\in D}P(z|x)Z_k \tag{2}$$

然后,生成器基于 x 和 Z_k 来生成最终答案 y,近似计算 $P(y|x,Z_k)$. 这一过程的有效性,从源头上就受到知识库 D 中文档质量和可处理性的深刻影响.

2.2 四组件基础框架与形式化

本文提出的面向文档的 RAG 基础框架(如图 1 所示),将整个流程解构为文档处理、嵌入与索引、检索、生成这四大核心组件(见表 1). 该框架特别强调了文档处理作为流程起点和应对文档复杂性关键环节的重要性,因为文档识别的质量将直接影响后续所有组件的性能.

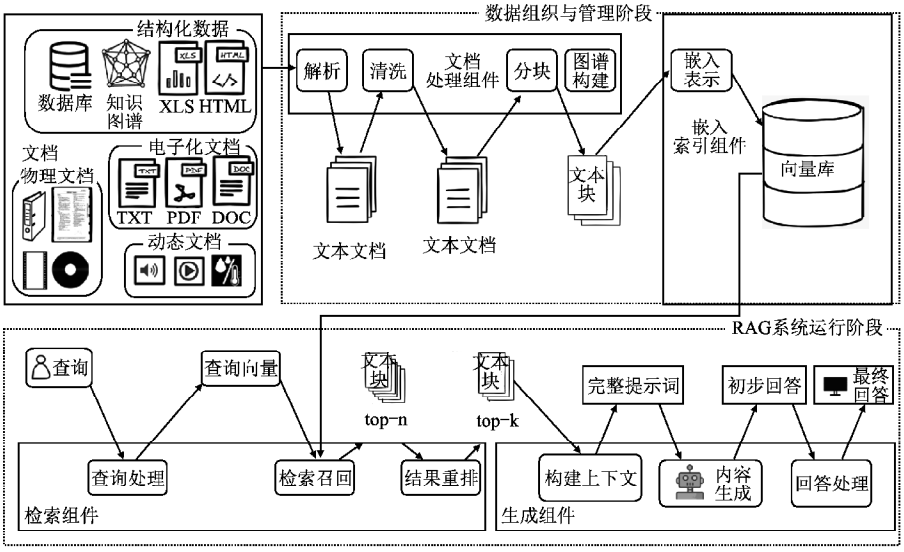

图 1 面向文档的 RAG 框架和基本流程

Fig. 1 Document-oriented RAG Framework and basic process

在需要与外部工具或服务交互以执行更复杂任务时,模型上下文协议(Model Context Protocol,MCP)[27] 可作为标准接口,实现 RAG 与外部系统的交互.

2.2.1 文档处理组件

作为 RAG 流程的起点,文档处理组件肩负着将来自"文档全谱系"的各种原始文档 D 转换为结构化的、适合后续处理和检索的知识单元集合 Z 的关键任务. 该组件旨在实现一个强大的文档处理函数 f_p,使得 $Z=f_p(D)$,其中每个知识单元 $z_i\in Z$ 通常是一个结构化的元组 (c_i,m_i),c_i 代表提取出的核心文本内容,而 m_i 则封装了丰富的元数据. 其关键子任务包括:

1)文档获取与解析,针对不同格式(物理、PDF、音视频等)采用相应技术(如 RPA、OCR[24,25]、STT[26])提取内容;

2)文档清洗与规范化,去除噪声,纠正错误,统一格式;

3)文本分块(Chunking),将文档切割为适配 LLM 上下文窗口的文本单元,如固定长度分块[5]或更智能的策略.

2.2.2 嵌入与索引组件

在文档内容经过处理转化为知识单元后,嵌入与索引组件(f_e,f_{idx})接续工作,其核心职责是将知识单元转换为计算机可高效检索的数值表示(嵌入向量),并构建支持快速相似性搜索的索引结构. 形式上,嵌入函数 f_e 将文本块 c_i 映射到高维向量空间 $v_i=f_e(c_i)$,索引构建函数 f_{idx} 则将所有知识单元及其向量组织成高效的索引结构 I. 其关键子任务包括:

表 1　面向文档的 RAG 框架核心组件概览

Table 1　Components of the RAG framework

组件	主要功能	主要文档相关挑战	关键技术方向举例
文档处理组件	从各类原始文档中提取、清洗、转换、组织知识;进行文本分块.	文档格式多样性、复杂布局、扫描/OCR质量、多模态内容、长文本处理、分块策略.	复杂文档解析(VLLMs, Layout-aware)、智能分块、文档图谱构建.
嵌入与索引组件	将处理后的文本块转换为可检索的向量表示(嵌入);构建高效的索引结构.	领域术语语义捕捉、多模态信息嵌入对齐、大规模向量索引的效率与存储.	领域适应性嵌入、跨模态嵌入、高效向量数据库与索引技术.
检索组件	根据用户查询,从索引库中定位并召回最相关的文本块;可能进行结果重排序.	查询意图理解、模糊查询处理、多源信息融合、复杂问题多跳推理、召回结果的相关性与多样性.	查询理解与改写、多路径/混合检索、迭代检索与多跳推理、上下文感知重排序.
生成组件	整合检索到的信息与原始查询,利用 LLM 生成回答;进行答案优化并提供溯源.	确保生成内容的事实一致性、避免幻觉、提供清晰准确的溯源、与外部动态知识/工具的交互.	事实一致性增强、幻觉缓解与纠错、细粒度溯源、模型上下文协议(MCP)应用.

1)文本/多模态嵌入生成,选择或训练合适的嵌入模型(如稀疏的 BM25[28]或稠密的 BERT[29]、CLIP[30]);

2)索引结构选择与构建,如倒排索引或近似最近邻(ANN)索引(如 HNSW[31]);

3)索引更新与维护.

2.2.3　检索组件

检索组件(f_r)的核心职责是根据用户查询 x,利用索引 I 快速定位并召回最相关的一组知识单元 Z_k. 形式上,该组件实现了一个检索函数 f_r,使得 $Z_k = f_r(x, I)$. 其关键子任务包括:

1)查询处理与编码,理解并编码用户查询为向量 v_x;

2)相似度计算与召回,利用索引计算相似度并筛选 Top-K 个结果[5];

3)结果重排序(Reranking),使用更精确的模型(如交叉编码器[32])对初步结果进行二次排序;

4)混合与多路径检索,融合多种检索策略以提升鲁棒性.

2.2.4　生成组件

作为 RAG 流程的最后一环,生成组件(f_g)的核心职责是整合用户查询 x 与检索到的知识 Z_k,并借助 LLM 生成最终回答 y. 形式上,该组件实现了一个生成函数 f_g,使得 $y = f_g(x, Z_k)$.

其关键子任务包括:

1)上下文构建,使用提示(Prompt)模板[33]组织输入;

2)内容生成,调用 LLM 生成初步回答;

3)回答优化与溯源,进行事实性校验[34]并提供清晰的引用[18,35,36];

4)与外部工具交互,在需要时通过 MCP[27]等接口调用外部服务.

3　关键技术与进展:优化策略

在前述基础框架之上,为了有效应对"文档全谱系"带来的复杂性挑战,并充分发挥RAG在提升LLM能力方面的潜

表 2　不同优化策略性能对比

Table 2　Comparison of performance of different optimization strategies

技术类别	传统/基线方法	改进技术	创新点	基准/任务	性能提升
分块策略	固定长度分块	基于文档结构元素分块[19]	利用标题、段落、表格等进行自适应切分	FinanceBench Q&A	ROUGE: +24%, BLEU: +152%
多模态文档解析	传统 OCR + 布局分析	Qwen2.5-VL[7]	统一 HTML 格式化表示(含 bbox),处理图文表公式等	OmniDocBench, OCRBench	SOTA, SOTA
领域术语增强	通用密集检索器	ITEM 术语增强检索[37]	利用领域术语词典增强嵌入和检索	CMMD(制造业文档)	Acc: +17%, Token: -20%
领域模型微调	通用嵌入模型	微调领域嵌入模型[21]	在特定领域(法律)语料上微调嵌入模型(GIST-Law-Embed)	多个法律问答任务	Hit Rate: +13%, MRR: +15%
多模态检索	OCR/单视觉检索	ViDoRAG[38]	GMM 混合策略 + 多智能体推理处理视频/多模态文档	ViDoSeek benchmark	> +10%
查询改写	原始查询	Query2Query[39]	TAO 框架 + 黄金法则(What/How/Why)提问优化	学科试题	F1: +16.5%
混合检索	BM25 + DPR	Meta-RAG[40]	混合编码 + 重排序	电力规范问答	Acc: +80.43%
混合检索	单一检索方式	HyPA-RAG[41]	融合稠密 + 稀疏 + 知识图谱检索,查询复杂度自适应	LL144(法律政策)	Correctness +0.06, Faithfulness: +0.9, Recall: +0.9
混合检索	向量检索或 KG 检索	混合架构[42]	结合向量检索与知识图谱用于金融领域	财务问答任务	Recall: +0.15%
多跳/迭代检索	单跳/固定多跳	DeepRAG[17]	MDP 建模检索决策,链式校准	HotpotQA, 2WikiMultihopQA 等	Avg EM: +21.99%, 检索次数: -30% ~ -60%
多跳/迭代检索	固定检索流程	Adaptive-RAG[43]	根据查询复杂度动态选择检索策略(单步/多步/无检索)	多跳 QA 任务	F1: +10%, 时间开销: -40%
领域知识层次检索	单一源检索	BioRAG[22]	基于 MeSH 知识层次的混合与迭代检索	GeneTuring, MedMCQA	Acc: 提升 +30%, SOTA
精细化重排序	检索固定数量块	DSLR[44]	句子级重排 + 动态上下文构建	NQ, TQA, SQD 等 QA 任务	Acc: +6.3%, Token 数: -53.4%
事实验证纠错	固定检索 + 直接生成	CoVe[34]	"规划-生成-验证"的迭代交叉验证流程	Llama 65B On Multi-SpanQA	FACTSCORE: 55.9 to 71.4
细粒度溯源(文档内)	无/文档级溯源	HiQA[35]	溯源至文档 ID 和章节路径(通过 HCA 模块实现)	MasQA	Adequacy: 4.89 to 4.96

术需更关注其输出如何优化服务于下游 RAG 任务.

3.1.2　智能分块策略:平衡粒度与语义完整性

将长文档或解析内容切分为适合检索与 LLM 处理的文本块(Chunking),是RAG预处理的关键环节,直接影响检索

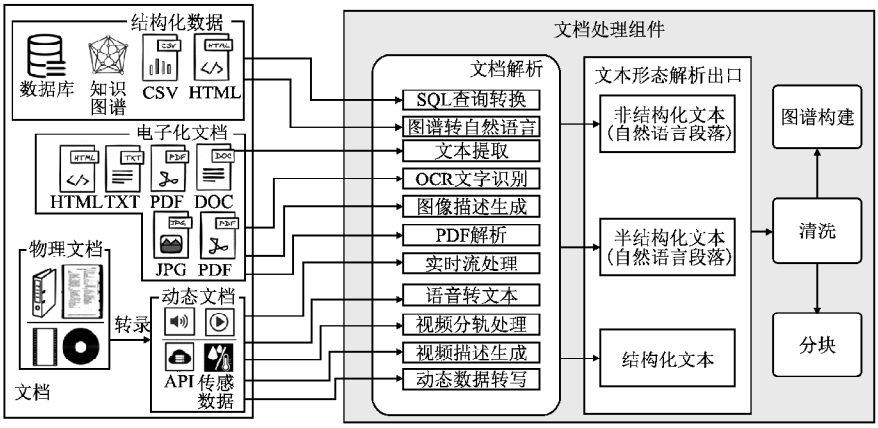

图 2　复杂文档处理示意图

Fig. 2　Processing complex documents

精度及生成上下文质量. 确定最优分块策略本身即为复杂挑战. 基础的固定长度分块[5]虽实现简单、计算高效,但其主要缺陷在于忽略文本内在语义与结构,常在不当位置切断句子或段落,导致语义割裂,严重影响后续处理. 若文档解析阶段因格式噪声导致段落边界识别错误,即便更智能的分块策略也可能失效. 因此,研究重点已转向发展更智能、自适应的分块策略,旨在切分时最大限度保持语义单元完整性,尊重文档原始结构(假设解析准确),并据下游任务需求优化分块粒度. 表 4 对常用的文档分块策略进行了对比分析.

表 4　常用分块策略对比分析

Table 4　Analysis of common chunking strategies

分块策略	描述	优点	缺点	适用场景
固定长度分块	按固定字符数或 Token 数切分成块	实现简单	截断语义,丢失上下文	简单问答
递归分块	先大后小逐步切分	平衡效率与可读性	上下文可能被割裂	通用框架默认
滑动窗口分块	使用滑动窗口重叠切分	减少信息丢失	增加冗余和开销	高精度问答
按语义分块	利用 NLP 模型识别语义边界	保留语义完整性	实现较复杂	高质量问答
按结构分块	根据文档结构(如 HTML 标签)切分	保留原始结构信息	对非结构化文本无效	PDF 解析、网页抽取
长分块策略	单个检索单元较长	极大保留语义完整性	资源消耗大	复杂推理、长文档分析
领域适应分块	根据领域文档特点设计	保留语义、提升相关性	设计复杂	垂直领域问答

为保持语义连贯,基于自然语言处理技术的分块策略应运而生. 例如,利用句子边界检测分块是常见改进,能确保至少句子层面的语义完整. 更进一步的语义分块则尝试利用词嵌入相似度变化、主题模型等 NLP 技术识别语义连贯的文本段落边界,以在段落级别保持语义完整性,但这通常计算成本更高,且对文本质量(如 OCR 引入的语义噪声)较为敏感. 研究如 LongLLMLingua[56]探索提示压缩,而 LongAgent[57]等工作致力于扩展有效上下文长度.

处理具明确结构信息的文档时(如 HTML 网页、Markdown 文件、准确解析的 PDF[19,35]、JSON 对象),利用文档固有结构标记(如标题、段落、列表项、表格单元格)进行分块是一种高效且自然的方式,这对于结构复杂的数字原生文档和长篇幅、结构化文本尤为重要(对应表 3). 此种结构化分块能良好保留文档原始逻辑层次,使检索到的信息块更具上下文意义. 然而,其前提是文档解析阶段能准确识别和提取这些结构标记,否则格式噪声将直接破坏其有效性.

实践表明,最优分块策略常需据具体文档类型和应用场景定制化设计. 例如,针对金融报告这类结构复杂的数字原生文档,研究发现,利用文档章节标题、表格标识等结构元素进行自适应分块,相比固定长度分块,在下游问答任务(FinanceBench Q&A)上性能显著提升(ROUGE + 24%,BLEU + 152%)[19]. 为应对 LLM 处理长文档的挑战,ChuLo[58]提出通过提取文档关键短语指导分块,旨在减少输入长度同时保留核心信息. HyPA-RAG[41]的研究亦证实,在法律领域这类领域专业文档根据其文档特点设计分块策略能有效提升 RAG 系统性能. 在保险领域,有研究尝试利用领域特定词根表和标签库进行更精细化分块[59]. 此外,滑动窗口分块[60]通过设置块间重叠区域,可在一定程度上缓解边界切割问题,但会增加数据冗余. 文献[61]提出先检索后分块的“后分块”策略.

最后,分块粒度的选择亦应考虑下游任务需求. 对于需广泛上下文进行复杂推理的任务,LongRAG[49]探索了使用超长分块(如 > 4096 Tokens)并配合专门的长上下文 LLM,这适用于处理长篇幅文本. 而在事件抽取等需捕捉特定触发词及其完整上下文的任务中,为避免关键信息被分割,可能选择文档级分块或采用确保事件提及不被切分的策略[62].

综上,文档分块技术正从简单固定规则向更智能、灵活和任务导向发展. 如何在保留语义完整性、适应文档结构(并顾及文档识别质量影响)、优化检索粒度及平衡计算效率间找到最佳平衡点,是当前研究的核心. 结合 LLM 的理解能力与长上下文处理技术的进步,智能分块有望显著提升 RAG 系统捕获和利用知识的效率与准确性.

3.1.3　知识图谱构建与利用:赋予知识结构化深度

传统 RAG 主要依赖检索非结构化文本块,常忽略文档中

力,学术界和工业界发展出了一系列精巧的优化技术.然而,这些技术并非孤立存在,其选择与应用必须紧密围绕待处理文档的具体特性.本节将建立一个从“文档特性”到“优化策略”的映射视角,系统梳理在不同文档挑战下,各环节的关键技术进展.

首先,表2展示了部分代表性技术相较于基线方法的性能提升情况,为本节的讨论提供了一个直观的成果概览.

为了更清晰地展示本节的核心思想,即如何根据文档特性选择优化策略,本文总结了典型文档特性、其带来的核心挑战以及相应的RAG优化策略,如表3所示.后续各小节的讨论将围绕这些映射关系展开.

表3 文档特性-优化策略映射表

Table 3 Features-strategy mapping

文档特性	核心挑战	关键优化策略(对应章节)
低质量扫描件/手写稿	OCR字符错误(语义噪声)、版面丢失(格式噪声)、图像伪影	文档处理:增强型OCR+LLM校正(如ERPA[25])、视觉大语言模型(VLLM)直接解析[7].(3.1.1)检索:鲁棒的稀疏检索(对错字不敏感)、查询扩展[39].(3.3.1,3.3.2)生成:强化事实验证与纠错(CoVe[34]).(3.4.2)
结构复杂的数字原生文档(如金融/科研PDF)	嵌套表格、数学公式、多栏布局、矢量图表解析困难	文档处理:结构感知解析(Nougat[45],Qwen-VL[7])、基于文档结构(标题/章节)的智能分块[19].(3.1.1,3.1.2)嵌入:表格专有嵌入模型(TableGPT[46])、多模态嵌入.(3.2.2)
长篇幅、结构化文本(如法律文书、技术手册)	上下文丢失、长距离依赖关系捕捉、分块粒度难以权衡	文档处理:语义/结构化分块、知识图谱构建[47,48]以捕捉实体关系.(3.1.2,3.1.3)检索:多跳/迭代检索[17,43]、句子级重排序(DSLR[44]).(3.3.3,3.3.4)生成:长上下文LLM应用[49]、细粒度溯源[35,36].(3.4.3)
领域专业文档(如医疗记录、工程规范)	大量专业术语、隐晦知识、通用模型语义理解偏差	嵌入:领域适应性嵌入微调(GIST-Law-Embed[21])、注入领域词典/知识图谱(BioRAG[22],ITEM[37]).(3.2.1)检索:混合检索(关键词+语义)、基于领域知识的查询增强.(3.3.1,3.3.2)
多模态混合内容(如网页、产品介绍)	图文音视频信息对齐与融合、跨模态语义理解	文档处理:VLLM/Speech LLM进行统一内容提取[7,26].(3.1.1)嵌入:跨模态嵌入对齐(CLIP[30]).(3.2.2)检索:多模态检索(ViDoRAG[38]).(3.3.2)
半结构化/结构化数据(如JSON,数据库)	数据到自然语言的转换、结构化查询的生成	文档处理:模式(Schema)提取.检索:文本到SQL/Cypher的转换、知识图谱查询[48].(3.1.3,3.3.2)生成:约束性生成[50],确保输出格式合规.(3.4.1)

3.1 文档处理与表示优化

此阶段优化直接作用于RAG流程的知识库基石.其核心目标在于,通过更智能的文档处理(f_p)及更具表征力的嵌入与索引技术(f_e,f_{idx}),将形态各异的原始文档高效转化为高质量、易检索且语义丰富的知识单元集合.Pre-Retrieval阶段的成效,无疑直接设定了后续检索与生成性能的上限.图2示意性地展示了针对不同类型复杂原始文档,如何通过解析、智能分块乃至可选的图谱构建,将其转化为RAG系统可利用的知识单元.

3.1.1 先进文档解析:跨越模态与结构鸿沟

从多样化、结构复杂的原始文档中准确提取信息,是构建高质量RAG知识库的首要挑战.传统的OCR与基于规则的布局分析方法,在处理低质量扫描件、手写体、复杂图文混排或非标准格式时常显不足[25](对应表3),易导致信息提取错误或结构丢失.这些错误,如OHRBench[51]研究指出的语义噪声(字符识别错误致词义改变)与格式噪声(布局元素表示不一致致结构误解),会级联影响RAG的后续所有组件.为此,研究趋势正转向利用大型预训练模型进行端到端、多模态感知的文档理解,力求更整体、深入地把握文档内容与结构,从源头减少此类噪声.

布局感知预训练模型的出现是一项重要进展.以LayoutLM系列[52]为代表的研究,通过在预训练阶段融合文本语义及关键的二维布局信息(如文字块边界框坐标),使模型能理解文档空间结构,显著提升了在表格理解、表单抽取等视觉丰富文档理解任务上的性能[53].这类模型为处理版式相对固定的电子文档提供了有力工具,有助于更准确地切分和理解文档结构,减少因布局误判引入的格式噪声.

当前,视觉大语言模型(VLLMs)正成为文档解析领域的新兴力量,它们凭借强大的视觉与语言理解能力直接处理文档图像.例如,针对含有复杂公式和表格的科研PDF这类文档,Nougat[45]专注于将含复杂公式、表格的科学文献PDF精准转换为结构化Markdown,极大便利了学术信息的后续利用.先进的多模态大模型如Qwen2.5-VL[7],则能端到端识别文档图像中的文本、图表、公式等元素,并统一输出为保留丰富结构信息的HTML格式,在OmniDocBench等基准上表现出色[7].

VLLMs的端到端特性有望减少传统多阶段处理流程中的错误累积.同时,研究者亦探索利用通用LLM(如GPT系列)直接解析PDF为结构化文本的可行性[54],例如HiQA[35]框架便采用GPT-4作为PDF结构化解析器.Unstructured库也提供了用于提取和预处理图像和各类文档的开源组件,例如文献[19]就通过该组件将金融报告输出为含有页面元素信息的JSON格式文档.若这些方法能准确捕捉文档结构,将助力后续更精确的分块与信息提取.

为应对大量低质量文档(如历史文献、扫描模糊档案),研究者提出了传统技术与LLM结合的增强型自动化方案.ERPA模型[25]即结合OCR与LLM,利用LLM校验、修正OCR初步结果并进行结构化处理,同时动态适应不同文档布局,在提升准确性的同时显著缩短了处理时间[25],可视为对OCR引入语义噪声的有效后处理.值得注意的是,即使在视觉模型快速发展的今天,专门的OCR扫描器(如PaddleOCR[55])结合先进的提示技术(如SFT)在某些文档处理任务上仍然能够优于先进的视觉模型[53].对于日益重要的音频文档(如会议录音,属于多模态混合内容),语音大语言模型(Speech LLMs)[26]通过端到端建模与多模态融合,显著提升了语音识别的准确性和鲁棒性[20],使RAG能有效利用此类信息源,减少因ASR错误引入的语义噪声.

综上,文档解析技术正从基于规则、分离处理向基于深度学习、端到端多模态理解深刻变革.其核心目标在于提升从各类文档中提取结构化、准确信息的质量,最大限度减少因文档识别(尤其OCR)引入的语义和格式噪声,为构建高质量RAG知识库奠定坚实基础.正如OHRBench[51]所强调,评估OCR对RAG各组件的级联影响至关重要,未来文档解析技

丰富的实体、概念及其结构化关系.将这些信息显式构建为知识图谱(Knowledge Graph,KG)并融入 RAG 流程,能为系统提供超越纯文本语义相似性的深层语义理解与推理能力,是提升 RAG 处理复杂知识能力的重要方向,尤其适用于长篇幅、结构化文本中蕴含的复杂关系和半结构化/结构化数据的利用(对应表 3).知识图谱通过节点(实体)和边(关系)清晰表达组织架构、产品依赖等复杂关系,对需结构化知识或多步推理的问题尤为重要.文档识别(尤其命名实体识别和关系抽取)的准确性,直接影响从文档构建知识图谱的质量.

然而,传统知识图谱构建流程复杂(含 NER、RE、EL、本体设计等),成本高昂且难扩展.当前显著趋势是利用 LLM 的理解与生成能力简化乃至自动化知识图谱构建.例如,GraphRAG[47]探索从文本直接构建图谱索引;FastRAG[50]尝试从半结构化数据中学习模式自动构建图谱;KG-RAG[48]则提出利用 LLM 从文本灵活提取三元组,无需预设复杂本体,极大降低了构建门槛.国内亦有研究尝试用大模型(如 GLM-4)据预设规则从特定领域文本(如中医古籍《伤寒论》)自动提取三元组[63],为特定领域知识图谱快速构建提供了新思路.

构建的知识图谱可通过多种方式增强 RAG 系统.其一为图谱增强检索:将用户自然语言查询转换为图查询语言(如 SPARQL,Cypher),在图谱中精确查找相关实体、关系或子图,并将结果与传统文本检索结果融合,提供更全面精确的上下文.其二为利用图谱进行推理:图谱结构信息可指导多跳推理,如沿关系路径探索信息.此外,图嵌入技术可将图谱节点与关系表示为向量,与文本嵌入融合于同一向量空间,或用于链接预测、关系推理等,为生成模型提供更丰富的结构化上下文.例如,G-Retriever[64]将图理解任务建模为图上检索问题,展现了图谱理解复杂关系的潜力.

总之,将文档知识转化为结构化知识图谱并有效融入 RAG 框架,正成为增强系统语义理解深度与推理能力的重要途径.利用 LLM 简化知识图谱构建,并通过图查询或图嵌入优化检索与推理,有望使 RAG 系统更深入把握文档间的复杂关系和隐含知识,尤其在需深度领域知识和结构化推理的场景中.

3.2 嵌入与索引优化

文档内容经解析分块后,如何将其有效转换为机器可理解和检索的表示,并构建高效索引结构,是 Pre-Retrieval 阶段的另一核心.此环节优化聚焦于提升嵌入向量(f_e)对特定领域和多模态内容的表征精度,及索引结构(f_{idx})的效率与可扩展性.文档识别阶段引入的语义噪声(如 OCR 错字)会直接扭曲文本语义,使嵌入向量偏离其应有语义位置,严重影响后续检索准确性.

3.2.1 领域适应性嵌入:精准捕捉专业语义

通用预训练嵌入模型(如 BERT[29]及其变体)虽在通用语料上学习了丰富语言知识,但在处理特定专业领域(如医疗、法律、金融)文档时(对应表 3),常难准确捕捉领域术语、概念及其微妙语义关系,导致检索偏差或遗漏.因此,发展领域适应性嵌入 技术,提升嵌入向量对专业知识的表征能力,对 RAG 在垂直领域的成功应用至关重要.

主要策略之一是将领域知识显式注入嵌入或检索过程.例如,BioRAG[22]处理生物医学问题时,除使用 PubMedBERT 外,还利用权威 MeSH 术语词典构建知识层次,辅助查询理解与检索过滤,显著提升了 GeneTuring 等任务性能[22].ITEM 方法[37]针对制造业文档术语密集问题,构建专门术语词典增强嵌入与检索,在领域问答任务上实现准确率提升 17% 和处理 Token 数减少 20%[37].Meta-RAG[40]通过在嵌入中融合文档元数据(如标题、作者、日期)提供更丰富上下文,在电力领域问答中大幅提升准确率[40].智能运维领域有研究在查询故障时引入故障树分析等领域知识图谱指导嵌入与检索[65],利用结构化领域知识提升相关性判断.

另一直接常用方法是在目标领域文档语料上对预训练嵌入模型进行微调(Fine-tuning).通过在大量领域相关文本上继续训练,可使嵌入向量更好捕捉领域特定语义.Legal Query RAG[21]通过在大量法律文本上微调得 GIST-Law-Embed 模型,在多个法律问答基准上较通用模型显著提升(Hit Rate + 13%,MRR + 15%)[21].Kim 等人[66]亦探讨了金融问答文档中优化检索策略,同样涉嵌入表示的领域适应性.文献[67]则在金融领域微调了 6 种嵌入模型进行对比.领域微调效果显著,但挑战在于获取足够规模和质量的领域数据及所需计算资源.为缓解数据稀缺,研究者探索了 LLM 辅助微调:Promptagator[68]利用 LLM 基于少量示例生成更多高质量查询-文档对,用于监督微调检索器;LLM-Embedder[69]则利用 LLM 为下游任务数据生成奖励信号,结合硬标签协同微调嵌入模型,使其更好拟合特定应用需求,在 MMLU 等基准上效果优于通用 BGE 等模型[69].另外,文献[70]还通过优化视觉模型底层架构实现低计算资源下的性能改进,为嵌入模型的优化提供了不同的视角.

提升嵌入模型对领域知识的敏感度与表征精度,无论是通过知识注入还是领域微调,都是确保 RAG 系统在专业领域提供高质量服务的关键.策略选择需据领域特点、数据可得性及计算资源限制等方面进行权衡.

3.2.2 跨模态嵌入对齐:统一多元信息表示

“文档全谱系”中含大量混合文本、图像、表格、音视频等多模态信息的文档(对应表 3).为使 RAG 系统能全面理解利用这些文档,仅依赖文本嵌入远不足够.因此,发展跨模态嵌入(Cross-Modal Embedding)技术,将不同模态信息映射至同一共享语义向量空间,实现跨模态信息对齐与统一表示,成为重要研究方向.唯此,方能基于语义相似性对含不同模态内容的知识单元进行统一检索与比较.文档识别在此环节的作用体现在,如从图像准确提取文本描述(用于图文对齐),或从视频准确转录语音并识别关键视觉对象,均依赖高质量的文档(多模态)解析.

图文嵌入是跨模态嵌入研究中较成熟方向.CLIP[30]模型是里程碑式工作,其通过在海量图文对上进行对比学习,成功学习到共享嵌入空间,使图像视觉特征向量与描述文本语义向量能相互匹配.CLIP 及其后续变种(如 ALIGN,Florence)极大推动了图文检索、图像标注等任务发展,也为 RAG 处理图文混排文档奠定基础.

对于含更复杂模态组合的文档,如视频文档(含时序变化的视觉帧、音频、字幕、场景文字等),跨模态嵌入挑战更大.ViDoRAG[38]针对视频文档检索,提出基于高斯混合模型(GMM)的混合策略有效融合视频中视觉与文本信息(提取

自音频或字幕),并结合多智能体推理框架检索,在 ViDoSeek 等多模态文档检索基准上取得超 10% 性能提升[38]. 这表明需更精巧模型设计处理时序信息和多模态特征的动态融合.

处理含大量表格的文档亦为重要跨模态问题(结构与文本结合),这在结构复杂的数字原生文档中常见. 表格内容虽主为文本,但其行列结构蕴含关键语义信息. TableGPT[46] 等工作探索如何设计模型从全局视角理解表格结构与语义,支持对表格内容进行复杂自然语言查询与操作. SANTA[71] 框架则通过结构感知预训练任务,助模型学习结构化数据有效表示,提升了结构化数据检索效率.

实现有效跨模态嵌入对齐,是 RAG 系统充分利用"文档全谱系"中日益增长多模态信息的关键. 未来研究需探索更强大多模态融合机制,处理更复杂模态组合(如音视频、文本、表格、图像混合),并确保不同模态信息在共享语义空间的一致性与可比性.

3.2.3　高效索引与向量数据库:支撑规模化检索

当知识库含海量文档,处理嵌入后产生数百万乃至数十亿级向量时,如何高效存储这些向量并支持快速相似性搜索,成为决定 RAG 系统响应速度与可扩展性的关键. 这需依赖高效索引结构(f_{idx})与专门向量数据库技术.

对传统稀疏嵌入(如 TF-IDF 或 BM25[28] 向量),倒排索引是标准高效方案,记录词项所在文档,可快速定位含查询关键词文档.

然对 RAG 中更常用稠密嵌入(高维浮点数向量),倒排索引不再适用. 此时需采近似最近邻(ANN)搜索技术. ANN 旨在牺牲一定精度(找到的可能非绝对最近邻,而是极近邻居)前提下,大幅提升高维空间搜索相似向量速度. 主流 ANN 索引算法分几类:基于树(如 KD-Tree,Annoy)、基于哈希(如 LSH)、基于量化(如乘积量化(PQ)[72] 及其变种)及基于图(如 HNSW[31]). 其中,基于图的 HNSW 通常在精度与速度间能取得较好平衡,是目前多场景首选.

为方便开发者使用这些复杂索引技术并提供向量数据存储、管理、查询等功能,涌现了许多专门向量数据库(Vector Database). 例如,FAISS 是广泛使用的开源库,提供多种高效 ANN 索引实现. Milvus,Pinecone,Weaviate,Qdrant,Chroma 等则是更完整向量数据库系统,提供数据持久化、标量过滤、分布式扩展、多租户、实时更新等更丰富功能. 选择合适 ANN 索引算法和向量数据库,需据数据规模、查询负载、精度要求、延迟容忍度、成本预算等综合考量.

高效索引与向量数据库技术是支撑大规模文档 RAG 系统实现快速响应和良好扩展性的基石. 随向量数据规模持续增长和查询需求日益复杂,对索引构建速度、查询延迟、内存占用、扩展能力及与现有数据生态集成的要求将不断提高,推动该领域技术持续创新.

3.3　高级检索策略优化

构建高质量知识库表示后,检索阶段(f_r)的核心任务是据用户查询 x,智能、准确、高效地从索引 I 中定位并召回最相关知识单元 Z_k. 简单基于向量相似度的检索难应对复杂查询意图、多方面知识需求或需推理场景. 因此,发展更高级检索策略,提升检索过程"智能",是优化 RAG 性能的关键. 文档识别质量在此阶段影响显著:若知识库充斥 OCR 错误导致的无意义或错误文本块(如低质量扫描件带来的问题),即便最先进检索策略也可能召回此类"垃圾"信息,污染后续生成. 图3直观地呈现了高级检索与推理的若干策略,包括如何通

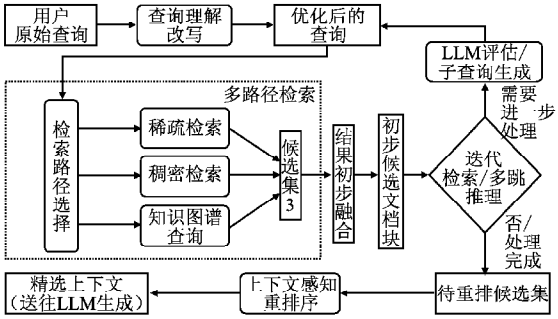

图 3　高级检索策略示意图

Fig. 3　Diagram of advanced retrieval strategies

过查询增强、多路径检索、迭代决策以及最终的重排序来优化信息获取过程.

3.3.1　查询理解与增强:精准把握用户意图

用户原始查询 x 常存模糊性、不完整性,或其表述与知识库文档语言风格存差异(词汇鸿沟). 这致简单向量匹配难准确捕捉用户真实意图,影响检索结果相关性. 因此,检索前对用户查询进行深入理解与适当增强,是提升检索精度的重要首步.

利用 LLM 强大自然语言理解与生成能力处理查询是当前主流. 常见策略是查询重写:让 LLM 将原始查询改写为一个或多个更清晰、具体、信息更丰富的版本. 例如,Query2Query[39] 则用精心设计提示(如 TAO 框架结合 What/How/Why 提问)引导 LLM 生成多个语义等价但表述不同的查询变体,融合其检索结果以提高召回率和鲁棒性,在学科试题检索任务上 F1 值提升 16.5%[39]. Iter-RetGen[73] 则采迭代方式,利用上轮生成结果优化下轮查询,逐步逼近用户意图.

另一策略是查询扩展,即在原始查询中补充相关关键词、同义词或概念. 可通传统伪相关反馈技术实现,或利用 LLM 生成相关扩展词.

还有独特思路是生成假设性文档,如 HyDE[74]. 它不直接改写查询,而是让 LLM 基于查询生成"假设性"答案文档,再用此假设文档嵌入向量检索知识库中语义相似真实文档. Query2doc[75] 也让 LLM 据原始查询生成简短伪文档,再用此伪文档与原始查询一起嵌入检索,相当于用更丰富上下文指导检索. 其逻辑在于,这种答案文档(即便虚构)应在语义上接近含真实答案文档.

此外,特定场景下查询处理可能更复杂. 例如,BioRAG[22] 先对生物医学查询分类,再据不同查询类型(事实型、定义型等)采不同优化策略. 结合知识图谱的 RAG 中,还需将自然语言查询转为结构化图查询[48]. 某些系统中,查询增强亦可作补救措施,如初步检索结果不佳或被过滤时,触发查询重写机制二次尝试[76]. 文献[77]为了减少上下文窗口限制,设计了文本过滤器来过滤最终的候选文档,若候选文档全被过滤,则采用查询重写的方法进行重新检索.

运用 LLM 深度分析与生成能力,对用户原始查询进行理解、重构、扩展或转换,正成为提升后续信息召回准确性与全

面性的重要手段. 将模糊用户意图转为机器更易理解执行的精准检索指令,能有效减少因表述不清或词汇不匹配导致的检索偏差.

3.3.2 混合与多路径检索:融合多元优势提升鲁棒性

单一检索策略常有固有限制,难适应所有查询与知识库特性. 例如,传统稀疏检索(如 BM25[28])基于词频与逆文档频率,擅长捕捉关键词精确匹配,对含特定术语或实体名查询效果好,但难理解词语深层语义与同义关系. 而稠密检索(基于向量嵌入)通过高维空间度量语义相似性,能好处理同义词、近义词和概念匹配,理解查询整体语义意图,但可能对具体关键词不够敏感,有时会召回语义相关但非精确答案文档.

为克服单一策略不足,混合检索应运而生. 其核心是结合稀疏与稠密检索优势,以期获更鲁棒、全面检索结果. 实现方式多样,如并行执行两种检索后据得分加权融合(如简单线性组合或更复杂学习排序算法如 RRF),或采两阶段策略(如先用稀疏检索快筛候选集,再用稠密检索精排). 实践表明,即便简单混合策略,通常亦优于单一策略[76].

更进一步,研究者开始探索多路径检索与动态路由机制. 此法不仅实现稀疏与稠密检索结合,还可能引入其他信息源或检索方式,如知识图谱查询[50,65](利用图谱结构化关系精确查找)、基于元数据过滤[40](如据文档日期、作者、类型筛选)、乃至外部搜索引擎调用. HiQA[35] 设计了复杂多路径检索器,能动态调整稀疏、稠密及基于实体识别关键词检索三路径混合参数. HyPA-RAG[41] 更进一步,用机器学习模型(如逻辑回归、SVM、BERT)对输入查询进行复杂度或类型分析,然后自适应选择或融合稠密、稀疏与图谱检索路径,在 LL144 法律政策问答基准上显著提升答案正确性、忠实度与召回率[41]. 特定应用如工业故障诊断[23],可通串行意图分类模型判断故障现象,并选择后续检索或处理流程. BioRAG[22] 亦用基于 Llama-8B 训练分类器决定后续调用哪些检索组件或知识源. 文献[78]就提出更复杂的静态、动态两阶段检索流程,并设计了 7 种不同的检索算法,并采用分类器对最终检索结果进行相关性过滤. 文献[62]也在事件抽取流程中,增加了预检索流程,并采用了轻量化的触发词过滤与贝叶斯概率筛选的预检索方法,快速获取与输入文本相关的候选信息集合用于后检索流程.

融合多检索范式并据查询特性动态路由的混合与多路径检索策略,正成提升信息获取鲁棒性与覆盖面主流方向. 此灵活性使 RAG 系统能更好适应不同信息需求、知识库特性及对精度效率的不同要求,从而在复杂多变真实场景中表现出更强适应能力.

3.3.3 迭代检索与多跳推理:攻克复杂知识密集型任务

许多现实问题,尤其需深入分析、比较或综合信息的复杂问题,常无法通过单次简单信息检索解答. 此类问题通常需从多个不同文档片段收集证据,并进行多步逻辑推理方能得出结论. 为应对此挑战,RAG 系统需发展超越单轮检索的迭代检索与多跳推理能力. 若文档识别阶段引入噪声致关键信息片段缺失或错误,多跳推理链条易在早期中断或走向错误方向.

迭代检索核心是将检索过程变为多轮、动态调整过程. 系统可据初步检索信息和当前理解状态,主动、迭代地生成新查询或调整检索策略,以搜集更全面相关证据,或对初步结果验证补充. 例如,系统发现初步检索信息不足以回答问题时,可生成更具体子查询二次检索[74]. 或利用 LLM 评估初步检索结果,判断信息相关性与缺失性,并据此指导下轮检索[22]. 此迭代过程允许系统逐步逼近答案,尤其适用知识分布分散或需探索性查找场景[76].

多跳推理更侧重在检索到的多个知识片段间建立逻辑联系. 它要求系统不仅找到相关单个信息点,还能理解其间关系(如因果、时序、从属),并沿关系进行一步或多步推理. 例如,回答"A 药与 B 药同用风险?"可能需先检索 A 药副作用,再检索 B 药副作用,最后检索两者已知相互作用信息,并综合评估. MultiHop-RAG[79] 针对此类问题提出基准数据集和典型多跳推理流程,包括初始检索、推理规划(判断是否需更多信息及如何获取)、递进检索、证据链整合等.

为实现更智能迭代检索与多跳推理,研究者探索了多种先进框架与方法. Self-RAG[16] 赋予 LLM 在生成中自主决策何时需检索及如何评价检索结果能力,实现检索与生成深度融合. DeepRAG[17] 将复杂检索推理建模为马尔可夫决策过程(MDP),训练策略网络动态规划检索步骤(继续检索、检索何物、或停止生成),在 HotpotQA 等多跳问答数据集上显著提升性能并减少检索次数[17]. Adaptive-RAG[43] 据用户查询复杂度预估,动态选择不同 RAG 策略(无需、单步或多步迭代检索),在保证效果同时优化效率[43]. Chain-of-Verification (CoVe)[34] 虽主旨提升事实性,但其"规划-生成-验证"迭代流程,亦内在地支持多步推理与信息整合.

赋予 RAG 系统迭代搜集信息、评估证据、进行深度多跳推理能力,是其从简单信息查找器向真正知识处理器和问题解决者转变的关键. 这通常需检索与生成组件更紧密协同,甚至引入规划、决策和自我反思机制,代表 RAG 技术未来重要发展方向. 未来可以结合 LLM 智能体架构[14],赋予系统更强的自主规划、迭代检索和多步推理能力.

3.3.4 上下文感知重排序:精选检索结果提升生成质量

初步检索(召回)阶段,为保证高召回率,系统常返回较多(如 Top-K,K 可达数十上百)候选文档块 Z_k. 然这些候选块相关性参差不齐,既有高度相关核心信息,亦可能混杂部分相关乃至无关噪声. 若将所有候选块不加区分直接输入后续生成模型(LLM),不仅增加 LLM 处理负担与计算成本(尤其上下文窗口有限时),还可能因噪声干扰降低最终答案质量与准确性. 文档识别引入的格式噪声可能致一些不相关文本片段因表面结构相似被误召回,此时重排序更显重要.

为解决此问题,重排序(Reranking)作为检索流程关键优化环节应运而生. 其目标是在初步召回候选集 Z_k 基础上,利用更强大、精细模型,结合原始查询 x 上下文,对候选块进行重新打分排序,筛选出与查询意图最匹配、信息含量最高知识片段,并置于前列.

与召回阶段常用计算相对简单的相似度度量(如向量内积)不同,重排序阶段可使用计算更密集但通常更准确模型,如交叉编码器(Cross-Encoder)模型[32]. 交叉编码器将查询 x 与每候选文档块 z_i 拼接作输入,让模型(通常基于 Transformer 架构)能充分交互建模,从而更精确判断两者相关性. 虽交叉编码器计算开销远大于双编码器(召回常用),但因其仅处

理初步召回的少量候选集(K 个),故整体效率通常可接受.

研究表明,引入重排序环节能显著提升 RAG 系统端到端性能. DSLR 框架[44]提出句子级重排策略:先将召回文档块分解为句子,再用微调 MonoT5 交叉注意力模型对每句与查询相关性打分排序,最后动态选择得分最高句子构建最终输入 LLM 上下文. 实验表明,此精细化重排不仅在 NQ、TQA 等多标准 QA 任务上平均准确率提升 6.3%,还能显著减少输入 LLM 的 Token 数(平均减 53.4%),从而在提升效果同时优化效率[44]. Meta-RAG[40]在电力领域问答任务的显著成功亦部分归功其混合编码与重排序策略[40].

需注意,训练高性能重排序模型常需大量相关性标注数据(如给定查询与文档块,标注相关等级)[80],这在某些特定领域可能是获取瓶颈. 因此,探索无监督或弱监督重排序方法,或利用 LLM 自身判断能力进行零样本重排序[76],亦是当前研究方向.

总之,初步召回后引入上下文感知重排序模块,利用更精细模型评估筛选与用户查询最相关知识片段,是提升最终生成内容质量、减少噪声干扰、优化 LLM 输入效率的重要步骤,已成构建高性能 RAG 系统常用实践.

3.4　生成控制与可信度增强

检索到相关知识片段 Z_k 后,最终挑战在于如何让生成组件(f_g)有效利用这些信息,生成流畅自然、忠实可靠且满足特定任务需求的答案 y. 此阶段优化聚焦于增强生成内容事实一致性、缓解幻觉、提供透明溯源,并确保生成过程安全可控. 若检索上下文 Z_k 中含文档识别错误(如 OCR 语义噪声,常见于低质量扫描件)导致的错误信息,生成器在缺乏有效校验机制时,很可能将这些错误复述或整合入最终答案,严重损害答案可信度.

3.4.1　增强事实一致性与忠实度:让生成紧贴证据

确保 LLM 生成答案忠实于检索文档内容 Z_k,而非凭空捏造或引入其内部参数知识中可能存在的错误信息,是提升 RAG 系统可信度的核心. 为此,研究者探索多种策略强化生成过程对所供证据的依赖性.

常用方法之一是通过精心设计提示工程(Prompt Engineering)引导 LLM. 例如,提示中明确指示 LLM“请仅据以下所供上下文信息回答,若无相关信息,则答‘信息不足’”. 或采思维链(CoT)[33]提示,要求 LLM 生成最终答案前,先显式列出其基于哪些检索证据片段进行何种推理步骤[23],使生成过程更透明有据.

模型参数微调是增强生成控制的有效范式. RAFT[18]通过在训练中混合黄金文档与干扰文档,迫使模型获得忽略无关内容并精准引用原文的能力,显著提升噪声环境下的生成鲁棒性. PA-RAG[81]则采用“基础 SFT 微调 + 多阶段 DPO 偏好优化”的策略,先构建高可信引用样本进行指令微调奠定生成基础,再通过信息完备性、抗干扰性、和引用精确性三维度精细化调整生成偏好. 这种分层渐进式的两阶段优化策略,实现了生成行为与可信度要求的深度对齐,对比基线模型性能显著提升,其中 EM 分数平均提升 13.97%,引用召回率提升 49.77%,引用精确度提升 39.58%.

另一方法是约束生成过程. 例如,FastRAG[50]处理半结构化数据时,利用从数据中学到的 JSON Schema 严格控制 LLM 输出格式,确保输出符合预期结构. 然而,仅依赖这类基于格式的硬约束,在处理更普遍的非结构化文本时,往往难以解决语义层面的幻觉问题.

为确保生成内容在语义上忠实于检索证据,一个更精细化的前沿方向是应用“可控文本生成”(Controllable Text Generation,CTG)领域的先进技术,在解码时动态施加语义约束. 其核心思想是在生成每个词元(Token)时,实时地干预模型的概率分布. 这类“即插即用”(Plug-and-Play)方法无需微调生成器本身,灵活性很高. 例如,可控文本生成领域的 PPLM[82]等代表性工作,利用一个外部的鉴别器模型来引导生成方向,惩罚那些可能偏离证据内容的词元. 更直接地,Keyword2Text[83]等工作探索了通过计算词汇与证据之间的语义相似度(如余弦距离),来直接增强相关词元的生成概率. 将这种思想迁移至 RAG 系统时,其“目标约束”便是检索到的外部证据 Z_k.

不过,直接应用此类细粒度的解码控制,需审慎考虑两个潜在的挑战:其一是效率考量,对每个生成的词元都进行动态的语义计算和权重调整,可能显著增加计算开销,影响系统的实时响应能力;其二是生成质量风险,若过度或不当地干预,可能导致模型输出在局部看似忠实,但牺牲了全局的连贯性和可读性.

为应对上述挑战,DeAL[84]将文本生成视为一个启发式引导的搜索问题. 它采用前瞻(Lookahead)机制来提升效率,并通过将外部的语义对齐奖励 h 与模型原始的生成概率 $\log p$ 进行加权融合,来平衡控制强度与生成质量.

$$Score = \log P + \lambda h \tag{3}$$

实验证明了这类先进框架的有效性. 在长度约束的摘要任务中,DeAL 能将满足约束的摘要比例从 16% 大幅提升至 73%,同时在流畅性、相关性等核心质量指标上与基线方法相比无统计学上的显著差异. 在更严格的“无害性”对齐任务中,该框架甚至能将有害内容的生成率从 57% 降至 0%. 这些成果表明,将这类先进的解码控制技术应用于 RAG 的生成环节,为解决其事实一致性与幻觉问题,提供了一条极具潜力的技术路径.

此外,可引入外部验证机制. 如生成初步答案后,设计验证模块(可能亦为 LLM),据检索证据检查答案陈述事实性,并打分或反馈. Chain-of-Verification(CoVe)[34]就通过生成验证问题交叉检查初步答案.

强化生成过程对检索证据的忠实度,无论是通过参数微调、提示工程、可控解码还是外部验证,皆为构建可靠 RAG 系统基础. 这有助降低模型偏离事实依据、产生不准确或误导性信息风险,提升用户对系统输出信任. 未来还应思考如何将这些异构的控制方法进行协同(不应固守某一种策略),根据查询的复杂性、任务的具体要求动态地选择和组合最合适的控制手段.

3.4.2　缓解幻觉与确保安全:应对不可靠与恶意风险

尽管 RAG 旨在通过引入外部知识减少 LLM 幻觉[4],但幻觉在 RAG 系统中依然可能发生,尤其在检索信息不完整、矛盾或 LLM 未能正确理解上下文时. 此外,RAG 系统亦面临新安全风险,如可能检索到恶意、污染或带偏见信息并将其不加批判用于生成,或系统本身遭对抗性攻击干扰. 文档识别阶

段错误,如 OCR 未能正确识别文档中免责声明或警告,可能致生成器输出不完整或误导性内容,带来安全隐患. 为主动检测和缓解生成内容中潜在幻觉,研究们提出更为主动的机制. Self-RAG[16] 在其模型设计中融入自我反思与批判能力,允许模型生成中评估检索信息是否相关、生成内容是否获证据支持、是否存在幻觉,并据此修正. CoVe[34] 通过其“规划-生成-验证”迭代流程,系统性发现并纠正潜在事实性错误,在 MultiSpanQA 任务上将 Llama 65B 的 FACTSCORE 从 55.9 显著提升至 71.4[34]. CRAG[85] 则在检索后增设轻量级评估器判断检索结果质量,并据此决定是直接利用结果生成、进行网络搜索获取更可靠信息,还是重新检索.

针对安全性问题,研究日益关注 RAG 系统对对抗性攻击的脆弱性. GARAG[86] 研究表明,即便文档中引入少量精心设计拼写错误(一种 OCR 语义噪声模拟),亦可能显著干扰 RAG 系统检索与生成,致性能大幅下降[86]. BadRAG[87] 则揭示通过在知识库注入少量“毒化”数据进行后门攻击的可能性. 这些发现凸显研发鲁棒防御机制的紧迫性. 例如,RAAT[88] 等工作探索通过对抗训练提升 RAG 系统对噪声与扰动抵抗能力. 未来研究需从架构层面设计内在安全机制,如开发能识别对抗性扰动的检索器和具“批判性思维”的生成器. 当验证过程需要调用外部数据库或专业 API 时,MCP[27] 还可为此提供标准化的交互机制,允许 RAG 系统在必要时查询外部事实核查服务或动态更新的知识源,进一步提升其纠错和幻觉缓解能力,这代表了 RAG 向更开放、动态的知识交互演进的重要方向.

多模态文档还为生成可信度带来一些独特挑战:图文/音画不一致导致的跨模态冲突易引发矛盾描述,如配文与图像主体错位造成的语义割裂[38,89];VLLM 对长上下文的位置偏差使检索结果增多反而降低生成质量,ViDoRAG 实验观测到 ROUGE-L 随 K 值增加呈反比下降的异常现象[38];文本描述与视觉细节的粒度失配则可能诱发细节幻觉,尤其当问题需精细视觉信息而检索内容仅提供宏观描述时[89]. 为应对这些挑战,ViDoRAG[38] 通过动态探索-总结-反思循环的多智能体策略实现迭代式跨模态校验. mRAG[89] 则提出了一种统一的智能体框架,通过自反思机制将重排序与生成动态整合并,实现噪声自适应过滤,这些创新方案通过建立跨模态理解与自省能力的协同机制,为多模态可信生成开辟了新方向.

构建含自我评估、事实核查、迭代修正的闭环机制,并增强系统对潜在风险(含内部幻觉和外部攻击)识别与防御能力,是确保 RAG 系统输出可靠、安全的关键. 这要求 RAG 系统具备更强自我反思、纠错能力与安全意识. 值得注意的是,采用多智能体协作框架正成为应对这些挑战的重要趋势. 无论是单模态还是多模态场景下的优化策略,其核心都在于通过多个具备特定功能(反思、验证、过滤、防御)的智能体协同工作. 这种协作架构将反思、验证、纠错与安全防护等能力模块化并动态联动,显著提升了系统处理复杂信息、抵御风险的内生能力.

3.4.3 细粒度溯源与引用:构建透明可信的答案

为使用户理解答案来源、验证准确性并建立信任,RAG 系统需提供清晰、准确、易追溯的来源信息,即溯源或引用能力. 理想溯源机制应能将生成答案中具体陈述或事实,精确追溯至原始知识库中一个或多个具体文档片段或句子. 若文档识别阶段未能准确划分文档结构(如页码、章节、段落),即便生成器欲提供准确溯源亦会非常困难.

早期 RAG 系统[5] 或仅简单指出答案主要基于哪些检索文档块,或利用解码器注意力机制粗略关联生成内容与输入上下文. 随技术发展,溯源机制日益细粒度化. 例如,RAFT[18] 在微调阶段即训练模型生成答案同时,直接引用原始文档相关片段. HiQA[35] 则致力将溯源信息精确到文档 ID 与文档内部章节路径(通过其设计层次化上下文增强模块 HCA 实现),并在 MasQA 基准测试中提升答案充分性(Adequacy)指标[35].

追求更细粒度溯源是当前研究趋势. LongCite[36] 探索让 LLM 生成句子级引用,即将答案中每关键句链接至支持其的源文档句子. 当 RAG 系统结合知识图谱时,还可提供基于图谱路径或三元组的结构化溯源[90],如指出某结论基于图谱中实体 A 与实体 B 间某种关系得出.

然而实现高质量细粒度溯源仍面临挑战. 例如,当答案是综合多来源信息推理得出结果时,如何清晰展示此融合过程及各来源贡献? 当检索信息存细微差异或矛盾时,溯源应指向何版本? 如何设计用户友好界面展示溯源信息,方便用户快速核查? 这些均需进一步研究.

增强系统透明度与可解释性,是构建负责任且可信赖 AI 系统关键一环. 细粒度溯源机制正从简单文档级链接,向更精确段落、句子乃至知识图谱实体关系级别发展. 为用户提供可靠、易核查答案来源,不仅提升用户信任,亦是未来实现更复杂人机协作与知识探索基础.

表 5 关键技术环节的演进趋势与核心问题

Table 5 Trends and issues of strategies in RAG

技术环节	演进趋势	核心问题/难点
复杂文档解析	端到端多模态大模型解析(VLLMs, Speech LLMs),增强对复杂结构(图文表、音频)的理解.	低质文档鲁棒性,细粒度结构提取精度,多模态信息深度融合.
智能分块策略	语义感知、结构自适应、动态分块,结合长上下文处理.	平衡语义完整性、检索粒度与计算效率,适应不同文档与任务.
文档图谱构建与利用	LLM 简化图谱构建,图结构增强检索深度与推理.	大规模图谱自动化构建效率与质量,图谱与向量检索有效融合.
领域适应性嵌入	微调、注入领域知识、跨模态对齐,提升专业领域与多模态语义表达.	高效获取利用领域知识,克服通用模型局限,处理多模态语义鸿沟.
高级检索与推理	查询增强、混合检索、迭代搜集与多跳推理,处理复杂查询和知识密集任务.	复杂查询意图理解与分解,多源冲突信息消解,长链推理可靠性与效率,避免错误累积.
可信生成与溯源	强化证据忠实度,幻觉检测与自动纠错,提供更细粒度溯源.	兼顾事实性与流畅性,应对对抗干扰,溯源准确性与易用性.
与外部系统交互(MCP)	标准化协议(MCP)调用外部工具、服务或实时数据,扩展能力边界.	交互安全性、效率,多智能体协作管理,无缝集成外部能力.

对复杂文档处理、领域适应性嵌入、高级检索与推理以及可信生成与溯源等关键技术的深入探讨,揭示了面向文档的 RAG 技术在不断演进以应对日益复杂的需求. 表 5 对本节所

讨论的这些核心技术环节的共同发展趋势及关键问题进行了总结.

4　挑战与展望

4.1　当前挑战

尽管面向文档的RAG技术取得了显著进展,但在迈向更广泛、更深入应用的道路上,依然面临多方面的核心挑战,可归纳为以下几个层面:

1)知识源头的根本性挑战. 文档理解的深度与广度瓶颈依然突出. 现有技术在精准解析具有复杂结构(如图表、公式、嵌套布局)的文档,尤其是低质量扫描件[25]时,仍显力不从心. 深度融合并理解文档内嵌的多模态信息(如文本与图像、音频的语义关联)[57,91],以及高效处理并从中提取关键信息的超长文档[2,58,92],均存在显著局限. 尤为关键的是,文档识别(特别是OCR)过程引入的语义噪声(如词义篡改)和格式噪声(如结构误判)[51],会对整个RAG流程的每一个后续环节构成级联式的负面影响,从根本上制约系统性能.

2)知识的复杂性挑战. 当检索模块返回多个可能包含冲突、冗余甚至错误信息的知识片段时,如何有效地进行信息筛选、去伪存真、权重分配,并在此基础上进行可靠的多跳逻辑推理以形成连贯答案[35,79],对现有模型(比如私有部署的大模型)仍是严峻考验. 同时,通用预训练的嵌入模型在应用于特定专业领域时,往往因缺乏领域知识而表现出适应性不足,难以准确捕捉细微的语义差别[21,22].

3)策略与场景的适配鸿沟(Strategy-Scenario Mismatch). 当前RAG研究往往聚焦单个环节的技术优化,但缺乏一个系统性的方法论来指导实践者根据具体的文档类型、业务需求和问题复杂度,选择和组合最优的技术栈. 这导致在实际落地时,常常出现设计低效、技术选型不当无法解决核心问题等困境. 如何建立自适应、可配置的RAG领域适配方法体系,实现从文档分析到策略推荐的自动化推荐,是当前面临的重大工程挑战.

4)系统层面的综合性挑战. 当前的RAG系统对输入噪声(包括查询噪声和文档本身含有的噪声)以及潜在的对抗性攻击[86,87]表现出一定的脆弱性,鲁棒性亟待提升. 在追求生成内容事实性的同时,如何维持回答的流畅性、完整性和信息量,往往需要在可信性与回答质量间进行微妙权衡[34]. 此外,尽管溯源机制有所发展,但提供真正细致入微、准确无误且用户易于理解和验证的溯源信息[35,36],仍然是一个尚未完全解决的难题. 效率与可扩展性也成为制约实际部署的关键因素,复杂的处理流程可能导致较高的查询延迟[17,44],而大规模知识库的索引与维护成本依然高昂[31,72].

5)评估与应用的生态挑战. 目前缺乏能够全面、标准化、多维度地衡量RAG系统性能的基准测试集和评估方法[79,93-95],特别是在特定垂直领域和面向复杂真实任务场景时,现有评估难以反映真实效能. 更重要的是,缺乏系统性评估文档识别质量对RAG各核心组件具体级联影响的标准化基准[51]. 此外,如何将RAG技术与特定垂直领域的深层业务流程、专有知识体系以及合规性要求进行有效且深度的结合[15],以及复合型专业人才的匮乏,是技术能否真正落地并创造价值的关键.

4.2　未来发展方向

面对当前挑战,面向文档的RAG技术有望在以下几个关键方向持续演进与突破,推动其向更智能、更可靠、更高效的未来迈进.

1)深化文档理解,迈向文档认知层面. 未来的研究将致力于研发更为强大的端到端多模态基础模型[7,57,91],这些模型不仅能处理文本,更能原生理解和融合图像、表格、版式乃至音视频等多种模态信息. 同时,深化领域自适应嵌入技术[22,66,69],使其能更精准捕捉特定行业的细微语义. 核心目标是实现对"文档全谱系"中各类复杂文档结构与内容的深层认知,而非仅仅是表面信息的提取. 在此过程中,需特别关注持续提升文档识别(尤其是OCR和复杂版面分析)的准确性与鲁棒性,并探索开发专门针对RAG下游任务需求进行优化的文档解析与表示方案[51],从源头上保障知识质量.

2)融合检索与推理,构建统一的智能框架. 未来的RAG系统将不再是简单的检索与生成模块的串联,而是会朝着更深度的融合方向发展. 这包括研究更有效的跨文档知识融合机制,解决信息冲突与冗余问题,并从多源信息中提炼一致性知识. 同时,发展具备自主规划能力的迭代式、自适应检索与推理框架[17,43],使系统能根据问题复杂度和当前信息状态动态调整策略,进行多步推理. 最终目标是探索检索、理解、推理与生成过程的深度一体化,形成一个更为统一和智能的知识处理引擎.

3)构建自适应与可组合的RAG框架 . 未来的RAG系统将不再是固化的流水线,而是一个高度模块化、可动态编排的智能系统. 该系统能够根据文档和查询,自动识别其类型、结构复杂度、领域归属等关键特征,然后动态地从选择并组合最合适的策略组件(如选择特定的解析器、分块策略、检索模型组合). 这要求研究者们要继续研究技术之间的协同与编排机制,最终实现自适应的智能化RAG服务.

4)提升系统韧性,兼顾鲁棒、效率与可信. 为应对现实应用中的不确定性,需要显著增强RAG系统对各类噪声(包括源于文档识别的噪声[51])和恶意对抗攻击的防御能力与恢复能力[88]. 在效率方面,通过模型压缩、知识蒸馏、更先进的高效索引技术[96]、智能缓存策略以及硬件加速等手段,持续优化系统响应速度和处理吞吐量. 在可信性方面,除了发展更强大的事实验证[34]与自动纠错机制外,还需构建更精确、更具解释性且支持用户交互探索的溯源体系[36],让用户不仅知其然,更能知其所以然.

5)扩展交互边界,拥抱开放的智能生态. 未来的RAG系统将更加开放,能够与外部世界进行更广泛和深入的交互. 利用诸如模型上下文协议(MCP)[27]等标准化接口和框架(如图4所示,MCP作为桥梁连接RAG系统与多样化的外部能力和数据源),RAG系统将能更安全、高效地调用外部API、查询实时动态数据库、执行代码、使用计算工具,甚至与其他专门的AI服务进行协作,从而突破自身知识边界和能力局限,解决更复杂和动态的问题. 这不仅能极大扩展RAG获取和利用实时信息的能力,还可能催生出全新的、基于RAG的组合式AI应用,使其在因果推理、反事实推理等高级认知任务中发挥更大作用. 实现更具个性化和交互性的RAG系统,

允许用户参与到检索和生成过程中，也是一个重要的发展方向.

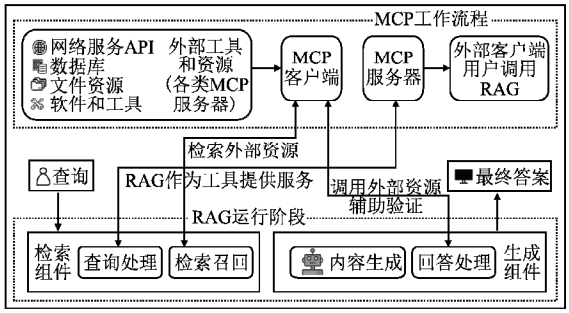

图 4 模型上下文协议(MCP)应用场景示意图

Fig. 4 Schematic diagram of application scenarios for model context protocol(MCP)

6)完善评估体系与构建健康生态. 亟需构建更全面、更细粒度、更贴近真实应用场景的标准化评估基准与方法论. 这不仅包括对 RAG 整体性能的评估，还应特别关注对文档识别质量及其对后续各环节具体级联影响的细致量化评估(例如借鉴 OHRBench[51] 的理念，并将其扩展至更广泛的文档类型和噪声模式). 同时，加强跨学科人才培养，鼓励开源社区建设，并积极研究能够进行在线学习、持续学习和快速适应新知识、新领域、新任务的自适应 RAG 系统，是推动该技术领域健康、快速发展的基石.

5 总 结

检索增强生成(RAG)已成为增强 LLM 应用效能的关键手段. 本文从文档的视角出发，构建了一个形式化的四组件核心框架，并围绕文档处理与表示优化、嵌入与索引优化、高级检索策略优化、生成控制与可信度增强等关键环节，深入梳理和凝练了各项技术的最新进展，强调了“文档处理”作为 RAG 系统首要且关键环节的重要性，认为 RAG 成功应用的关键在于深入任务需求，深刻把握应用领域的文档特性与多重挑战，有效设计最优策略组合. RAG 技术正向着一个更智能、更动态、更具适应性、具备初步认知能力的复杂系统演进. 未来，面向文档的 RAG 技术将不断演进，有望实现从“检索增强”到“认知增强”的飞跃，为人类处理和利用日益增长的海量文档知识提供更强大、更可靠、更智能的支持.

需要指出，本综述侧重于技术框架、挑战与发展方向，对 RAG 的评估方法论和基准测试未作深入探讨. 同时，对多模态 RAG 应用的具体场景和细节的讨论也相对有限，这些均可作为未来工作的补充.

References:

[1] Vaswani Ashish, Shazeer Noam, Parmar Niki, et al. Attention is all you need[C]//31st International Conference on Neural Information Processing Systems, 2017: 5998-6008.

[2] Gao Yunfan, Xiong Yun, Gao Xinyu, et al. Retrieval-augmented generation for large language models: a survey[J]. arXiv preprint arXiv: 2312.10997, 2023.

[3] Ji Ziwei, Lee Nayeon, Frieske Rita, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[4] Huang Lei, Yu Weijiang, Ma Weitao, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions[J]. arXiv preprint arXiv: 2311.05232, 2023.

[5] Lewis Patrick, Perez Ethan, Piktus Aleksandra, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//34th International Conference on Neural Information Processing Systems, 2020: 9459-9474.

[6] Touvron Hugo, Lavril Thibaut, Izacard Gautier, et al. LLaMA: open and efficient foundation language models[J]. arXiv preprint arXiv: 2302.13971, 2023.

[7] Bai Shuai, Chen Keqin, Liu Xuejing, et al. Qwen2.5-VL technical report[J]. arXiv preprint arXiv: 2405.13923, 2024.

[8] Liu Aixin, Feng Bei, Xue Bing et al. DeepSeek-V3 technical report [J]. arXiv preprint arXiv: 2406.19437, 2024.

[9] Goyal Tanya, Li Junyi Jessy, Durrett Greg, et al. News summarization and evaluation in the era of GPT-3[J]. arXiv preprint arXiv: 2209.12356, 2022.

[10] Workshop BigScience, Scao Teven Le, Fan Angela, et al. BLOOM: a 176B-parameter open-access multilingual language model[J]. arXiv preprint arXiv: 2211.05100, 2022.

[11] Achiam Josh, Adler Steven, Agarwal Sandhini, et al. GPT-4 technical report[J]. arXiv preprint arXiv: 2303.08774, 2023.

[12] Sun Yu, Wang Shuohuan, Feng Shikun, et al. Ernie 3.0: large-scale knowledge enhanced pre-training for language understanding and generation[J]. arXiv preprint arXiv: 2107.02137, 2021.

[13] Zhang Tianyi, Zhang Chenxi, Peng Xin, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21(1): 5485-5551.

[14] ZHANG T Y, ZHANG C X, PENG X, et al. A review of research on software configuration based on large language models[J]. Computer Applications and Software, 2025, 42(3): 1-12.

[15] Spirling Arthur. Why open-source generative AI models are an ethical way forward for science[J]. Nature, 2023, 616(7957): 413, doi: 10.1038/d41586-023-01295-4.

[16] Asai Akari, Wu Zeqiu, Wang Yizhong, et al. Self-RAG: learning to retrieve, generate, and critique through self-reflection[C]//12th International Conference on Learning Representations, 2024: 41056-41085.

[17] Guan Xinyan, Zeng Jiali, Meng Fandong, et al. DeepRAG: thinking to retrieval step by step for large language models[J]. arXiv preprint arXiv: 2502.01142, 2025.

[18] Zhang Tianjun, Patil Shishir G, Wen Kai, et al. RAFT: adapting language model to domain specific RAG[J]. arXiv preprint arXiv: 2403.10131, 2024.

[19] Yepes Antonio Jimeno, You Yao, Milczek Jan, et al. Financial report chunking for effective retrieval augmented generation[J]. arXiv preprint arXiv: 2402.05131, 2024.

[20] Fan Wenqi, Ding Yujuan, Ning Liangbo, et al. A survey on RAG meeting LLMs: towards retrieval-augmented large language models [J]. arXiv preprint arXiv: 2405.06211, 2024.

[21] Wahidur Rahman S. Legal query RAG[J]. IEEE Access, 2024, 13: 36978-36994, doi: 10.1109/access.2025.3542125.

[22] Wang Chengrui, Long Qingqing, Xiao Meng, et al. BioRAG: a

RAG-LLM framework for biological question reasoning[J]. arXiv preprint arXiv:2408.01107,2024.

[23] HE Z,JIANG B,WANG X. Improved retrieval-augmented and llm of chain-of-thought maintenance strategy generation[J]. Computer Applications and Software,2024,42(3):1-6+83.

[24] Sharma Shagun,Kataria Anjali,Sandhu Jasminder Kaur,et al. Applications,tools and technologies of robotic process automation in various industries[C]//International Conference on Decision Aid Sciences and Applications,2022:1067-1072.

[25] Abdellaif Osama,Nader Abdelrahman,Hamdi Ali,et al. ERPA:efficient RPA model integrating OCR and LLMs for intelligent document processing[J]. arXiv preprint arXiv:2402.19840,2024.

[26] Peng Jing,Wang Yucheng,Li Bohan,et al. A survey on speech large language models[J]. arXiv preprint arXiv:2403.18908,2024.

[27] Hou Xinyi,Zhao Yanjie,Wang Shenao,et al. Model context protocol(MCP):landscape,security threats,and future research directions[J]. arXiv preprint arXiv:2503.23278,2025.

[28] Robertson Stephen,Zaragoza Hugo. The probabilistic relevance framework:BM25 and beyond[J]. Foundations and Trends in Information Retrieval,2009,3(4):333-389.

[29] Devlin Jacob,Chang Ming Wei,Lee Kenton,et al. BERT:pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805,2018.

[30] Radford Alec,Kim Jong Wook,Hallacy Chris,et al. Learning transferable visual models from natural language supervision[C]//38th International Conference on Machine Learning,2021:8748-8763.

[31] Malkov Yu. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(4):824-836.

[32] Glass Michael,Rossiello Gaetano,Chowdhury Md Faisal Mahbub,et al. Re2G:retrieve,rerank,generate[J]. arXiv preprint arXiv:2202.03629,2022.

[33] Wei Jason,Wang Xuezhi,Schuurmans Dale,et al. Chain-of-thought prompting elicits reasoning in large language models[C]//36th International Conference on Neural Information Processing Systems,2022:24824-24837.

[34] Dhuliawala Shehzaad,Komeili Mojtaba,Xu Jing,et al. Chain-of-verification reduces hallucination in large language models[J]. arXiv preprint arXiv:2309.11495,2023.

[35] Chen Xinyue,Gao Pengyu,Song Jiangjiang,et al. HiQA:a hierarchical contextual augmentation RAG for multi-documents QA[J]. arXiv preprint arXiv:2402.01767,2024.

[36] Zhang Jiajie,Bai Yushi,Lv Xin,et al. LongCite:enabling LLMs to generate fine-grained citations in long-context QA[J]. arXiv preprint arXiv:2409.02897,2024.

[37] Bei Yijun,Fang Zhibin,Mao Shenyu,et al. Manufacturing domain QA with integrated term enhanced RAG[C]//International Joint Conference on Neural Networks,2024:1-8.

[38] Wang Qiuchen,Ding Ruixue,Chen Zehui,et al. ViDoRAG:visual document retrieval-augmented generation via dynamic iterative reasoning agents[J]. arXiv preprint arXiv:2502.18017,2025.

[39] WANG H,SHI Y. Research on query extension method based on large language model[J]. Computer Technology and Development,2024,35(3):148-155.

[40] WANG H,WEI J,JING H,et al. Meta-RAG:a metadata-driven retrieval augmented generation framework for the power industry[J]. Computer Engineering,2024,50(2):1-11.

[41] Kalra Rishi,Wu Zekun,Gulley Ayesha,et al. HyPA-RAG:a hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications[C]//1st Workshop on Customizable NLP,2024:237-256.

[42] Sarmah Bhaskarjit,Hall Benika,Rao Rohan,et al. HybridRAG:integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction[J]. arXiv preprint arXiv:2408.04948,2024.

[43] Jeong Soyeong,Baek Jinheon,Cho Sukmin,et al. Adaptive-RAG:learning to adapt retrieval-augmented large language models through question complexity[J]. arXiv preprint arXiv:2403.14403,2024.

[44] Hwang Taeho,Jeong Soyeong,Cho Sukmin,et al. DSLR:document refinement with sentence-level re-ranking and reconstruction to enhance retrieval-augmented generation[C]//3rd Workshop on Knowledge Augmented Methods,2024:73-92.

[45] Blecher Lukas,Cucurull Guillem,Scialom Thomas,et al. Nougat:neural optical understanding for academic documents[J]. arXiv preprint arXiv:2308.13418,2023.

[46] Zha Liangyu,Zhou Junlin,Li Liyao,et al. TableGPT:towards unifying tables,nature language and commands into one GPT[J]. arXiv preprint arXiv:2307.08674,2023.

[47] Peng Boci,Zhu Yun,Liu Yongchao,et al. GraphRAG:a survey on retrieval augmented generation with knowledge graphs[J]. arXiv preprint arXiv:2402.04331,2024.

[48] Sanmartin Diego. KG-RAG:bridging the gap between knowledge and creativity[J]. arXiv preprint arXiv:2405.12035,2024.

[49] Jiang Ziyan,Ma Xueguang,Chen Wenhu,et al. LongRAG:enhancing retrieval-augmented generation with long-context LLMs[J]. arXiv preprint arXiv:2406.15319,2024.

[50] Abane Amar,Bekri Anis,Battou Abdella,et al. FastRAG:retrieval augmented generation for semi-structured data[J]. arXiv preprint arXiv:2411.13773,2024.

[51] Zhang Junyuan,Zhang Qintong,Wang Bin,et al. OHRBench:a benchmark for evaluating OCR's cascading impact on retrieval-augmented generation[J]. arXiv preprint arXiv:2407.10701,2024.

[52] Xu Yiheng,Li Minghao,Cui Lei,et al. LayoutLM:pre-training of text and layout for document image understanding[C]//26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,2020:1192-1201.

[53] Peer David,Schöpf Philemon,Nebendahl Volckmar,et al. ANLS*-a universal document processing metric for generative large language models[J]. arXiv preprint arXiv:2402.03848,2024.

[54] Guerrero Maximo. Use GPT4-Vision for PDF to JSON data extraction[EB/OL]. https://dev.to/maximoguerrero/use-gpt4-vision-for-pdf-to-json-data-extraction-311c,2024.

[55] PaddleOCR Team. PaddleOCR:an open-source OCR library for document automation[EB/OL]. https://github.com/paddle-ocr/paddle,2024.

[56] Jiang Huiqiang,Wu Qianhui,Luo Xufang,et al. LongLLMLingua:accelerating and enhancing LLMs in long context scenarios via prompt compression[J]. arXiv preprint arXiv:2305.14627,2023.

[57] Zhao Jun, Zu Can, Xu Hao, et al. LongAgent: scaling language models to 128k context through multi-agent collaboration[J]. arXiv preprint arXiv:2402.11550,2024.

[58] Li Yan, Han Soyeon Caren, Dai Yue, et al. ChuLo: chunk-level key information representation for long document processing[J]. arXiv preprint arXiv:2403.11119,2024.

[59] CHENG Y, ZHANG Z, YANG L, et al. A research on interactive application of index structure optimized retrieval enhanced generation technology in insurance field[J]. Journal of the Hebei Academy of Sciences, 2024, 42(1): 13-20.

[60] BI F, ZHANG Q, ZHANG J, et al. A retrieval-augmented generation system based on a sliding window strategy in large language models[J]. Journal of Computer Research and Development, 2024, 61(5): 1-10.

[61] Ram Ori, Levine Yoav, Dalmedigos Itay, et al. In-context retrieval-augmented language models[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 1-18.

[62] SHI D, ZENG J. Retrieval-augmented document-level multi-event extraction with fine-tuned large language models[J]. Journal of Chinese Computer Systems, 2024, 45(4): 1-8.

[63] MENG X, WANG H, LI Y, et al. Prompt learning based on retrieval-augmented generation of fine-grained knowledge graph[J]. Data Analysis and Knowledge Discovery, 2024, 8(5): 1-12.

[64] He Xiaoxin, Tian Yijun, Sun Yifei, et al. G-retriever: retrieval-augmented generation for textual graph understanding and question answering[J]. arXiv preprint arXiv:2402.07630,2024.

[65] LIN B, GAO J, LI H. Intelligent operation and maintenance method based on rag finetuned and enhanced LLM[C]//Frontiers of Intelligent Operation Technology, Chinese Institute of Command and Control, 2024: 89-96.

[66] Kim Sejong, Song Hyunseo, Seo Hyunwoo, et al. Optimizing retrieval strategies for financial question answering documents in retrieval-augmented generation systems[J]. arXiv preprint arXiv: 2503.15191,2025.

[67] Chen Jungang, Gildin Eduardo, Kompantsev Georgy, et al. Optimization of pressure management strategies for geological CO2 sequestration using surrogate model-based reinforcement learning[J]. International Journal of Greenhouse Gas Control, 2024, 138: 104262.

[68] Dai Zhuyun, Zhao Vincent Y, Ma Ji, et al. Promptagator: few-shot dense retrieval from 8 examples[J]. arXiv preprint arXiv: 2209.11755,2022.

[69] Zhang Peitian, Xiao Shitao, Liu Zheng, et al. Retrieve anything to augment large language models[J]. arXiv preprint arXiv: 2310.07554,2023.

[70] Alabdulmohsin Ibrahim, Zhai Xiaohua, Kolesnikov Alexander, et al. Getting ViT in shape: scaling laws for compute-optimal model design[C]//37th International Conference on Neural Information Processing Systems, 2023.

[71] Li Xinze, Liu Zhenghao, Xiong Chenyan, et al. Structure-aware language model pretraining improves dense retrieval on structured data[C]//Findings of the Association for Computational Linguistics, 2023: 11560-11574.

[72] Jégou Herve, Douze Matthijs, Schmid Cordelia, et al. Product quantization for nearest neighbor search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(1): 117-128.

[73] Shao Zhihong, Gong Yeyun, Shen Yelong, et al. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy[C]//Findings of the Conference on Empirical Methods in Natural Language Processing, 2023: 9248-9274.

[74] Gao Luyu, Ma Xueguang, Lin Jimmy, et al. Precise zero-shot dense retrieval without relevance labels[C]//61st Annual Meeting of the Association for Computational Linguistics, 2023: 1762-1777.

[75] Wang Liang, Yang Nan, Wei Furu, et al. Query2doc: query expansion with large language models[J]. arXiv preprint arXiv: 2303.07678,2023.

[76] Zhuang Shengyao, Liu Bing, Koopman Bevan, et al. Open-source large language models are strong zero-shot query likelihood models for document ranking[J]. arXiv preprint arXiv: 2310.13243,2023.

[77] ZHANG Y P, CHEN M F, TIAN C H, et al. Multi-strategy retrieval-augmented generation method for military domain knowledge question answering systems[J]. Journal of Computer Applications, 2025, 45(3): 746-754.

[79] Tang Yixuan, Yang Yi. MultiHop-RAG: benchmarking retrieval-augmented generation for multi-hop queries[J]. arXiv preprint arXiv: 2401.15391,2024.

[80] Jabal Mohamed Sobhi, Warman Pranav, Zhang Jikai, et al. Language models and retrieval augmented generation for automated structured data extraction from diagnostic reports[J]. arXiv preprint arXiv: 2409.10576,2024.

[81] Wu Jiayi, Cai Hengyi, Yan Lingyong, et al. PA-RAG: RAG alignment via multi-perspective preference optimization[C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2025: 9091-9112.

[82] Dathathri S, Madotto A, Lan J, et al. Plug and play language models: a simple approach to controlled text generation[C]//8th International Conference on Learning Representations, 2020: 3411-3444.

[83] Pascual Damian, Egressy Beni, Meister Clara, et al. A plug-and-play method for controlled text generation[C]//Findings of the Association for Computational Linguistics, 2021: 3973-3997.

[84] Huang James Y, Sengupta Sailik, Bonadiman Daniele, et al. DeAL: decoding-time alignment for large language models[J]. arXiv preprint arXiv: 2402.06147,2024.

[85] Yan Shiqi, Gu Jiachen, Zhu Yun, et al. Corrective retrieval augmented generation[J]. arXiv preprint arXiv: 2401.15884,2024.

[86] Cho Sukmin, Jeong Soyeong, Seo Jeongyeon, et al. Typos that broke the RAG's back: genetic attack on RAG pipeline by simulating documents in the wild via low-level perturbations[C]//Findings of the Conference on Empirical Methods in Natural Language Processing, 2024: 2826-2844.

[87] Xue Jiaqi, Zheng Mengxin, Hu Yebowen, et al. BadRAG: identifying vulnerabilities in retrieval augmented generation of large language models[J]. arXiv preprint arXiv: 2404.09279,2024.

[88] Fang Feiteng, Bai Yuelin, Ni Shiwen, et al. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training[C]//62nd Annual Meeting of the Association for Computational Linguistics, 2024: 10028-10039.

[89] Hu Chanwei, Wang Yueqi, Xing Shuo, et al. mRAG: elucidating the design space of multi-modal RAG[J]. arXiv: 2505.24073,2025.

[90] PENG J, YU F, ZHAO Y. From patient-generated health data to medication recommendations: a graph retrieval-augmented genera-

tion model for patient medication information Q&A[J]. Journal of Modern Information,2024,44(5):1-10.

[91] Hui Yulong, Lu Yao, Zhang Huanchen, et al. UDA: a benchmark suite for retrieval augmented generation in real-world document analysis[C]//38th International Conference on Neural Information Processing Systems,2024:67200-67217.

[92] Liu Nelson F, Lin Kevin, Hewitt John, et al. Lost in the middle: how language models use long contexts[J]. Transactions of the Association for Computational Linguistics,2024,12:157-173,doi:10.1162/tacla_0_0638.

[93] Saad Falcon Jon, Khattab Omar, Potts Christopher, et al. ARES: an automated evaluation framework for retrieval-augmented generation systems[C]//Conference of the North American Chapter of the Association for Computational Linguistics,2024:338-354.

[94] Es Shahul, James Jithin, Espinosa Anke Luis, et al. RAGAs: automated evaluation of retrieval augmented generation [C]//18th Conference of the European Chapter of the Association for Computational Linguistics,2024:150-158.

[95] Chen Jiawei, Lin Hongyu, Han Xianpei, et al. Benchmarking large language models in retrieval-augmented generation [C]//38th AAAI Conference on Artificial Intelligence,2024:17754-17762.

[96] GUO Q, CHEN Q. PGCA-RAG: a parallel graph caching architecture for large language model retrieval-augmented generation[J]. Journal of Chinese Computer Systems,2024,45(6):1-8.

附中文参考文献:

[14] 张添翼,张晨曦,彭　鑫,等. 基于大语言模型的软件配置研究综述[J]. 计算机应用与软件,2025,42(3):1-12.

[23] 何　哲,姜　博,王　晓. 改进检索增强与 LLM 思维链维修策略生成[J]. 计算机应用与软件 2024,42(3):1-6 +83.

[39] 王　慧,石　云. 基于大语言模型的查询扩展方法研究[J]. 计算机技术与发展,2024,35(3):148-155.

[40] 王　浩,魏　佳,景　浩,等. Meta-RAG:基于元数据驱动的电力领域检索增强生成框架[J]. 计算机工程,2024,50(2):1-11.

[59] 程　宇,张　震,杨　璐,等. 一种索引结构优化的检索增强生成技术在保险领域的交互应用研究[J]. 河北省科学院学报,2024,42(1):13-20.

[60] 毕　飞,张　琦,张　杰,等. 基于滑动窗口策略的大语言模型检索增强生成系统[J]. 计算机研究与发展,2024,61(5):1-10.

[62] 石　东,曾　杰. 利用微调大语言模型的检索增强文档级多事件抽取[J]. 小型微型计算机系统,2024,45(4):1-8.

[63] 孟　祥,王　海,李　阳,等. 基于细粒度知识图谱检索增强生成的提示学习研究[J]. 数据分析与知识发现,2024,8(5):1-12.

[65] 林　博,高　健,李　辉. 基于大模型 RAG 微调与增强的智能运维方法[C]//中国指挥与控制学会·智能运维技术前沿,2024:89-96.

[77] 张艳萍,陈梅芳,皿昌海,等. 面向军事领域知识问答系统的多策略检索增强生成方法[J]. 计算机应用,2025,45(3):746-754.

[90] 彭　杰,于　飞,赵　刚. 从患者生成健康数据到用药建议生成:基于图检索增强生成的患者用药信息问答模型[J]. 现代情报,2024,44(5):1-10.

[96] 郭　强,陈　琦. PGCA-RAG:面向大语言模型检索增强的并行图缓存架构[J]. 小型微型计算机系统,2024,45(6):1-8.