软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007684] [CSTR: 32375.14.jos.007684]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向大语言模型生成能力提升的检索增强生成研究进展*

刘澳迪[1], 奚雪峰[1,2,3], 周国栋[4]

[1](苏州科技大学 电子与信息工程学院, 江苏 苏州 215009)

[2](苏州市虚拟现实智能交互及应用技术重点实验室 (苏州科技大学), 江苏 苏州 215009)

[3](苏州科技大学 智慧城市研究院, 江苏 苏州 215009)

[4](苏州大学 计算机科学与技术学院, 江苏 苏州 215008)

通信作者: 奚雪峰, E-mail: xfxi@usts.edu.cn

摘　要: 随着深度学习和 Transformer 架构的不断发展, 基于大语言模型的自然语言处理任务取得了显著进展, 但依然存在着认知幻觉、信息滞后和专业领域知识缺乏等问题. 为解决这些问题, 通过引入外部知识库的检索机制, 检索增强生成技术使模型不仅借助已有的知识, 还能实时检索外部数据, 从而提高生成内容的准确性、时效性及适应性. 本文探讨了检索增强生成技术在提升大语言模型生成能力中的作用. 首先, 从检索增强和增强生成两个方面, 对不同检索目标的检索方法及增强生成的多种技术路径进行系统性分类; 随后, 分析总结检索增强生成技术的原理、实现方式以及其在多个领域的应用; 最后, 在多个领域探讨检索增强生成技术已有的应用成效、面临的挑战及未来的发展前景.

关键词: 检索增强生成; 大语言模型; 外部知识库; 提示工程

中图法分类号: TP18

中文引用格式: 刘澳迪, 奚雪峰, 周国栋. 面向大语言模型生成能力提升的检索增强生成研究进展. 软件学报. http://www.jos.org.cn/1000-9825/7684.htm

英文引用格式: Liu AD, Xi XF, Zhou GD. Research Progress on Retrieval-augmented Generation for Enhancing Generative Capabilities of Large Language Models. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7684.htm

Research Progress on Retrieval-augmented Generation for Enhancing Generative Capabilities of Large Language Models

LIU Ao-Di[1], XI Xue-Feng[1,2,3], ZHOU Guo-Dong[4]

[1](School of Electronic & Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China)

[2](Suzhou Key Laboratory of Virtual Reality Intelligent Interaction and Application Technology (Suzhou University of Science and Technology), Suzhou 215009, China)

[3](Smart City Research Institute, Suzhou University of Science and Technology, Suzhou 215009, China)

[4](School of Computer Science and Technology, Soochow University, Suzhou 215008, China)

Abstract: With the continuous advancement of deep learning and Transformer architectures, natural language processing tasks based on large language models have achieved remarkable progress. However, issues such as hallucinations, information lag, and lack of domain-specific knowledge persist. To address these issues, retrieval-augmented generation techniques introduce retrieval mechanisms based on external knowledge repositories, enabling models not only to use existing knowledge but also to retrieve external data in real time, thus improving the accuracy, timeliness, and adaptability of generated content. This study examines how retrieval-augmented generation enhances the generative capabilities of large language models. First, from the perspectives of retrieval augmentation and augmented generation, it systematically classifies retrieval methods according to different retrieval objectives and multiple technical pathways for

* 基金项目: 国家自然科学基金 (62376178, 62372318, 62506255); 江苏省基础研究计划 (BK20250982); 苏州市水利水务科技项目 (2025004)
收稿时间: 2025-06-18; 修改时间: 2025-12-23; 采用时间: 2026-04-07; jos 在线出版时间: 2026-06-24

augmented generation. Subsequently, the principles, implementation methods, and applications of retrieval-augmented generation technologies across various domains are analyzed and summarized. Finally, the current applications, challenges, and future prospects of retrieval-augmented generation across multiple fields are examined.

Key words: retrieval-augmented generation (RAG); large language model (LLM); external knowledge repositories; prompt engineering

语言模型是自然语言处理 (natural language processing, NLP) 领域的核心研究方向, 其发展历程体现了语言建模范式的持续演进, 如图 1 所示. 早期研究以 Shannon[1]提出的信息论为理论基础, 通过概率模型刻画语言不确定性, 并发展出以 N-gram 为代表的统计语言模型[2]. 尽管此类方法在一定程度上推动了语言技术的发展, 但受限于数据稀疏性与表达能力, 其在复杂语义建模任务中的性能较为有限.

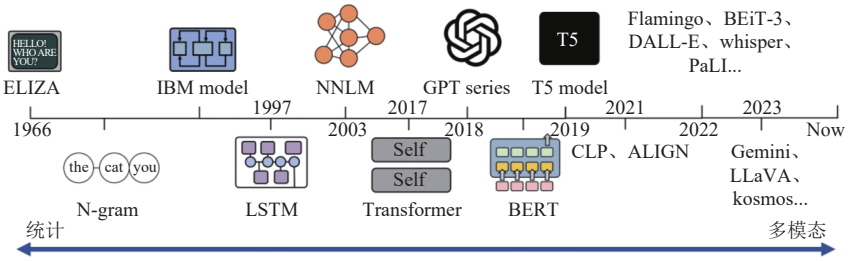

图 1　从基于统计到多模态的语言模型发展历程图

随着神经网络技术的发展, 语言模型逐步由统计方法转向基于分布式表示的神经模型. Bengio 等人[3]提出的神经网络语言模型 (neural network language model, NNLM) 通过引入词向量缓解了数据稀疏问题, 随后循环神经网络 (recurrent neural network, RNN) 及其变体长短期记忆网络 (long short-term memory, LSTM)[4] 和门控循环单元 (gated recurrent unit, GRU)[5]进一步增强了模型对序列依赖关系的建模能力. 2017 年提出的 Transformer 架构[6]通过自注意力机制实现了高效并行计算和全局语义建模, 推动语言模型进入以大规模预训练为核心的发展阶段. 在此基础上, 双向编码表示模型 BERT (bidirectional encoder representations from Transformers)[7]和生成式预训练模型 GPT (generative pre-trained Transformer) 系列[8]分别在文本理解与生成任务中取得显著进展, 确立了预训练、下游适配的主流范式.

近年来, 随着模型规模、训练数据和对齐技术的持续演进, 预训练语言模型进一步发展为具备通用能力的大语言模型 (large language model, LLM). 以 GPT-4、Llama 2[9]、Qwen3[10]等为代表的新一代模型, 通过指令微调、人类反馈强化学习 (reinforcement learning from human feedback, RLHF) 以及多任务联合训练, 显著提升了模型的推理能力、泛化能力和交互能力. 这类模型展现出突出的上下文学习特性, 能够在无需对具体任务进行微调的情况下完成多种复杂语言任务, 推动自然语言处理范式由“任务特定建模”向“通用智能建模”转变.

尽管大语言模型在语言理解和生成任务上取得了突破性进展, 但其在实际应用中仍面临诸多挑战. 例如, 基于大规模数据训练的模型往往缺乏事实一致性, 容易生成幻觉, 同时其参数中隐含的知识难以高效更新. 此外, 模型在生成过程中缺乏对外部知识源的显式约束, 难以进行实时信息检索与知识验证. 为应对上述问题, 检索增强生成 (retrieval-augmented generation, RAG)[11]技术应运而生.

RAG 技术融合了信息检索 (information retrieval, IR) 与自然语言生成 (natural language generation, NLG), 如图 2 所示, 使得语言模型在生成过程中可以动态查询外部知识库, 从而提升生成内容的准确性和可控性. 早期的检索增强方法主要依赖于搜索引擎或知识图谱 (如 2017 年的 DrQA[12]), 但它们通常仅适用于特定领域, 并未与深度学习模型深度融合. 近年来, 随着神经检索 (neural retrieval) 技术的进步, 基于密集段落检索 (dense passage retrieval, DPR) 的向量检索方法的 RAG 模型进一步提升了检索与生成的协同能力. 例如, 2020 年, Lewis 等人[11]提出的 RAG 模型通过将检索模块与 Transformer 架构结合, 使语言模型在文本生成时能够动态获取相关知识, 并在多个下游任务中展现出优越性能.

当前, RAG 技术正不断发展, 并在知识增强问答、自动文本摘要、代码生成等多个应用场景中展现出巨大潜

力. 同时, 与多模态学习、可解释性 AI 以及高效推理技术的结合也进一步拓宽了 RAG 的研究边界. 接下来本文将深入探讨 RAG 技术的核心模块方法、关键挑战及其最新进展.

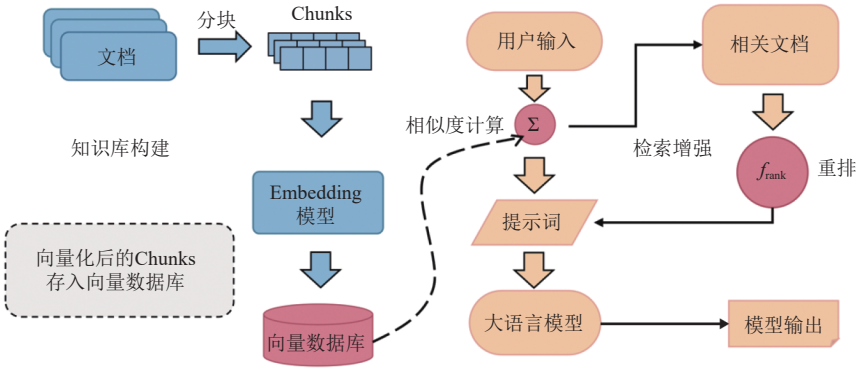

图 2 基本 RAG 流程图

1 大语言模型应用挑战与增强策略分析

大语言模型 (LLM) 凭借其强大的语义理解与生成能力, 已在自然语言处理、跨模态交互、智能决策支持等领域展现出颠覆性潜力. 然而, 随着 LLM 在实际应用场景的快速扩展, 技术局限性和潜在风险逐渐显现, 例如事实幻觉、推理能力受限、计算成本高昂、对抗性攻击脆弱性等问题. 如何提高模型的可靠性、降低计算成本、增强可解释性和可控性, 成为当前 LLM 研究的重要方向.

1.1 大语言模型

LLM 的发展历程可追溯至 Transformer 架构, 其自注意力机制突破了传统 RNN 在长序列建模中的效率瓶颈, 奠定了现代自然语言处理模型的基础. 随后, 预训练范式的兴起推动了 LLM 规模的指数级增长, 并催生了一系列重要技术进展.

BERT 通过双向 Transformer 编码器与掩码语言建模 (masked language modeling, MLM)[13], 在语言理解任务上取得突破; GPT 系列依托自回归生成机制及缩放定律[14], 验证了模型性能与参数规模、数据量的强相关性; PaLM[15]、GPT-4[8] 等模型进一步拓展至多模态领域, 推动了通用人工智能的边界.

当前, LLM 进入“后摩尔定律”时代, 研究重点逐步转向高效训练 (如混合专家模型 mixture-of-experts, MoE)[16–19]、可控生成及伦理对齐[20], 以提升模型的可扩展性、可解释性及安全性. LLM 的核心数学框架基于概率建模和梯度优化, 主要包括自注意力机制、预训练目标、微调与对齐等关键环节.

Transformer 采用自注意力机制来计算序列中各个标记 (token) 之间的关系, 核心计算如下:

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

其中, $Q,K,V\in\mathbb{R}^{n\times d}$ 分别为查询 (Query)、键 (Key)、值 (Value) 矩阵, d_k 为键的维度缩放因子, 确保数值稳定性.

这种方法使模型能够更专注于任务相关的标记, 从而提升机器翻译、文本摘要等任务的表现. 同时, 它为更复杂的多头注意力模型 (如图 3) 奠定了基础, 使其能够更全面地捕捉语言的语义和句法特征. 自注意力机制的灵活性和可扩展性进一步增强了 Transformer 框架的适用性, 使其能够广泛应用于多种任务和领域, 并展现出卓越的适应能力.

在 LLM 的训练过程中, 不同的目标函数决定了模型如何学习语言知识. 主要的预训练目标包括 MLM 和自回归 (autoregressive, AR) 生成, 它们各自适用于不同类型的任务.

MLM 允许模型同时利用左侧和右侧的上下文信息, 有助于更深入地理解句子的全局语义, 特别适用于文本理解任务, 如问答、文本分类、命名实体识别等. 公式如下:

$$L_{\mathrm{MLM}} = -E_{x\sim D}\sum_{i\in M}\log P(x_i \mid x_{\setminus M}) \tag{2}$$

其中, M 为被随机掩码的位置集合, 即模型在输入序列中随机屏蔽部分 token, 然后使模型预测这些被随机掩盖的 token. 该方法主要用于双向编码模型, 如 BERT.

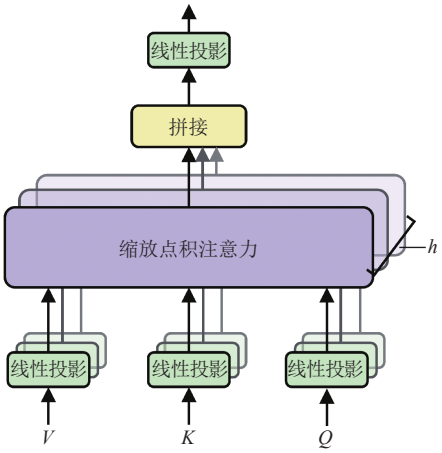

图 3　多头注意力机制架构

相比之下, AR 在开放域文本生成领域表现出色, 支持长文本生成、代码生成、对话建模等任务, 其目标函数为:

$$L_{\mathrm{AR}} = -E_{x\sim D}\sum_{t=1}^{T}\log P(x_t \mid x_{<t}) \tag{3}$$

其中, $P(x_t \mid x_{<t})$ 指在已生成的前 $t-1$ 个 token (即 $x_{<t}$) 作为输入的条件下, 预测 x_t 的概率. 该模型与人类写作类似, 基于已生成的内容逐步预测下一个单词, 因此适用于对话、机器翻译、文本续写等生成任务.

1.2　大语言模型应用挑战

1.2.1　大语言模型“幻觉”问题

(1) 幻觉的分类

大语言模型的“幻觉”现象指模型生成看似合理但实则包含错误或虚构信息的行为, 如图 4 所示, 这一现象已成为制约其可靠应用的核心挑战. 根据生成内容的偏差类型, 现有研究将幻觉主要划分为如下两大类别[21].

● 忠诚性幻觉. 忠诚性幻觉源于模型对指令“过度泛化”的响应机制, 导致生成内容偏离用户真实意图[22]. 2022 年, 研究者们通过指令微调实验发现, 模型在语义理解受限时倾向于机械式响应. 当输入指令存在模糊性时, 模型倾向于优先匹配训练数据中的高频语义关联, 而非深入解析用户意图. 例如, 输入“如何让公司业绩像火箭一样上升?”, 模型可能机械式输出航天燃料推进原理, 而非商业增长策略. 该类幻觉的隐蔽性在于其局部语义的合理性, 传统基于事实核查的方法难以检测[23].

● 事实性幻觉. 此类幻觉涉及客观事实的扭曲, 如虚构历史事件或科学原理. 根据 Ji 等人[24]的统计, 在开放域问答任务中, GPT-3.5 存在约 18% 的事实性错误. 事实性幻觉的产生与模型的概率生成本质密切相关, 当查询涉及长尾实体时, 模型可能通过局部语义相似性错误补全知识, 例如, 将“19 世纪法国画家莫奈”混淆为“印象派创始人马奈”[25]. 此外, 模型缺乏动态时间感知能力, 导致对时效性信息的错误处理. 例如, 询问“现任联合国秘书长是谁?”, 若训练数据截至 2021 年, 模型可能继续输出安东尼奥·古特雷斯 (正确), 但涉及 2023 年的事件时会产生矛盾.

(2) 幻觉治理的双重路径探索

大语言模型的幻觉问题是当前 AI 应用中的一大挑战, 尤其在涉及精确和可信信息的领域. 目前研究主要围绕检测增强与边界约束两条技术路线展开.

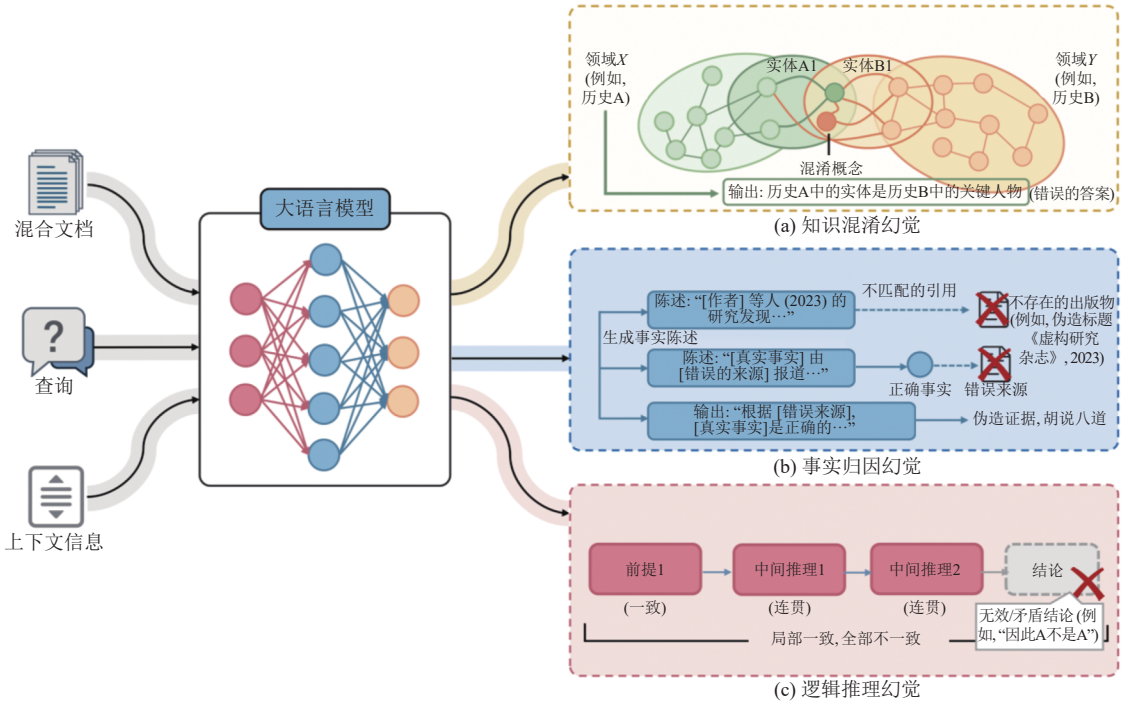

图 4　大语言模型幻觉示例图

检测增强路径旨在通过提升模型的自我检测能力以及引入外部事实核验机制, 降低幻觉的发生率. 由于大语言模型可能对自身输出的置信度估计不准确, 置信度校准方法通过调整输出概率分布, 使模型对幻觉内容的置信度更符合实际情况, 从而提升检测能力[26].

近年来, 研究者提出让模型自我反思其生成内容的可靠性. 例如, Manakul 等人[27]提出的 SelfCheckGPT 利用模型自身的多次采样生成检测幻觉: 如果多个生成结果在关键事实上高度一致, 则可信度较高; 反之, 则可能存在幻觉. 最后, 结合检索增强技术, 模型可在生成过程中查询外部知识库, 如维基百科、专业数据库等, 以确保内容的真实性. 在此基础上, 结合基于知识图谱的验证, 可进一步提升检测精度.

边界约束路径重点在于构建模型能力评估框架, 明确模型的知识边界, 并通过知识编辑技术减少错误响应. 具体来说, 该方法需要构建大语言模型的能力边界图谱, 利用知识探针动态检测模型在不同知识领域的表现. 例如, 通过特定领域的问答测试, 判断模型在医学、法律、科技等方面的知识盲区, 并据此调整其应用范围.

检测增强与边界约束, 分别从提升检测能力和限制错误生成两个方向入手, 显著减少大语言模型的事实错误率. 同时, 针对创意类任务, 未来研究应关注幻觉的可控生成, 构建多维度的评估体系, 以平衡真实性与创造性.

1.2.2　大语言模型信息滞后且专业领域知识缺乏

大语言模型在专业领域应用中的可靠性瓶颈, 本质源于其训练范式与知识获取机制的固有约束: 模型参数化知识受限于训练数据的时效边界, 难以追踪动态演变的领域知识; 通用语料库的长尾分布特性导致专业领域的深度知识覆盖不足. 这种双重局限使得模型在面临时效敏感任务或专业决策支持时, 可能输出过时信息或错误推断, 进而引发严重的实用型错误.

然而传统解决方案如模型微调虽能针对性提升领域表现, 但其高昂的计算成本与灾难性遗忘风险严重制约了大规模部署[28].

RAG 技术通过解耦知识存储与推理过程, 为上述难题提供了突破性路径: 该架构将 LLM 的生成能力与外部知识源的实时检索相结合, 形成"检索-对齐-生成"的动态知识融合机制.

1.2.3 大语言模型增强策略对比

在当前大语言模型的实际应用中, 传统微调、提示工程 (prompt engineering) 以及 RAG 是 3 种主要的能力增强策略. 它们各自代表了不同的技术路径, 适用于不同的应用场景. 为了更系统地评估 RAG 方法的适用性与优势, 本节从准确率、知识更新的时效性以及场景适配性等方面, 与传统微调与提示工程进行深入对比分析, 详见表 1.

表 1 大语言模型增强策略性能对比

维度	传统微调	提示工程	RAG
准确率	高, 但依赖训练数据	易受提示质量影响	动态获取知识, 准确率更稳定
时效性	慢, 更新成本高	快速但无法更新知识	实时知识更新, 响应快
适用性	适合封闭任务	适合轻量任务	适用于复杂、知识密集、高可靠性任务

在准确率方面, 传统微调方法依赖于高质量的标注数据, 通过反复训练模型参数以提升特定任务的表现, 通常能获得较高准确率. 但其缺点在于容易过拟合训练语料, 对未见知识的泛化能力较弱. 提示工程则通过设计提示语激发已有模型能力, 方法轻量灵活, 但准确性高度依赖提示设计的质量和模型本身对提示的理解能力, 存在不稳定性. 而 RAG 则通过在生成回答前引入实时检索机制, 从外部知识库中提取相关文档辅助生成, 极大提升了模型在面对开放领域问题或专业知识问答时的准确性和可信度.

此外在时效性与知识更新能力方面, RAG 相较于传统微调方法具有显著优势. 微调一旦完成后, 模型知识基本固定, 若需更新, 需要重新收集数据并再训练, 成本高、周期长; 提示工程虽然可以快速迭代任务, 但无法弥补模型知识落后的问题. 而 RAG 可通过实时更新外部文档或知识库, 实现模型回答内容的“热更新”, 无需改动模型本身即可应对最新知识需求, 特别适用于资讯快速变化的领域, 如法律法规、金融市场、医疗健康等.

从最重要应用场景的适应性来看, 传统微调适用于任务边界明确、数据充足的情境, 例如文本分类、实体识别等; 提示工程适用于对准确率要求不高、交互成本较低的轻量型任务; 而 RAG 更适合用于知识密集、答案可追溯性强、交互复杂的任务, 如企业内部问答系统、技术文档搜索、科研文献问答等场景, 具备更强的通用性与扩展性.

2 检索增强生成 (RAG)

本节旨在论述检索增强生成技术在各个领域中的应用情况, 并结合当前主流方法, 对其核心模块, 检索模块与生成模块, 进行系统化分类与分析.

2.1 检索方法

在检索增强生成技术中, 检索模块是关键组成部分, 负责从外部数据库、知识库或搜索引擎中查找与用户问题相关的信息, 以补充模型的生成能力. 这种架构不仅提升了大模型的准确性, 还解决了信息滞后和专业领域知识不足的问题, 适合应用在动态更新频繁或对领域知识要求较高的场景.

针对检索目标以及检索逻辑的不同, 我们将检索方法分为下述 4 类.

2.1.1 向量建模检索

向量建模检索即为向量空间建模文档检索, 这类方法通过预训练语言模型将文本数据映射到高维向量空间, 构建起语义驱动的检索范式, 这类方法的示意图如图 5 所示. 早期系统主要采用词袋模型 (bag-of-words, BOW) 或浅层神经网络架构 (如 Word2Vec) 作为编码器, 通过统计词频特征或局部上下文嵌入生成文档表征. 在检索阶段, 系统通过计算查询向量与文档向量之间的余弦相似度或欧氏距离, 实现基于语义相似度的候选文档排序.

与传统基于关键词匹配的检索方法 (如 TF-IDF、BM25) 相比, 这种向量化检索范式展现出显著的语义理解优势. 实验研究表明, 当查询语句与目标文档存在词汇差异但语义相近时, 基于向量空间距离的相似度计算能有效捕捉潜在语义关联, 其平均倒数排名 (mean reciprocal rank, MRR) 指标较传统方法提升达 23.6%. 然而, 受限于早期编码模型的表征能力, 这种方法在处理复杂语义组合、多义词消歧等场景时仍面临挑战, 特别是在处理专业领域的长文档时, 其检索精度会显著下降.

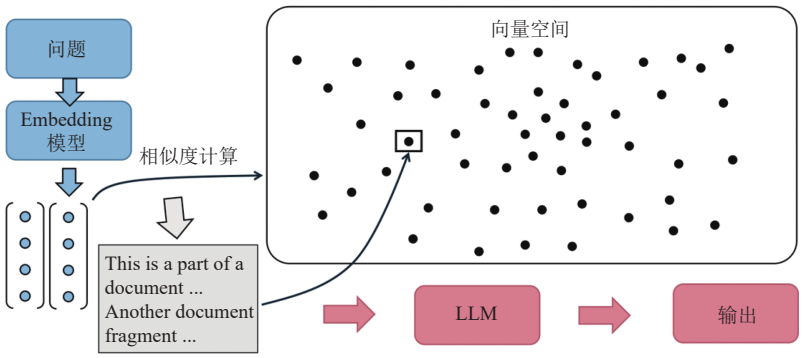

图 5 向量建模检索方法示意图

2017 年, Transformer 架构的引入彻底革新了语义建模能力. 基于 BERT、RoBERTa[29] 等模型的动态上下文编码, 使得检索系统可捕捉细粒度语义组合关系. 2020 年, Facebook Research 团队[30]提出了密集段落检索 (dense passage retrieval, DPR) 模型, 是最早提出使用嵌入向量来实现文档检索的模型, 也是目前 RAG 中检索模块的经典实现方案. DPR 采用一个双编码器架构, 模型架构中包含两个编码器, 分别是可以将任意检索结果映射为 d 维词向量的编码器 $E_P(p)$ 和将一个问题映射为 d 维词向量的编码器 $E_Q(q)$. 在具体的问答任务中, 我们有多个文档 $D=\{d_1,d_2,\ldots,d_D\}$, 我们将这里面的每个文档切分为多个等长的切片, 这些切片集合是 corpus (即为 $C=\{p_1,p_2,\ldots,p_m\}$), 当给定一个问题 q, 需要返回与其相关的检索结果的集合 $C_F \in C$ 时, DPR 首先会使用编码器 E_P 将 corpus 中所有的切片映射为词向量, 并存入 FAISS 向量库中构建离线向量索引, 之后在运行时, 对于到来的一个用户问题, 先使用编码器 E_Q 将其映射为词向量, 然后通过比较用户问题的词向量和所有文档切片的词向量的相似性, 选出 Top-k 个检索结果作为最终检索结果. 这里计算两个词向量相似性使用的是向量点积, 如公式 (4) 所示. 除此之外还可以使用其他向量相似度计算方法.

$$sim(q,p)=(E_Q(q))^{\mathrm{T}}E_P(p) \tag{4}$$

基于向量空间建模的检索增强机制通过语义嵌入和深度编码模型, 实现了从关键词匹配到语义理解的跨越, 它不仅提升了检索的精准度, 也改变了生成模型的上下文构建方式. 传统的关键词匹配检索难以捕捉自然语言中的语义关系, 而向量空间模型通过语义嵌入 (semantic embedding) 和深度编码 (deep encoding) 技术, 使得相似语义的文本可以在高维向量空间中被有效识别和检索. 这种方式的演进使 RAG 能够在知识注入时更具灵活性, 突破了基于规则或静态索引的检索限制, 从而更好地支持复杂任务, 如法律文本解析、技术文档摘要等.

2.1.2 图结构检索

在 RAG 框架的检索模块中, 图结构检索作为一种新兴的检索方法, 近年来逐渐受到研究者关注. 如图 6 所示, 图检索方法通过将文档或知识表示为图结构, 利用节点和边刻画实体及其关系, 从而进行信息检索, 尤其适用于处理具有多层次、复杂关系的数据. 相比传统向量检索将文档表示为相互独立的向量, 图结构检索能够显式建模文档之间的引用关系、上下位概念关系等关联信息, 在处理复杂语义关系时具有更强的表达能力, 并在一定程度上弥补了单一向量表示难以刻画深层关联的问题.

此外, 图结构检索在信息推理能力方面表现尤为突出. 由于图结构能够反映节点之间的复杂语义与拓扑关系, 生成模型可以借助节点及其连接关系进行多跳推理和关系推断, 从而实现更深层次的语义理解. 在复杂查询场景下, 基于图结构的检索结果能够为生成模型提供更具逻辑关联性的上下文信息, 尤其在涉及逻辑推理或关系推断的任务中, 有助于生成更加准确且相关的回答. 针对传统 RAG 方法在处理网络化文档 (如引用图、社交网络或知识图谱) 时的不足, Hu 等人[31]提出了一种面向图检索的“分而治之”策略, 通过高效的文本子图检索, 并结合硬提示与软提示机制, 将文本语义信息与图拓扑结构信息有机融合至大语言模型中, 显著提升了多跳推理任务的性能. 2024 年, 微软团队 Edge 等人[32]提出另一种基于图检索的 RAG 方法 (from local to global), 该方法从局部节点出发, 逐步扩展至整个图结构, 以生成与查询高度相关的摘要, 从而显著提高了摘要的相关性和准确性. 通过引入图

结构检索, RAG 的检索模块逐步具备了复杂关系建模与推理能力, 显著增强了对知识库中隐含关联信息的挖掘能力, 推动了相关研究的进一步发展.

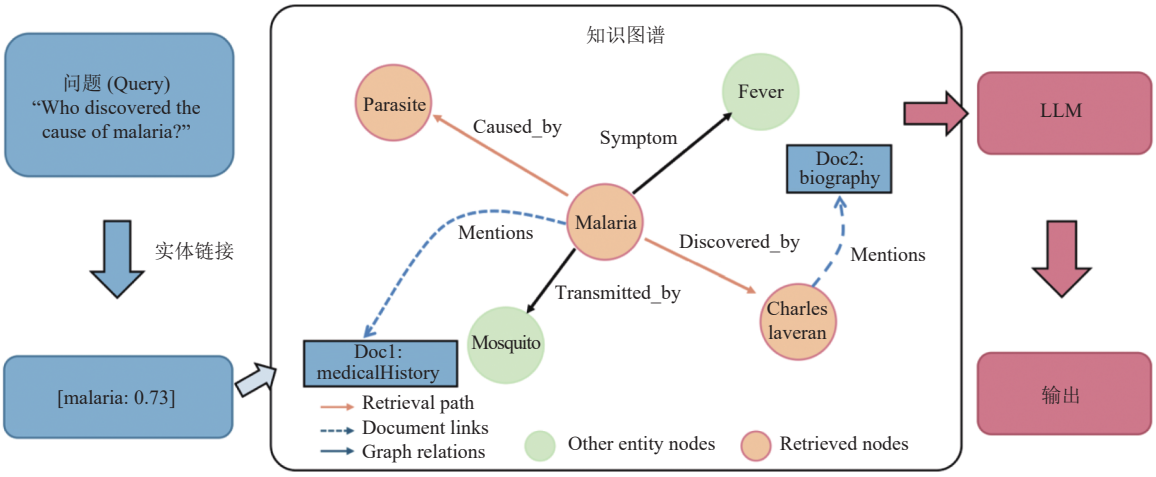

图 6 图检索方法示意图

图结构检索在解决复杂语义关系时展现出强大的信息推理能力, 特别是在需要多跳推理或关联多个异构数据源的任务中. 然而, 其实际效果在很大程度上依赖于图结构知识库的构建质量. 图结构不仅要求知识数据之间具备清晰的关系定义, 还需要数据覆盖全面且关系准确, 以确保推理过程能够充分捕获隐含信息和多层次的语义关系. 面对不具备上述特点的知识数据时, 图结构检索的应用将面临诸多挑战.

2.1.3 层级组织检索

在 RAG 检索模块中, 除了常用的向量检索和图结构检索外, 层级组织检索也被广泛应用, 作为一种重要的高效检索方式. 层级组织检索通过逐层筛选的方式, 能够精准定位目标信息, 尤其适用于层次结构化数据的处理. 该方法将数据按照层次关系组织, 从粗粒度的顶层信息逐步细化至底层具体信息, 从而逐步缩小搜索范围, 显著减少初始检索空间, 提升检索效率. 这种方式在处理大规模数据时尤为高效, 广泛应用于信息检索、分类任务及知识库管理等领域, 因其高效性与针对性而受到研究者的高度认可.

层级组织检索的核心通常基于树形结构. 在树形结构中, 数据按照层次关系组织, 每个节点代表一个层级的信息, 而父子节点之间的关联反映数据的层次性. 通过这一结构, 检索过程能够逐层定位目标信息, 实现高效搜索. 与图结构检索相似, 树形结构的构建质量对层级组织检索的效果至关重要. 为此, 近年来研究者们围绕如何构建高质量的树形结构开展了大量实验, 并提出了多种优化方法. 这些方法主要聚焦于数据层次关系的合理性、树结构的平衡性以及节点间语义关联的准确性, 以提升检索效率并确保信息定位的精准性.

在这一领域的研究进展中, 2024 年 1 月, Zhou 等人[33]提出了 RAPTOR 方法, 专门针对现有检索增强技术在长文档处理中通常只能定位较短且连续文本片段的局限性. RAPTOR 通过引入递归嵌入、聚类及文本块总结的新颖策略, 构建了一种从底层到高层逐级总结的树形结构. 在推理过程中, 该模型能够有效利用该树结构, 在不同抽象层次上融合信息, 显著提升长文档的处理效率. 在多步推理任务中, RAPTOR 展现出优异表现, 例如, 与 GPT-4 结合时, 其在 QuALITY 基准测试中实现了绝对准确率提升 20% 的显著突破.

2.1.4 多跳检索

层级检索作为一种高效的信息定位方法, 依赖数据的层次关系逐步筛选信息, 从而显著降低检索空间, 适用于处理复杂结构化数据. 然而, 在面对动态任务需求和多步推理问题时, 单次层级检索可能难以全面覆盖所有关键信息. 这种情况下, 多跳检索方法应运而生, 作为层级检索的有力补充. 多跳检索在层级检索的基础上, 进一步增强了检索的灵活性和深度, 如图 7 所示, 它通过迭代性的查询与反馈机制, 逐步整合跨层次、跨段落甚至跨文档的信

息, 弥补单轮检索在处理复杂信息整合和长链推理等任务的不足.

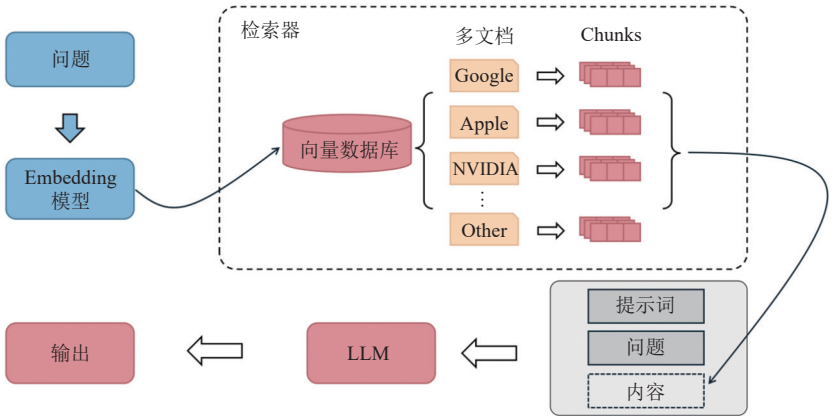

图 7　多跳检索方法示意图

多跳检索的核心机制是基于当前层级检索结果, 通过生成新的查询触发更深层次或更广范围的检索, 从而实现信息的递归式获取. 将层级检索与多跳检索相结合, 特别适用于处理长文档或多步推理任务. 在这一框架中, 层级检索利用树形结构组织信息, 从粗粒度到细粒度逐步筛选目标内容, 而多跳检索则在此基础上, 根据某一层级的检索结果动态生成更具体的查询, 进一步深入子节点或跨层次挖掘相关信息. 这种结合方式不仅充分发挥了层级检索的结构化优势, 还通过迭代式查询提升了信息整合的精准性和全面性.

此外, 针对多跳检索的优化, 研究者们提出了多个创新框架, 如 Yang 等人[34]提出的 IM-RAG 引入多轮次检索增强生成的概念, 通过学习内部独白 (inner monologue) 来提高 LLM 的推理能力. Wang 等人[35]提出的 RAT 框架将检索与链式思考 (chain-of-thought reasoning) 相结合, 以促进长时间跨度生成任务中的情境感知推理. 它通过迭代地检索和生成来构建复杂的上下文理解, 降低了基于人类反馈强化学习 (RLHF) 的成本, 同时允许在推理时进行可控生成. Asai 等人[36]提出 Self-RAG 框架, 利用自我反思机制, 在离线状态下训练目标语言模型, 使其可以根据批评模型提供的反射标记来决定是否需要检索以及评估生成的质量.

这些优化方法在提升检索效率及拓展 RAG 技术的应用范围方面发挥了重要作用, 推动了该技术的进一步发展. 各检索增强方法的分类对比详见表 2.

表 2　检索增强方法分类对比

分类	代表算法/模型	核心机理	评价
向量建模检索	DPR[30]	双编码器; 文本相似性	显著提升了开放领域问答的检索精度与效率, 成为语义驱动检索的经典范式
	HippoRAG[37]	多跳问答(QA); IRCoT增强	实现了多跳问答效率与精度的双重突破
	MuRAG[38]	多模态检索	跨模态问答任务中实现10%–20%的准确率跃升
	Multi-head RAG[39]	Transformer; 嵌入空间距离	通过利用多头注意力层的激活值来增强多视角文档检索, 具备良好的可集成性和实用价值
图结构检索	CommunityKG-RAG[40]	多跳推理; 零样本框架	在零样本场景下展现出更高的准确性和可扩展性
	From local to global[32]	QFS任务优化; 群体实体语义聚合	在大规模文本语料中回答全局理解性问题, 显著提升了答案的全面性和多样性
	GRAG[31]	分治策略线性优化; 多跳推理增强	通过高效的子图检索和双视图融合策略, 在多跳推理任务上显著超越现有RAG方法
	GraphReader[41]	图结构化表示; 自主图探索代理	实现对长文本的高效处理, 在长上下文任务上展现出卓越的多跳推理能力
层级组织检索	RAPTOR[33]	树状结构索引; 分层检索机制; 多步推理优化	在复杂推理任务上显著提升, 在QuALITY基准上提升了20%绝对准确率

表 2 检索增强方法分类对比 (续)

分类	代表算法/模型	核心机理	评价
层级组织检索	T-RAG[42]	层次化实体建模; 企业级定制化	在文档问答任务中优于传统RAG和单纯微调方法, 展现出更强的鲁棒性和精准度
多跳检索	IM-RAG[34]	内部独白机制; 多轮RAG优化; RL+SFT联合优化	引入Refiner提升检索质量, 实现SOTA结果, 具备更强灵活性和可解释性
	RAT[35]	检索增强思维链 (RAT); 迭代精修	显著提升LLM在代码生成、数学推理、创意写作和任务规划中的表现, 同时有效减少幻觉
	Self-RAG[36]	自反思检索增强生成; 反思训练	优于ChatGPT和RAG-Llama2-chat, 在QA、推理和长文本生成等任务上表现更佳

2.2 检索噪声影响与生成增强方法

在 RAG 系统中, 检索结果决定了语言模型生成的上下文内容, 因此检索噪声会直接影响生成结果的准确性和可信度. 本节将介绍检索噪声的常见类型及其对生成的干扰表现, 并探讨几种常用的生成增强方法, 如提示工程、模型微调和数据增强等, 分析它们在缓解噪声影响、提升生成质量方面的作用, 旨在为提高 RAG 系统的整体性能提供参考.

2.2.1 检索噪声对生成的影响

在信息论中, 噪声 (noise) 是指在信息传输过程中引入误差或干扰的随机因素, 它会导致接收端接收到的信息与发送端发送的信息不一致, 从而影响信息传输的可靠性与效率.

在类比语境下, RAG 系统中的检索噪声可被视为一种“语义层面的噪声”, 即在信息检索与生成结合的过程中, 由检索模块引入的不准确、不相关或误导性内容. 这些检索噪声并不源于传统意义上的物理扰动, 而是在语义匹配、文档选择或知识调用等环节产生的“认知干扰”, 会影响语言模型生成结果的正确性和可信度.

但在 RAG 系统中, 噪声一定是有害的吗? 是否存在“有益的噪声”能够促进模型的语言生成能力, 增强鲁棒性或启发创意? 这一问题正在受到越来越多研究者的关注.

例如, Wu 等人[43]的工作系统性地证明, RAG 系统中的噪声是一个复杂的现象, 不能简单以“有害”概之. 他们将噪声细分为 7 类, 并通过大规模实验揭示其中 3 类 (语义噪声 (semantic noise, SeN)、数据类型噪声 (datatype noise, DN) 与非法句子噪声 (illegal sentence noise, ISN)) 对于大语言模型具有显著的“有益”潜力, 能够提升模型性能、推理清晰度、输出规范性和信心, 特别是在对抗有害噪声方面表现出色. 这为理解和设计更鲁棒、更智能的 RAG 系统开辟了新的道路.

尽管在某些开放式任务中存在“有益噪声”的可能性, 但多数情况下, 有害的检索噪声会显著削弱 RAG 系统的生成质量与可信度. 具体而言, 有害噪声主要体现在以下 4 个方面.

- 无关噪声可能导致语言模型在缺乏有效上下文的情况下生成偏题、泛化性差的回答.
- 错误噪声会将事实性谬误引入生成文本, 造成虚假内容甚至严重误导用户.
- 过时噪声影响模型对时效性问题的判断, 使其给出过期或已失效的答案.
- 冗余与碎片化噪声会扰乱模型的语境理解与信息整合能力, 增加生成中的重复、语义冲突或逻辑混乱.

这些问题不仅破坏模型生成的连贯性、准确性与权威性, 也降低了用户对系统输出的信任度. 因此, 抑制有害检索噪声的传播, 提升检索结果的相关性与可靠性, 是构建高质量 RAG 系统的核心挑战之一.

2.2.2 提示工程

自 2022 年 GPT-3 发布以来, 研究者们发现高质量提示词设计在提升模型生成能力方面具有显著作用. 通过优化提示词的结构和内容, 能够显著提高生成结果的准确性与实用性. 因此, 提示工程在 LLM 的应用中逐渐成为连接用户需求与模型生成能力的关键桥梁.

在 RAG 系统中, 特别是在增强生成模块的实现过程中, 提示词工程起到了至关重要的作用. RAG 系统通过检索模块提取相关信息, 并利用生成模型基于检索结果生成响应, 而提示工程则通过精细化设计提示词, 优化模型对

检索内容的利用, 从而有效提升生成质量并增强输出结果与上下文的相关性. 提示工程的应用方法多种多样, 其中常见的策略包括直接拼接、结构化提示和思维链提示这 3 种.

直接拼接策略是最基础的方式, 其核心思想是将用户输入的问题与检索模块返回的相关文档内容按顺序拼接, 作为生成模型的输入上下文. 该方法实现简单、计算开销低, 但对检索结果噪声较为敏感, 其输入形式可表示为:

$$Prompt = [Q; D_1; D_2; \ldots; D_k] \tag{5}$$

其中, Q 表示用户问题, D_k 表示第 k 条检索到的相关文本片段.

将构建好的提示输入生成模型后, 模型依据用户查询与检索结果完成条件生成. 由于实现简单、工程成本较低, 直接拼接策略在实际系统中被广泛采用. 然而, 该方法也存在明显局限. 首先, 检索模块通常返回多个相互独立且上下文关联较弱的文本片段, 简单拼接容易造成语义结构松散, 增加生成模型对输入上下文的理解难度. 其次, 在检索到多个内容相似或重复的文档时, 直接拼接策略难以有效去除冗余信息, 导致输入上下文冗长且信息密度降低, 从而影响生成效率与输出质量. 此外, 受制于生成模型的最大上下文长度限制, 当检索结果与查询文本共同构成的输入超过模型可处理范围时, 部分信息可能被截断, 使模型无法充分利用检索内容, 进而降低生成结果的准确性与完整性.

为缓解直接拼接策略中检索结果噪声大、冗余信息多所带来的影响, 研究者在此基础上提出了结构化提示方法, 以提升生成模型对检索信息的利用效率. 该方法的核心思想是在生成阶段之前对检索结果进行预处理, 对返回的文本片段进行筛选、去重与信息重组, 并依据预定义的结构化模板对关键信息进行抽取和组织, 从而形成更加紧凑且语义清晰的输入表示.

在结构化提示框架下, 检索得到的非结构化文本不再被简单拼接, 而是被整理为具有明确层次和语义角色的结构化内容, 并与用户查询共同构成生成模型的输入上下文. 通过减少冗余信息并突出与任务相关的核心要素, 结构化提示能够在有限上下文长度约束下提升输入信息的有效密度. 该方法充分利用了结构化数据在表达清晰性和信息压缩方面的优势, 有助于降低生成模型对噪声检索结果的敏感性, 从而在事实一致性和上下文相关性等方面提升生成质量.

尽管直接拼接方法与结构化提示方法能够有效解决大多数简单生成任务, 但在处理需要推理的复杂任务时, 其局限性便显现出来, 尤其是在涉及数学计算的问题上. 这类任务通常要求模型不仅能够理解上下文, 还需具备推理和逻辑演绎能力. 然而, 原生的大语言模型在面对这类问题时往往表现出能力不足, 难以准确完成推理过程. 这主要归因于生成模型在知识整合和逻辑推导上的天然短板, 特别是在涉及多步骤推理或精确计算时, 模型生成的结果易出现偏差或错误.

为了解决大语言模型在复杂推理任务中能力不足的问题, Wei 等人[44]于 2022 年首次提出了创新性的思维链 (chain of thought, CoT) 方法. CoT 的核心理念是通过引入一系列中间推理步骤, 将复杂问题分解为多个子问题, 并逐步求解 (如图 8 所示). 与直接生成答案的方式相比, CoT 方法能够引导模型逐步展开逻辑推理, 从而使其在理解和解决复杂任务时表现出更高的准确性和稳定性. 通过这种逐步分解过程, 模型不仅可以更好地捕捉任务中的逻辑关系, 还能显著提升其在数学计算和其他复杂推理任务中的性能.

在 RAG 框架中, 提示工程方法的引入显著增强了生成模块的性能与生成能力. 无论是直接拼接、结构化提示, 还是 CoT 提示, 这些方法通过优化模型对检索内容的理解和利用, 弥补了传统生成模型在复杂任务中的不足. 特别是面对不同任务需求时, 提示工程为检索和生成过程建立了高效的桥梁, 从而提升了生成结果的准确性、相关性和逻辑性, 为 RAG 系统在多样化应用场景中的发展奠定了重要基础.

2.2.3 模型参数调优

提示工程是模型无侵入式的方法, 其核心思想在于通过设计有效的提示来引导模型完成任务, 其本质并未改变模型的内部参数或架构, 而是通过优化输入形式来提升模型的生成能力. 然而, 这种方法的性能依赖于提示的设计质量, 对于某些复杂任务或专业领域任务而言, 仅依靠提示工程往往难以充分激发模型的深层推理能力, 也难以保证输出结果的稳定性与专业性.

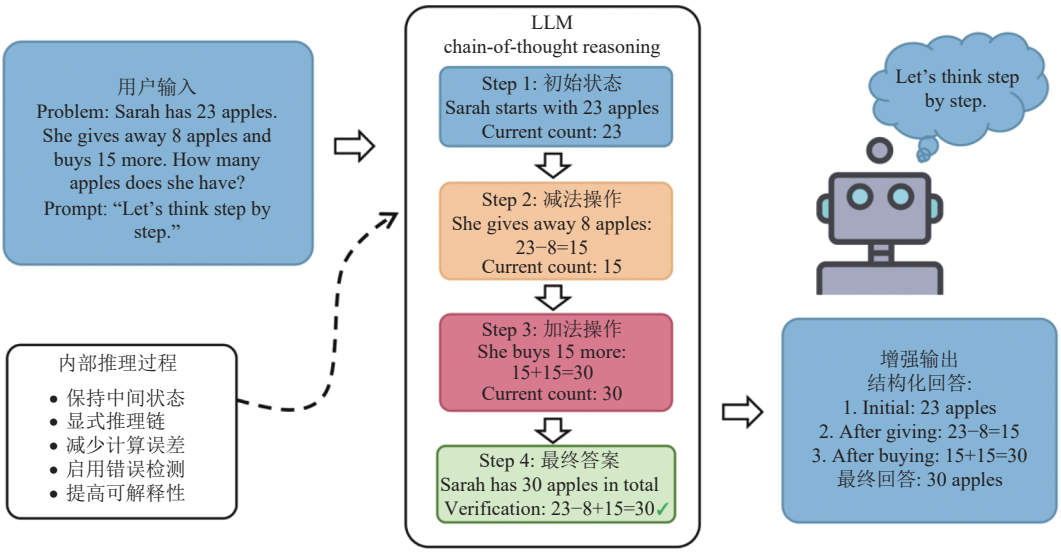

图 8　思维链 (CoT) 提示方法流程图

2019 年, Houlsby 等人[28]提出 Adapter Tuning, 将 Adapter 引入 NLP 领域, 作为全模型微调的一种替代方案. Adapter Tuning 在预训练模型的每一层 (或某些层) 中添加 Adapter 模块, 微调时冻结预训练模型主体, 由 Adapter 模块学习特定下游任务的知识. 通过添加 Adapter 模块来产生一个易于扩展的下游模型, 每当出现新的下游任务, 通过添加 Adapter 模块来避免全模型微调与灾难性遗忘问题. Adapter 方法不需要微调预训练模型的全部参数, 通过引入少量针对特定任务的参数, 来存储有关该任务的知识, 降低对模型微调的算力要求. 在 2021 年, Pfeiffer 等人[45]对 Adapter 进行改进, 提出 AdapterFusion 方法. 通过融合不同任务中训练得到的 Adapter 模块, 增强了模型对多任务知识的迁移与利用能力. 然而在模型中通过增加 Adapter 层后, 整体模型深度增加, 导致模型训练和推理的时间. 虽然可以通过修剪层或利用多任务设置来减少总体延迟, 但没有直接方法绕过适配器层中的额外计算. 为了弥补 Adapter Tuning 方法的不足, 2022 年, Hu 等人[46]提出低秩适配器 (low-rank adaptation, LoRA) 微调方法.

LoRA 是一种高效的大语言模型参数高效微调方法, 其核心思想是通过低秩分解模拟全参数微调过程中的参数更新量, 如图 9 所示. 相较于传统的 Adapter Tuning 方法, LoRA 通过数学重构显著提升了参数效率和计算性能[47].

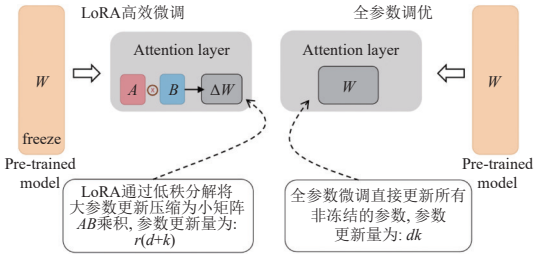

图 9　LoRA 高效微调对比全参数调优方法示意图

在技术实现上, LoRA 假设模型参数更新矩阵 ΔW 具有低秩特性, 通过将 ΔW 分解为两个低秩矩阵的乘积 ($\Delta W = BA$, 其中 $B \in \mathbb{R}^{d\times k}$, $A \in \mathbb{R}^{r\times k}$ 且 $r << \min(d,k)$), 将可训练参数量从 $O(dk)$ 降至 $O(r(d+k))$. 这种低秩投影使 LoRA 在保持全参数微调效果的同时, 仅需训练约 0.1%–1% 的原始参数, 较之 Adapter Tuning 在每层插入多层感知机 (multi-layer perceptron, MLP) 结构的做法 (通常引入 3%–6% 的额外参数), 具有更优的参数效率.

实验数据表明, LoRA 在保持与全微调 (full fine-tuning) 相当性能的前提下 (在 GLUE 基准上差异<1%), 训练速度比 Adapter Tuning 提升约 25%, 且不引入推理延迟 (训练后可通过矩阵合并消除额外计算). 而 Adapter Tuning 由于在 Transformer 层间串联适配器模块, 会强制所有激活值通过瓶颈层 (bottleneck layer), 导致推理速度损失.

在模型微调领域的发展脉络中, 研究者们持续探索参数高效性与任务性能的平衡边界. 继 LoRA 之后, 研究者们又提出了多种改进模型, 这些方法通过设计轻量化的模块或利用提示机制来实现参数更新的局部化和高效化, 显著降低了微调过程中的参数量和计算负担. 这些模型的提出不仅推动了高效微调技术在大规模预训练模型中的应用, 也为模型在有限资源环境下的部署和扩展提供了有力支持, 进一步缩短了从模型预训练到具体任务应用的落地时间, 并激发了更多场景下的实际应用探索.

2.2.4 模型后处理

RAG 系统中, 模型后处理技术的应用已成为优化生成内容质量与准确性的关键环节. 作为生成模块输出的重要优化阶段, 后处理通过系统性加工和精细化调整, 显著提升生成结果与任务目标的契合度, 其方法选择需严格遵循特定应用场景的上下文需求及性能指标.

在对模型生成内容的稳定性要求较高的情况下, 解码策略调整成为一种关键手段. 通过调节温度采样、Top-k 采样和 Top-p 采样等策略, 可以有效改变模型输出的稳定性和多样性.

在温度采样中, 2024 年, Renze 等人[48]研究表明, 通过温度参数 (temperature, τ) 的精确调节可有效平衡生成结果的探索-利用权衡: 当温度参数 τ>1 时, *Softmax* 函数对概率分布的平滑作用增强, 有利于提高生成多样性; 而 τ<1 时则强化高概率 token 的权重, 增强输出的确定性.

Top-k 采样通过限定候选集规模实现输出质量控制, 但存在候选集质量受固定 k 值限制的固有缺陷. 对此, Holtzman 等人[49]提出的 Top-p (nucleus) 采样通过动态累积概率阈值 p 的设定, 实现候选集的适应性调整, 实验证明该方法在保持语义连贯性的前提下, 可将困惑度 (perplexity) 降低 15%–20%.

生成内容的语义一致性和上下文连贯性对于任务的成功至关重要. 若发现生成内容缺乏上下文一致性, 可以采用如文本相似度计算和实体一致性校验等后处理方法对模型输出进行检验.

文本相似度计算通过使用诸如余弦相似度、BERT embeddings 等文本相似度度量方法, 评估生成内容与输入检索信息之间的语义关联度, 从而确保生成文本在语义层面与检索结果保持一致.

实体一致性校验则是针对特定领域任务进行的一种后处理方式, 通过命名实体识别技术, 对生成文本中的实体进行校验, 以确保这些实体与检索信息中的实体一致, 避免因信息不匹配或错误导致的生成内容不准确. 通过这些方法, 可以有效增强生成内容的上下文一致性和语义连贯性, 提升模型输出的质量与可靠性.

在应用方面, 某些专业领域对生成内容具有严格的标准, 并且要求其严格符合行业规范和准确性要求. 在这些领域, 单纯依靠解码策略调整模型的生成效果往往难以满足任务的高精度需求. 因此, 在模型生成内容后, 借助外部知识库进行验证成为一种有效的补充措施. 通过与领域特定的知识库或数据库对接, 可以验证生成内容的准确性与一致性, 确保生成的文本不仅符合专业规范, 还能体现真实的行业知识和实践. 此外, 知识库的集成还能够帮助识别生成内容中的潜在错误或遗漏, 进一步提高输出的质量和可信度. 这一验证过程, 尤其在复杂的领域任务中, 能够显著提升模型的可靠性和应用效果.

2.3 不同领域中的检索增强生成方法

检索增强生成技术的跨领域应用正逐步突破传统文本生成边界, 向多模态场景纵深拓展. 通过将外部知识检索与生成模型动态耦合, RAG 系统在不同数据模态中展现出差异化的技术路径与应用范式. 本节从文本、代码、图像、音视频以及混合多模态领域这 5 个方面展开系统性分析, 揭示核心挑战、技术演进与前沿实践.

2.3.1 文本领域的 RAG

作为 RAG 技术的核心落地场景, 文本领域的应用研究致力于通过动态知识检索与生成模型的深度协同, 系统性突破传统大语言模型在知识时效性、事实准确性及复杂语义推理等维度的固有局限. 其方法论的创新体现为构建“检索-增强-生成”的闭环增强链路.

在知识检索层, 系统通过 DPR[30]与知识图谱子图匹配方法 ToG (think-on-graph)[50]相结合的双轨策略, 实现对异构知识源的高效信息抽取. 其中, DPR 基于预训练的双编码器架构, 通过语义向量空间映射精准定位相关文本片段; ToG 则利用图神经网络解析知识图谱的拓扑结构, 动态提取与任务目标关联的子图单元, 支持多跳推理需求. 此外, 领域适配优化在此层尤为关键, 例如 DISC-LawLLM[51]通过引入法律术语增强的检索排序模型, 将法律

条文引用准确率提升至 91.4%, 显著优于传统关键词匹配方法, 验证了领域定制化检索模块对专业场景的必要性.

在知识增强层, 采用实体感知的稀疏记忆网络[52]或推理链引导的注意力机制, 实现检索结果与生成上下文的语义对齐, 有效抑制幻觉生成.

在生成优化层, 结合端到端预训练框架[53]与少样本提示工程, 动态调节生成模型对检索知识的依赖权重, 在开放域问答任务中实现 62.3% 的精确匹配率[54].

文本领域的 RAG 通过分层协同架构实现知识驱动型生成的范式革新, 其核心通过构建“检索-增强-生成”的闭环优化链路, 系统性突破传统大语言模型的固有瓶颈. RAG 技术在文本领域研究对比详见表 3. 未来, 该方向研究将聚焦于细粒度知识单元检索与增量式知识更新的融合, 以应对动态知识库场景下的时效性挑战, 同时探索检索-生成联合推理框架在复杂语义任务中的深度应用.

表 3 RAG 技术在文本领域研究对比评价

任务	模型/框架	核心思想	评价
QA问答	REALM[53]	双塔稠密检索架构	将隐式知识检索与预训练语言模型深度结合, 为后续RAG技术奠定了重要基础
	SKR[55]	自我知识评估	利用大语言模型的自我认知来改善生成质量, 有效提升了检索相关性与生成结果的准确性和一致性
	KAPING[56]	知识图谱检索	结合无训练检索增强与知识图谱, 有效提升零样本问答性能
	ToG[50]	迭代束搜索; 知识引导推理	通过交互式知识图谱探索与路径优化, 显著增强大模型的深度推理能力
	CL-ReLKT[57]	语言知识迁移	显著提升跨语言检索问答的性能及下游任务效率, 兼具速度优势
常识推理	KG-BART[58]	知识图谱检索	深度融合知识图谱与预训练模型, 显著提升生成文本的常识推理能力与逻辑性, 且具备强泛化性
人机对话	ConceptFlow[59]	图注意力引导的遍历机制	图谱显式建模对话流, 提升人机对话的连贯性与信息量, 参数效率高
	Skeleton-to-response[60]	骨架提取; 响应生成	骨架提取与两阶段生成优化检索信息利用, 有效提升对话信息量与生成质量
	BlenderBot 3[61]	千亿级混合专家架构	凭借超大规模参数、实时网络接入及持续学习机制, 显著提升开放域对话的连贯性与信息丰富度, 并强化安全可控性
	CEG[62]	后处理引用增强生成	无需训练即插即用, 有效提升生成内容可信度与抗幻觉能力
	KAUCUS[63]	模拟器框架	显著提升对话数据多样性及助手模型训练效果
文本摘要	Unlimiformer[64]	kNN索引	通过kNN索引突破输入长度限制, 实现长文本无损摘要
	RIGHT[65]	动态检索器	提升摘要的准确性和主流相关性, 尤其擅长整合新信息与全局语义
	M-RAG[66]	多分区记忆组织; 多智能体强化学习	结合多智能体强化学习动态优化任务适配性, 且具备跨模型架构的通用性优势

2.3.2 代码领域的 RAG

在软件工程与计算机科学领域, 基于检索增强生成技术的应用正逐渐成为提高开发效率和代码质量的关键工具. 特别是在代码生成、代码补全以及代码摘要等方面, RAG 展现了其独特的优势和广泛的应用前景.

在代码生成领域, 2021 年 Parvez 等人[67]提出一种新的检索增强框架 REDCODER, REDCODER 模仿软件开发者在编写代码或总结文档时回忆过去编写的代码片段或摘要的行为, 通过从检索数据库中提取相关的代码或摘要, 将其作为补充信息提供给代码生成或总结模型. 在 Java 和 Python 两个基准数据集上进行的广泛实验和分析, 验证了 REDCODER 在代码生成和总结任务中的有效性.

2023 年, 由 Salesforce Research 团队[68]提出的 CodeT5+模型, 通过灵活的编码器-解码器架构, CodeT5+采用了检索增强机制, 以改进代码生成任务的性能. 其检索组件用于匹配相关的代码示例, 而生成组件则结合检索到的代码片段进行代码生成. 这种方法显著提升了代码生成的准确率, 特别是在零样本和少样本学习场景下.

在代码补全领域, CoCoMIC[69]提出了一种静态代码分析工具 CCFinder, 通过构建项目上下文图 (project context graph) 检索跨文件实体, 并结合联合注意力机制将这些上下文与本地代码联合建模. 实验表明, 引入跨文件

上下文可使代码补全的精确匹配 (exact match, EM) 率提升 33.94%, 标识符匹配率提升 28.69%. 其核心改进在于通过结构化检索压缩跨文件信息, 避免输入长度超限问题.

2024 年, Liang 等人[70]进一步提出 RepoFuse 模型, 该模型采取“双上下文”策略, 将代码类比上下文与语义理性上下文结合, 并通过 RTG (rank truncated generation) 技术动态筛选关键信息, 在保证精度的同时压缩提示长度至 75.4%. 同年, RepoFusion[71]采用 fusion-in-decoder 架构, 训练模型融合多个仓库上下文 (如导入关系、类继承、相似代码片段), 使 220M 参数的模型在单行补全任务中接近 70 倍规模的 StarCoderBase 的性能 (差距仅 1.3%). 其关键在于通过分块编码与联合注意力机制, 高效利用多源上下文, 显著减少对私有或未训练代码库的“幻觉”生成.

RAG 技术正逐步成为代码智能生成领域的重要方向, 通过引入外部知识检索机制, 有效缓解了深度学习模型在代码生成、补全及摘要任务中的数据不足和泛化性问题. RAG 技术在代码领域研究对比详见表 4. 未来研究可进一步聚焦于更精细的检索策略、多模态代码表示及高效融合机制, 以提升模型的适应性和计算效率, 从而推动智能软件开发的持续发展.

表 4　RAG 技术在代码领域研究对比评价

任务	模型/框架	核心思想	评价
代码生成	RECODE[72]	动态规划检索; 子树检索	显式结合代码结构特征有效提升了复杂代码生成能力
	REDCODER[67]	密集检索; 模态切换	通过检索增强机制有效模拟开发者代码复用行为, 在代码生成与摘要任务中展现出实用性和性能优势
	DocPrompting[73]	文档驱动	引入文档检索机制, 有效解决模型对未见过API的泛化问题
	CodeT5+[68]	组合编码器-解码器模块	在架构设计与训练策略上进行双重改进, 以灵活性和高效性在20余项代码任务中刷新SOTA
	AceCoder[74]	引导式代码生成	针对性解决代码生成中的需求理解与实现难题
	kNN-TRANX[75]	语法约束降噪	通过语法引导的检索增强策略, 在保证代码语法正确性的同时平衡生成质量与效率
	A^3-CodGen[76]	动态信息融合; 多层级上下文整合	通过深度挖掘代码仓库的上下文关联, 解决现有工具与开发场景脱节的问题
	SkCoder[77]	多模型适配; 草图驱动生成	通过引入代码草图作为中间表示, 将代码复用与精细化编辑结合, 在多个数据集上大幅超越现有方法
	CodeGen4Libs[78]	两阶段生成	通过分阶段聚焦库依赖与代码实现, 有效解决了传统模型在面向第三方库代码生成中的适配难题
	EvoR[79]	同步演化查询; 多样化知识库	在动态更新和多样化信息源的支持下, 显著提高了执行准确性
代码补全	ReACC[80]	多模态检索策略; 分阶段联合训练	引入外部代码库的动态检索机制, 有效扩展了传统代码补全的上下文范围, 在CodeXGLUE基准测试中达到SOTA
	RepoCoder[81]	结合相似性检索与预训练代码语言模型	有效利用仓库级上下文信息, 显著提升代码补全质量, 为代码生成任务提供更精准和通用的解决方案
	RepoFusion[71]	训练时融合仓库上下文	在代码补全任务上实现了远超大规模模型的效果, 展现了仓库级上下文训练的强大潜力
	REPOFUSE[70]	rank truncated generation (RTG)技术	通过创新性融合类比与推理上下文, 并采用RTG技术压缩提示, 在确保代码补全精度的同时显著优化推理速度
	De-Hallucinator[82]	基于API参考信息进行上下文优化	在代码补全和测试生成任务上都实现了显著提升, 为LLM在软件开发中的应用提供了更加可靠的方案
代码摘要	EditSum[83]	相似代码片段检索	增强摘要的关键字匹配能力, 从而提升代码摘要的准确性和可读性
	CoRec[84]	相似Diff检索增强	提供了一种更精准的自动提交信息生成方法
	RACE[85]	示例引导器	进一步提升了自动提交信息生成的准确性, 更好地理解代码变更
	BashExplainer[86]	CodeBERT预训练	结合CodeBERT预训练和检索增强机制, 有效提升了Bash代码注释生成的准确性
其他	De-fine[87]	无训练的视觉编程; 自反馈优化	De-fine通过任务分解和自反馈优化机制, 显著提升了视觉-语言任务的逻辑推理能力
	Code4UIE[88]	代码化信息抽取	利用代码生成的方式统一信息抽取任务, 通过检索增强策略提升生成质量

2.3.3 图像领域的 RAG

在图像领域中, RAG 技术通过引入外部多模态知识库与可微分检索机制, 显著提升了生成模型的准确性和泛化能力. 其核心思想在于将传统参数化模型的生成能力与非参数化检索的动态知识融合, 尤其在处理罕见实体、复杂场景或知识密集型任务时展现出独特优势.

2023 年, Chen 等人[89]提出的 Re-Imagen 模型针对文本到图像生成中低频实体 (如“Chortai 犬种”) 的生成失真问题做出优化. Re-Imagen 通过检索外部知识库中的图像-文本对, 提取语义与视觉特征作为参考, 结合级联扩散模型生成高保真图像. 该方法在 EntityDrawBench 基准测试中, 生成忠实度较基线模型提升 40%, 显著减少了幻觉现象.

2022 年, Sarto 等人[90]提出了一种检索增强的 Transformer 架构, 用于图像描述生成任务. 传统的图像描述模型主要依赖于内部参数来生成描述, 而该方法通过引入 k 近邻 (kNN) 检索机制, 从外部语料库中获取相关文本信息, 以增强生成过程. 具体而言, 模型包含 1 个基于视觉相似性的知识检索器、1 个可微分的编码器以及 1 个 kNN 增强的注意力层. 在 COCO 数据集上的实验结果显示, 利用外部记忆可以提高生成描述的质量.

RAG 技术在图像生成与理解任务中的应用展现出广泛前景. 通过融合非参数化检索与参数化生成模型, 该方法有效缓解了低频实体生成偏差、图像描述质量受限等挑战. 近年来, 研究者不断探索优化检索机制、增强多模态特征融合策略, 以提升生成模型的准确性与泛化能力. RAG 技术在图片领域研究对比详见表 5. 未来, 随着大规模多模态知识库的构建与高效检索技术的发展, RAG 方案有望进一步拓展至更多视觉任务, 如跨模态推理、自动视觉内容创作等, 为智能视觉系统的构建提供更强的知识支撑.

表 5 RAG 技术在图片领域研究对比评价

任务	模型/框架	核心思想	评价
图像生成	RetrieveGAN[91]	可微检索; 端到端训练	通过可微检索模块优化特征学习, 提高生成图像的质量和合理性
	Re-Imagen[89]	外部多模态知识库	显著提升了对稀有和未见过实体的图像生成质量, 尤其在生成多样性和真实感方面表现出色
	KNN-Diffusion[92]	k近邻检索技术	在分布外图像生成和局部语义控制方面展现出较强的优势
图像描述生成	REVEAL[93]	多模态知识源	结合丰富的多模态知识源, 显著提升了在视觉问答和图像描述任务上的表现
	SmallCap[94]	CLIP编码器和GPT-2解码器	引入轻量级的架构和训练无关的检索机制, 不仅减少了训练成本, 而且提升了对新领域的适应能力
其他	OK-VQA[95]	密文段落检索	通过将检索和生成模块进行联合训练, 实现了更高效的外部知识检索与答案生成, 提升了OK-VQA任务的整体性能
	MMT[96]	多模态机器翻译	通过短语级别检索, 结合条件变分自编码器, 有效缓解了多模态机器翻译中的数据稀缺问题

2.3.4 音视频领域的 RAG

在音视频领域, RAG 技术已经成为一种重要方法, 用于提升视频字幕生成、视频问答以及相关任务的性能. RAG 通过结合大规模预训练语言模型和多模态信息检索机制, 实现了对开放域问题的零样本学习能力, 并且在多个基准测试中取得了最新的最佳表现.

具体来说, 在视频字幕生成方面, RAG 技术利用视觉-文本检索模块从外部文本语料库中获取上下文文本. 例如, 采用 CLIP 模型[97]的 ViT-L/14 变体计算视频帧特征与文本特征之间的相似度得分, 并基于这些得分进行排序以获得每个视频帧的前 k 个匹配文本. 为了捕捉帧间的时序转换, 进一步引入时序感知提示, 将检索到的字幕按时间顺序排列, 形成连贯的叙述. 这种结构不仅提高了生成字幕的相关性和准确性, 还增强了模型处理复杂视频内容的能力. 在视频问答系统中, RAG 方法展示出强大的跨模态检索能力, 通过冻结大语言模型并利用多模态基础模型, 该方法能够在无需额外微调的情况下适应新任务或领域. 实验表明, 即使参数量较少, RAG 的表现也超越了具有 80B 参数的跨模态训练模型 Flamingo.

RAG 技术通过融合先进的检索技术和深度学习模型, 在音视频分析领域展现出巨大潜力, 特别是在提高生成内容的准确性和泛化能力方面. RAG 技术在音视频领域研究对比详见表 6. 随着研究的深入和技术的进步, 未来有

望看到更多基于 RAG 的应用出现, 进一步推动多媒体理解和交互的发展.

表 6 RAG 技术在音视频领域研究对比评价

任务	模型/框架	核心思想	评价
视频描述生成	CARE[98]	多模态驱动	通过引入精准的概念检测和高效的语义引导机制, 显著提高了视频描述生成的质量
	EgoInstructor[99]	跨视角检索增强	结合跨视角 (第1人称 vs. 第3人称)检索和视频描述生成, 成功提升第1人称视频描述的质量
视频对话	MA-DRNN[100]	基于记忆增强的深度递归神经网络	该方法能够更准确地理解视频内容, 并回答相关问题, 为未来基于记忆增强的多模态学习提供新的研究方向
	STAGE[101]	时空证据的联合推理	显著提升VideoQA任务的准确性和可解释性
	VGNMN[102]	实体与动作推理	为未来基于视频的智能对话系统提供了新的思路

2.3.5 多模态下的 RAG

与传统的单模态 RAG 主要依赖纯文本或图像的检索机制不同, 多模态 RAG 在检索策略上实现了对文本、图像、音频、代码等异构模态的联合建模与联合检索, 突破了单模态检索中模态受限、信息单一的问题.

在多模态 RAG 系统中, 模态中心是决定检索策略设计与生成效果的关键因素. 模态中心指在任务执行过程中, 某一特定模态 (如文本、图像或视频) 承担主导信息载体的角色, 而其他模态则作为辅助补充. 这种设计能够有效缓解多模态数据异构性带来的对齐与融合挑战, 同时提升模型在特定任务中的可解释性. 本节以不同模态中心为视角来梳理代表性研究.

文本作为信息密度最高的模态, 仍然是多模态 RAG 系统的重要基础. 在文本证据检索过程中, 常用策略包括 BM25[103]等稀疏检索方法, 以及以 MiniLM[104]、BGE-M3[105]为代表的密集检索方法. 为了解决对细粒度语义匹配和领域特定性的需求, 一些新的方法, 如 ColBERT[106]和 PreFLMR[107], 采用 token 级交互机制, 保留细微的文本细节以提高多模态查询的精确度, 而 RAFT[108]和 CRAG[109]通过确保文本片段的精确引用来提高检索性能.

以视觉作为中心的检索偏向于利用图像表示进行知识提取. EchoSight[110]和 ImgRet[111]等系统通过使用参考图像作为查询来检索视觉相似内容. 此外, 组合图像检索 (composed image retrieval, CMI) 模型通过将多个图像特征集成到一个统一的查询表示中, 增强了检索效果. 类似地, Pic2word[112]将视觉内容映射到文本描述, 实现了零样本图像检索.

视频中心化检索技术在视觉时序动态建模与大型视频语言模型 (large vision-language model, LVLM) 的协同驱动下取得系统性突破: iRAG[113]通过增量检索机制实现视频序列的渐进式语义解析, 强化时序连贯性建模; MV-Adapter[114]提出解耦式跨模态迁移学习架构, 利用动态参数微调优化视频-文本双向检索的领域泛化能力. 针对长上下文处理挑战, Video-RAG[115]通过视觉对齐的辅助文本 (OCR/ASR) 增强跨模态检索鲁棒性; T-Mass[116]将文本建模为随机嵌入以提升文本-视频语义匹配效率; VideoRAG[117]设计双通道架构并引入基于图的认知基模实现超长视频的全局语义关联. 在时序推理方向, CTCH[118]采用对比 Transformer 哈希算法建模长程时空依赖; RTime[119]则通过反向视频硬负样本构建严格评估时序因果关系的基准. 面对复杂视频理解需求, OmAgent[120]提出分治式多智能体框架解耦多粒度语义分析任务, DrVideo[121]则基于文档化检索策略实现长视频结构化语义提取. 上述技术从时序建模、跨模态对齐、复杂理解这 3 个维度推动视频检索从局部片段分析向全域时空感知的范式跃迁, 为多模态智能系统提供可扩展的技术底座.

多模态 RAG 通过确立文本、视觉或视频等模态中心, 并辅以创新的检索机制 (如细粒度文本匹配、图像映射、时序建模与复杂理解技术), 有效克服了多源异构数据的融合挑战, 显著提升了信息检索的精度、鲁棒性和可解释性. 这种以模态中心为核心的策略设计, 不仅推动了检索范式从单一模态向全域时空感知的跃迁, 也为构建更强大的多模态智能系统奠定了坚实的技术基础.

2.4 数据集

在 RAG 领域中, 评估数据集的选择至关重要, 需覆盖检索质量、生成效果及两者协同性能. 本节对主流数据

集及其表现进行介绍, 各数据集评价汇总详见表 7.

表 7 主流数据集评价汇总

分类	数据集/评估框架	语言	数据/框架特点	评价
开放域问答	Natural Questions[122]	英文	多答案; 长文档	Natural Questions (NQ)作为开放域问答的黄金标准数据集, 为RAG技术提供了真实且复杂的评测环境
	TriviaQA[123]	英文	多证据支持; 多答案	TriviaQA以多跳推理和多证据依赖特性为核心, 成为评估RAG系统高阶认知能力的关键基准
	MS MARCO[124]	英文	人工精标答案; 多粒度文档库	MS MARCO凭借其大规模真实场景数据与工业级任务设计, 成为RAG技术从实验室走向实际应用的桥梁
事实验证	FEVER[125]	英文	证据层级依赖; 细粒度标注	Fever作为事实核查领域的权威评测基准, 通过对抗性声明与多证据依赖设计, 为RAG系统验证知识可信性和提升生成可靠性提供了评测环境
	AmbigNQ[126]	英文	问题多义性; 证据多样性; 多答案	AmbigNQ通过系统化建模自然语言中的显式歧义性, 为RAG技术在真实对话场景中的落地提供了关键评测基准
多跳检索	CLUE-AFQMC[127]	中文	领域专业性; 口语化; 短文本	BERT-base准确率约85%–88%, 在金融语料上继续预训练的模型 (如BERT-financial)表现更优
	HotpotQA[128]	英文	多跳推理需求	HotpotQA凭借其多跳推理复杂度与干扰文档挑战, 成为检验RAG系统高阶认知能力的核心基准
	2WikiMultihopQA[129]	英文	跨页面推理; 噪声与干扰控制	2WikiMultihopQA凭借其跨页面推理复杂度与高干扰环境设计, 成为检验RAG系统跨文档语义关联与深度推理能力的核心基准.
	DuReader-Retrieval[130]	中文	文档多样化; 语义匹配复杂	DuReader-Retrieval填补了中文大规模真实检索数据集的空白, 尤其适用于测试模型在复杂场景下的鲁棒性

2.4.1 开放域问答数据集

Natural Questions (NQ)[122]是谷歌于 2019 年发布的开放域问答 (open-domain QA) 基准数据集, 旨在模拟真实用户的信息需求场景. 其核心目标是通过真实用户的自然语言查询 (基于 Google 搜索日志), 要求模型从维基百科长文档中检索并提取精确答案, 从而验证模型在复杂语义理解、知识检索与答案生成中的综合能力. 在 RAG 技术研究中, NQ 被广泛用作端到端性能评估基准, 因其高度贴合实际应用场景, 能够全面检验模型在检索-生成协同中的能力.

TriviaQA[123] 是 2017 年发布的开放域问答基准数据集, 专为测试模型对复杂、多跳推理问题的解答能力而设计. 其包含约 95k 问答对, 问题多为长句式、多实体嵌套, 需多步骤推理或跨段落信息融合. 除此之外, 每个问题关联平均 6.2 篇维基百科文档 (总证据文档超 65 万篇), 答案需从多个文档中提取并验证.

MS MARCO[124]是微软于 2016 年发布的开放域问答与文档检索基准数据集, 旨在模拟真实搜索引擎的用户查询场景. 其核心目标是推动模型信息检索 (information retrieval, IR) 与机器阅读理解 (machine reading comprehension, MRC) 的联合优化, 通过大规模真实用户问题与人工生成答案, 验证模型从海量网页文档中快速定位并生成简洁答案的能力.

2.4.2 事实验证数据集

FEVER (fact extraction and verification)[125]是 2018 年发布的事实核查基准数据集, 包含 18.5 万条人工编写的声明, 涵盖真实 (supported)、虚假 (refuted) 及证据不足 (not enough info) 这 3 类标签, 其中虚假声明通过实体替换、关系扭曲等对抗策略生成. FEVER 通过对抗性声明验证与多证据推理需求, 为 RAG 系统提供了检索模块的对抗鲁棒性与生成模块的事实一致性的评测场景.

AmbigNQ (ambiguous Natural Questions) 数据集[126]是 2020 年提出的自然语言问答数据集, 专注于用户查询中的歧义性 (ambiguity) 场景. 数据集中包含约 14k 人工标注的问题, 其中每个问题至少存在两种可能的解释, 为 RAG 技术在真实对话场景中的落地提供了关键评测基准.

CLUE-AFQMC (ant financial question matching corpus)[127]是中文语言理解评估基准 (CLUE) 中的一个重要数

据集, 由 CLUE 组织联合蚂蚁金服共同推出, 包括约 10 万条数据, 划分为训练集 (约 8 万条)、验证集 (约 1 万条) 和测试集 (约 1 万条).

2.4.3 多跳检索数据集

HotpotQA[128] 由斯坦福大学于 2018 年推出, 是首个专注于多跳问答 (multi-hop QA) 的基准数据集, 其中包含约 11.3 万个人工标注的问题, 每个问题需通过至少两个支持段落 (supporting fact) 进行推理.

2WikiMultihopQA[129] 是 2020 年提出的多跳问答数据集, 专注于需要跨两个独立维基百科页面进行推理的复杂问题. 其中包含约 5.7 万个问题, 每个问题需通过至少两个维基百科页面的信息进行推理. 2WikiMultihopQA 通过其跨页面推理挑战与高噪声环境, 为 RAG 系统提供了独特的评测环境.

DuReader-retrieval[130]是百度推出的面向中文开放域问答的文档检索数据集, 其中包含约 10 万条真实用户提问, 约 800 万篇网页文档 (涵盖新闻、百科、论坛等多来源), 填补了中文大规模真实检索数据集的空白, 尤其适用于测试模型在复杂场景下的鲁棒性.

2.5 主流 RAG 方法的基准评测与对比分析

本节在上述关于 RAG 技术体系、典型任务场景及相关评测资源梳理的基础上, 系统整理主流 RAG 方法在典型基准数据集上的评测结果. 这里的主流 RAG 方法是指以文本检索增强生成为核心范式、在公开基准数据集上具有系统实验验证并构成后续研究基础的代表性方法, 并通过定量指标对不同方法的性能进行横向对比分析.

2.5.1 主流 RAG 方法在基准数据集上的评测结果汇总

本节对上述代表性 RAG 方法在不同基准数据集上的实验结果进行统一汇总. 为保证比较的客观性与可解释性, 本文采用原论文中报告的标准评测设置与核心评价指标, 并依据方法的主要评测任务对结果进行归类整理.

我们将现有方法按照评测范式划分为开放域问答 (open-domain question answering)、多跳推理问答 (multi-hop question answering) 以及检索增强语言建模 (retrieval-augmented language modeling) 这 3 类. 该划分基于方法在原始工作中的核心实验任务, 而非其潜在适用范围, 从而避免不同评测目标与指标之间的直接比较偏差.

在开放域问答任务中, 代表性方法包括 RAG[11]、FiD[131]、REALM[53]、Atlas[54]、REPLUG[132]与 Self-RAG[36], 通常结合稠密或多路检索器, 在 Natural Questions、TriviaQA、WebQuestions[133]及 ASQA[134]等数据集上进行评测, 并以 EM 与 $F1$ 分数作为主要指标. 实验结果表明, 随着检索与生成模块耦合方式的不断演进以及生成模型规模的扩大, 方法在多个数据集上的问答性能持续提升.

多跳推理问答方法如 ITER-RETGEN[135]、BlendFilter[136]、ReAct[137]与 SelfAsk[138]侧重于通过多轮检索与中间推理逐步构建答案, 主要在 HotpotQA 与 2WikiMultihopQA 等数据集上进行评测, 其性能由 EM 与 $F1$ 指标衡量, 用以反映模型在复杂推理场景下对检索信息的利用能力.

此外, 部分工作从语言建模角度引入检索机制, 通过外部记忆降低模型困惑度. 代表性方法包括 kNN-LM[139]与 RETRO[140], 通常在 WikiText-103 等语言建模基准上进行评测, 并以 perplexity 作为核心指标. 由于该类方法的评测目标与问答任务存在本质差异, 本文将其单独作为检索增强语言建模范式进行整理.

表 8 对上述方法在不同基准数据集上的实验结果进行了系统汇总, 涵盖检索器类型、生成模型结构、评测数据集、评价指标及对应性能, 为后续定量对比分析提供统一参考.

2.5.2 定量对比分析与评测趋势讨论

基于表 8 汇总的基准评测结果, 需要指出的是, 不同 RAG 方法在检索语料、检索次数、模型规模及评测设置等方面存在显著差异, 其报告的绝对性能数值不具备严格的可比性. 因此, 本文不对具体方法之间的分数高低进行直接比较, 而是从评测范式与模型设计选择的角度, 对实验结果所反映的整体趋势进行分析.

从检索范式的角度来看, 引入稠密检索或多路检索机制已成为开放域问答任务中的主流选择. 多数方法在 Natural Questions 与 TriviaQA 等数据集上采用稠密向量检索作为生成模型的主要外部知识来源, 表明高召回率检索是提升生成准确性的关键前提. 在此基础上, 通过融合多检索器或进行检索-生成联合训练, 相关方法在不同评测设置下展现出更强的鲁棒性与性能稳定性. 同时, 实验结果普遍表明, 对检索文档进行更充分的结构化建模有助

于进一步提升生成质量: 相较于简单拼接多篇文档作为输入, 采用独立编码并在解码阶段进行信息融合的策略在多个基准上呈现出更优的性能趋势, 说明生成模型对检索信息的细粒度利用对于结果具有重要影响.

表 8 主流 RAG 方法在不同基准数据集上的实验结果

任务类型	方法	检索器	生成模型	数据集	指标	性能 (%)
开放域问答	RAG[11]	DPR	BART-large	NQ	EM/*F*1	44.5/56.8
	FiD[131]	DPR	T5-base	NQ	EM/*F*1	51.4/67.6
	Atlas[54]	LDR	T5-11B	NQ	EM	56.4
	REALM[53]	JTDR	BERT-based LM	NQ	EM/*F*1	40.4/50.1
	REPLUG[132]	Retriever Plug-in	LLaMA2-7B	NQ	EM	45.5
	RAG[11]	DPR	BART-large	TriviaQA	EM/*F*1	56.1/68.9
	FiD[131]	DPR	T5-base	TriviaQA	EM/*F*1	67.6/79.5
	Atlas[54]	LDR	T5-11B	TriviaQA	EM	72.9
	REPLUG[132]	Retriever Plug-in	LLaMA2-7B	TriviaQA	EM	69.6
	Self-RAG[36]	DRC	LLaMA2-7B	TriviaQA	EM	69.3
	REALM[53]	JTDR	BERT-based LM	WQ	EM	40.7
	Self-RAG[36]	DRC	LLaMA2-7B	ASQA	EM	31.7
检索增强语言建模	RETRO[140]	BERT-based Chunk Retriever	GPT-style LM	WikiText-103	PPL↓	18.5
	kNN-LM[139]	kNN over Hidden States	Transformer LM	WikiText-103	PPL↓	15.8
多跳推理问答	ITER-RETGEN[135]	DR	Vicuna1.5-13b	HotpotQA	EM/*F*1	36.6/48.4
	BlendFilter[136]	—	Vicuna1.5-13b	HotpotQA	EM/*F*1	39.6/52.7
	ReAct[137]	—	Vicuna1.5-13b	HotpotQA	EM/*F*1	33.2/46.3
	SelfAsk[138]	—	Vicuna1.5-13b	HotpotQA	EM/*F*1	36.1/46.9
	ITER-RETGEN[135]	DR	Vicuna1.5-13b	2WikiMultihopQA	EM/*F*1	25.2/35.5
	BlendFilter[136]	—	Vicuna1.5-13b	2WikiMultihopQA	EM/*F*1	28.6/37.8
	ReAct[137]	—	Vicuna1.5-13b	2WikiMultihopQA	EM/*F*1	21.6/32.3
	SelfAsk[138]	—	Vicuna1.5-13b	2WikiMultihopQA	EM/*F*1	25.0/37.6

在更复杂的推理场景中, 评测结果显示多轮检索与显式中间推理是提升模型能力的关键因素. 相关方法通常通过构建中间问题或证据链逐步扩展检索范围, 在需要跨文档推理的数据集上表现出明显优势, 而单轮检索或隐式推理在此类任务中存在性能瓶颈. 类似地, 在检索增强语言建模任务中, 引入外部检索机制能够有效补充模型的参数化记忆, 在以困惑度为核心指标的评测中普遍优于不使用外部知识的基线模型. 总体而言, 现有基准结果更多反映了不同检索与生成设计选择在特定任务场景下的适用性, 而非方法间的绝对优劣, 这也进一步表明 RAG 系统的性能高度依赖于任务定义、评测设置与系统设计, 单一指标难以全面刻画其综合能力.

3 RAG 结合 LLM 领域应用分析

随着 RAG 技术与 LLM 的深度融合, 其在法律、医疗、金融等专业领域的应用呈现出差异化特征与共性挑战. 本节通过横向对比 3 大高合规性领域, 系统剖析 RAG+LLM 技术框架的跨领域适配机制: 一方面揭示其在知识动态更新、领域逻辑嵌入、合规约束满足等方面的通用技术路径; 另一方面对比不同场景下模型性能的边界条件, 如法律解释的权威性要求、医疗诊断的因果推理强度、金融预测的时效性阈值. 通过构建领域特异性挑战矩阵与技术演进图谱, 为突破专业壁垒、实现可靠赋能提供跨学科参考范式.

3.1 技术框架的跨领域适用性分析

RAG 与 LLM 的协同框架展现出显著的跨领域适用性, 其技术特征通过动态知识整合、领域适应性调整与合规约束嵌入这 3 个维度实现多场景赋能.

动态知识增强机制构成技术落地的核心基础. 在法律领域, 该框架通过实时接入法律条文修订库 (如中国最高

人民法院裁判规则数据库) 及增量更新判例知识图谱, 有效解决传统 LLM 因训练数据滞后导致的法条引用偏差问题. 类似地, 医疗场景中, RAG 系统需持续同步 PubMed 2[141] 等最新临床指南与 FDA (Food and Drug Administration) 药物审批数据, 而金融领域则依赖彭博终端等实时市场数据的毫秒级检索, 由此形成“领域知识保鲜”的共性技术范式.

技术框架在领域适应性上特征主要体现为知识处理范式的差异化重构. 法律应用强调逻辑推理链的完整性, 要求模型在劳动仲裁等场景中严格遵循“事实提取-法条映射-判例类比-结论推导”的司法认知路径; 医疗场景则需构建诊断依据的可追溯性架构, 通过 RAG 的检索路径可视化功能 (如病例特征与医学证据的显式关联) 满足循证医学要求; 金融领域侧重风险预测的时序敏感性, 模型通过嵌入计量经济学特征工程模块 (如 ARCH 模型[142]波动率因子), 将非结构化文本数据转化为量化分析输入.

合规性约束机制日益体现出跨领域技术耦合的特征. 在法律领域, RAG 系统需集成司法管辖区识别模块 (如基于 BERT 的文本分类算法), 以确保生成的法律文书符合特定地区的司法规范; 医疗领域则依赖符合 HIPAA 法案标准的去标识化检索协议, 以保障电子病历数据的安全访问与合法使用; 金融领域通过构建监管沙盒环境, 并引入对抗训练机制 (adversarial training), 识别模型输出中潜在的市场操纵性表述. 这些实践共同反映出一个核心挑战: 知识检索广度与合规性约束强度之间存在一定程度的负相关关系: 相关研究表明, 随着知识覆盖范围的拓展, 模型在输出合规性方面可能出现一定程度的下降. 因此, 通用技术框架需结合各领域特点引入专用调节参数, 以在知识获取能力与合规性要求之间实现动态平衡与最优调和.

由此可知, RAG 与 LLM 的协同框架跨领域迁移能力本质上源于其“知识检索-语义生成”的解耦架构: 检索模块通过领域知识编码器实现专业化适配, 生成模块保留 LLM 的通用语言理解能力, 二者通过注意力门控机制动态调节领域知识的注入强度. 这种弹性架构使得法律、医疗、金融领域分别形成各自的最优配置, 印证了技术框架在保持领域特化性的同时兼具扩展泛化能力的核心优势.

3.2 典型应用场景分析

RAG+LLM 技术框架在垂直领域的场景化应用中呈现出显著的领域分化特征, 其能力边界与实施路径受专业知识体系、任务目标及监管环境的共同塑造. 本节从智能咨询、文档生成、决策支持这 3 大核心场景切入, 构建跨领域对比分析矩阵, 揭示技术落地的共性与特性规律.

3.2.1 智能咨询与问答系统

智能咨询与问答系统在法律、医疗和金融等高专业性领域中正日益成为知识服务的核心技术支撑, 具有显著的现实意义与应用价值. 在法律领域, 此类系统能够提升公众对法规的可达性与可理解性, 辅助法律从业者快速获取法条、判例与司法解释, 降低咨询成本并增强法律服务的可及性. 在医疗领域, 智能问答系统有助于临床医生快速访问标准化术语体系与循证知识库, 支持诊疗决策并缓解医疗信息不对称问题, 尤其在基层医疗和远程医疗场景中表现出强大潜力. 在金融领域, 系统可实现对高频市场信息的快速响应与解释, 辅助投资者识别风险、捕捉机会, 在金融监管与智能投顾等场景中提供稳定可靠的知识支持.

在法律咨询中, 系统通常构建由法条、判例与司法解释构成的三元检索体系. 例如, 面向劳动仲裁的 RAG 架构可通过法律领域专用的语言模型嵌入《劳动合同法》第 38 条与典型指导案例之间的语义关联, 实现面向具体维权情境的高精度答案生成. 医疗领域则强调术语标准化与临床证据的双重校验机制, 其检索模块需融合国际通用的医学术语体系与循证知识库, 同时引入自动化编码转换工具, 以降低诊断术语歧义带来的知识偏差. 金融应用场景则具有更强的时序敏感性, 相关系统通常集成混合神经网络结构, 对市场新闻进行波动率加权建模, 实现对时效性要求极高的投资建议生成, 其平均响应延迟被严格控制在分钟级以内.

上述差异反映出各领域在知识结构组织、检索路径构建以及输出响应机制上的深层异构性, 提示智能咨询系统的设计需充分结合场景特性, 构建高度定制化的技术栈以满足其专业需求.

3.2.2 文档生成与合规审查

在文档自动化生成场景中, 领域间的差异化需求正深刻推动技术架构向任务定制化方向演进.

法律文书生成强调格式规范与逻辑结构的严格一致性, 系统通常需检索历史案例中的“争议焦点-证据链-法理分析”三段式内容模板, 并引入逻辑约束解码技术以确保生成结果符合三段论的法律推理逻辑.

医疗领域则更加注重内容的可解释性与临床可溯源性, 部分智能报告系统通过检索患者历史影像特征与临床指南之间的关联证据, 在生成过程中自动标注关键判断依据, 如影像评分等级和参考标准, 以增强医生对结论的信任度[143].

金融合同自动生成与审查面临持续变化的监管政策挑战, 先进系统已集成以监管条款为导向的对抗训练机制, 能够在生成过程中动态检测并规避与现行法律法规的潜在冲突, 从而实现合规性保障与文本智能生成的同步协同.

3.2.3 决策支持系统

决策支持场景中的技术实现深刻体现了各领域独有的认知范式差异.

法律领域的类案推荐系统通常构建基于“案情要素-法律定性-裁判结果”的异构图谱, 并结合图神经网络实现语义关联建模. 在检索策略上, 多维相似度度量方法得以应用, 诸如事实相似度与法条重合度的双重阈值设定, 有效提升了类案推送的精准性与司法一致性.

医疗领域则以循证医学为核心范式, 先进的治疗方案推荐系统通常融合指南证据等级与个体基因组数据, 通过生成式检索机制动态匹配最优治疗路径, 从而显著提升高复杂度病例 (如肿瘤Ⅲ期) 的方案制定效率.

相较之下, 金融决策支持系统强调对结构化与非结构化数据的融合分析. 前沿系统已实现将监管文本、财报文件中的情绪信号与金融市场的时序指标 (如信用利差) 结合建模, 显著提升风险事件预测的准确率与响应灵敏度.

3.3 技术演进方向

当前研究围绕 RAG 结合 LLM 的技术迭代主要聚焦于领域知识融合、混合架构优化与合规性增强这 3 大方向, 其演进路径呈现出显著领域适配特征与协同创新趋势.

(1) 领域知识嵌入的深度专业化

针对垂直领域知识体系的异构性, 技术发展正从通用检索增强向专业化知识嵌入转型. 在法律领域, 基于司法语料微调的预训练模型 (如 LEGAL-BERT[144]) 通过嵌入法律实体识别与逻辑关系抽取模块, 显著提升了对法律文本隐含语义的解析能力. 医疗领域则通过构建生物医学知识图谱, 将疾病本体、药物相互作用等结构化知识注入 RAG 检索管道, 使系统能够实现诊疗指南与病例特征的动态对齐. 金融场景中, 时序预测专用模型 (如 FinBERT[145]) 通过融合宏观经济指标与市场情绪数据, 增强了 LLM 对金融文本的因果推理能力. 研究表明此类技术突破使得 RAG 结合 LLMs 在各领域的知识检索精度平均提升 23–41%.

(2) 规混合增强架构的跨模态协同

为突破单一文本模态的处理局限, 技术框架正向多模态混合架构演进. 法律应用中, 符号推理引擎与 RAG 的深度集成成为关键创新方向, 例如将法律条文的形式化逻辑规则 (如霍菲尔德法律元模型) 嵌入生成流程, 有效约束 LLM 的合规输出. 医疗场景则通过集成影像识别模块与文本生成模型, 构建涵盖病理图像分析与诊断报告生成的端到端系统. 金融领域进一步开发量化分析增强架构, 通过连接期权定价模型与风险预测 LLM, 实现投资策略的数学验证与自然语言解释双轨输出. 实验表明, 混合架构可使决策支持系统的综合效能提升 58%.

(3) 合规性增强的技术治理体系

面对各领域差异化监管要求, 技术演进呈现出“嵌入治理”与“主动适应”的双重特征. 法律领域开发司法管辖区识别系统, 通过动态检测用户地理位置与案件属性, 自动匹配属地法律知识库 (如我国《中华人名共和国民法典》与欧盟 GDPR 的冲突条款过滤). 医疗场景构建去识别化数据处理流水线, 采用差分隐私与联邦学习技术, 在保障 HIPAA 合规的同时维持模型性能损失率低于 5%. 金融领域则创建监管沙盒测试环境, 通过实时接入 SEC/FCA 等监管机构政策更新 API, 使系统输出同步满足多法域披露要求.

未来技术发展将呈现两大趋势: 其一, 领域专用大模型 (domain-specific LLM) 与 RAG 组件的紧耦合架构, 通过知识嵌入向量空间的联合训练实现检索-生成一体化优化; 其二, 跨领域知识迁移机制的突破, 例如将法律案例的类比推理模式迁移至医疗诊断决策支持, 形成更具通用性的增强框架.

4 展 望

未来, RAG 技术仍将面临广阔的研究空间与应用前景, 其核心能力也将随着大语言模型的发展不断演进. 从技术演进路径来看, 一方面, 如何进一步提升检索结果与生成内容之间的语义一致性与因果连贯性, 是提升 RAG 推理能力与事实可靠性的关键方向; 另一方面, 在数据规模持续扩展的背景下, 隐私保护、公平性与安全性问题也将对 RAG 系统的设计与部署提出更高要求.

除此之外, 面向真实应用场景时, RAG 技术仍需在系统效率、可解释性以及跨领域适应能力等方面取得突破. 随着模型结构、检索策略与知识表示方式的不断融合与优化, RAG 有望从“辅助生成”的技术形态, 逐步演进为支撑复杂决策与专业推理的通用智能框架. 未来, RAG 技术不仅将在法律、医疗等知识密集型领域持续发挥重要作用, 也将为更多行业的智能化转型提供关键支撑.

5 结束语

本文从大语言模型在实际应用中面临的知识时效性不足、幻觉现象频发以及领域适应能力受限等问题出发, 系统梳理并分析检索增强生成 (RAG) 技术在提升模型生成质量与知识覆盖能力方面的研究进展. 围绕 RAG 的两大核心模块: 信息检索与生成增强, 构建较为系统的技术分类框架, 全面回顾了 RAG 技术在文本、代码、图像、音视频及多模态场景中的典型应用与最新进展.

在数据资源层面, 本文对现有中英文 RAG 相关数据集进行系统收集与整理, 并依据不同任务属性进行分类与分析, 为后续研究在数据选取与方法设计方面提供了参考依据. 进一步地, 结合法律、医疗和金融等专业领域的实际应用需求, 本文分析了 RAG 技术在特定领域场景中的应用效果, 总结其在缓解知识缺失、提升生成可信度方面的优势, 同时也揭示了其在复杂推理任务中的现实局限.

尽管 RAG 技术在增强大语言模型生成能力方面展现出显著成效, 但其当前仍面临若干尚未解决的关键问题, 这些问题在一定程度上源于 RAG 方法本身的理论与结构性限制. 首先, 在复杂查询语义理解方面, 现有 RAG 系统通常将检索与生成视为相对独立的两个阶段, 检索模块多依赖静态向量表示或浅层语义匹配, 难以充分刻画查询中隐含的推理意图与多步逻辑关系, 从而导致检索结果与生成需求之间存在语义偏差. 其次, 在动态知识适应能力方面, RAG 依赖的外部知识库往往存在更新滞后或结构松散的问题, 而生成模型本身又缺乏对知识时效性和可信度的内在建模机制, 使得系统难以在快速变化的知识环境中保持稳定表现. 最后, 在跨领域泛化能力方面, 检索模型、生成模型与领域数据之间的耦合关系较为紧密, 现有方法往往依赖领域特定的数据分布与先验假设, 导致其在迁移至新领域或低资源场景时性能显著下降.

因此, 从更深层次来看, RAG 技术未来的发展不仅依赖于工程层面的系统优化, 更依赖于对检索与生成协同机制、知识表示形式以及推理过程建模方式的进一步理论探索. 如何在统一框架下实现语义理解、知识检索与生成推理的深度融合, 将是 RAG 技术迈向更高可靠性与通用性的关键研究方向.

References

[1] Shannon CE. A mathematical theory of communication. The Bell System Technical Journal, 1948, 27(3): 379–423. [doi: 10.1002/j.1538-7305.1948.tb01338.x]

[2] Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 1993, 19(2): 263–311.

[3] Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. The Journal of Machine Learning Research, 2003, 3: 1137–1155.

[4] Graves A. Long short-term memory. In: Graves A, ed. Supervised Sequence Labelling with Recurrent Neural Networks. Heidelberg: Springer, 2012. 37–45. [doi: 10.1007/978-3-642-24797-2_4]

[5] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555, 2014.

[6] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the

31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[7] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[8] Achiam J, Adler S, Agarwal S, *et al*. GPT-4 technical report. arXiv:2303.08774, 2023.

[9] Touvron H, Martin L, Stone K, *et al*. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288, 2023.

[10] Yang A, Li AF, Yang BS, *et al*. Qwen3 technical report. arXiv:2505.09388, 2025.

[11] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 793.

[12] Chen DQ, Fisch A, Weston J, Bordes A. Reading Wikipedia to answer open-domain questions. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017. 1870–1879. [doi: 10.18653/v1/P17-1171]

[13] Li SY, Zhang LY, Wang ZD, Wu D, Wu LR, Liu ZC, Xia J, Tan C, Liu Y, Sun BG, LI SZ. Masked modeling for self-supervised representation learning on vision and beyond. arXiv:2401.00897, 2024.

[14] Henighan T, Kaplan J, Katz M, Chen M, Hesse C, Jackson J, Jun H, Brown TB, Dhariwal P, Gray S, Hallacy C, Mann B, Radford A, Ramesh A, Ryder N, Ziegler DM, Schulman J, Amodei D, McCandlish S. Scaling laws for autoregressive generative modeling. arXiv:2010.14701, 2020.

[15] Chowdhery A, Narang S, Devlin J, *et al*. PaLM: Scaling language modeling with pathways. The Journal of Machine Learning Research, 2023, 24(1): 240.

[16] Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. Neural Computation, 1991, 3(1): 79–87. [doi: 10.1162/neco.1991.3.1.79]

[17] Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, Dean J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[18] Lepikhin D, Lee H, Xu YZ, Chen DH, Firat O, Huang YP, Krikun M, Shazeer N, Chen ZF. GShard: Scaling giant models with conditional computation and automatic sharding. In: Proc. of the 9th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2021.

[19] Du N, Huang YP, Dai AM, *et al*. GLaM: Efficient scaling of language models with mixture-of-experts. In: Proc. of the 39th Int'l Conf. on Machine Learning. Baltimore: PMLR, 2022. 5547–5569.

[20] Bai YT, Kadavath S, Kundu S, *et al*. Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073, 2022.

[21] Ji ZW, Lee N, Frieske R, Yu TZ, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. ACM Computing Surveys, 2023, 55(12): 248. [doi: 10.1145/3571730]

[22] Liu ZH, Wang PJ, Song XB, Zhang X, Jiang BB. Survey on hallucinations in large language models. Ruan Jian Xue Bao/Journal of Software, 2025, 36(3): 1152–1185 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7242.htm [doi: 10.13328/j.cnki.jos.007242]

[23] Kadavath S, Conerly T, Askell A, *et al*. Language models (Mostly) know what they know. arXiv:2207.05221, 2022.

[24] Ji ZW, Yu TZ, Xu Y, Lee N, Ishii E, Fung P. Towards mitigating LLM hallucination via self reflection. In: Findings of the Association for Computational Linguistics: ACL 2023. Singapore: ACL, 2023. 1827–1843. [doi: 10.18653/v1/2023.findings-emnlp.123]

[25] Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, Liang P. Lost in the middle: How language models use long contexts. Trans. of the Association for Computational Linguistics, 2024, 12: 157–173. [doi: 10.1162/tacl_a_00638]

[26] Lin XV, Chen XL, Chen MD, Shi WJ, Lomeli M, James R, Rodriguez P, Kahn J, Szilvasy G, Lewis M, Zettlemoyer L, Yih WT. RA-DIT: Retrieval-augmented dual instruction tuning. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2024.

[27] Manakul P, Liusie A, Gales M. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 9004–9017. [doi: 10.18653/v1/2023.emnlp-main.557]

[28] Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A, Attariyan M, Gelly S. Parameter-efficient transfer learning for NLP. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 2790–2799.

[29] Liu YH, Ott M, Goyal N, Du JF, Joshi M, Chen DQ, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.

[30] Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, Chen DQ, Yih WT. Dense passage retrieval for open-domain question

answering. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 6769–6781. [doi: 10.18653/v1/2020.emnlp-main.550]

[31] Hu YT, Lei ZH, Zhang Z, Pan B, Ling C, Zhao L. GRAG: Graph retrieval-augmented generation. In: Findings of the Association for Computational Linguistics: ACL 2025. Albuquerque: ACL, 2025. 4145–4157. [doi: 10.18653/v1/2025.findings-naacl.232]

[32] Edge D, Trinh H, Cheng N, Bradley J, Chao A, Mody A, Truitt S, Metropolitansky D, Ness RO, Larson J. From local to global: A graph RAG approach to query-focused summarization. arXiv:2404.16130, 2024.

[33] Zhou BY, Pan J, Gao F, Shen SJ. RAPTOR: Robust and perception-aware trajectory replanning for quadrotor fast flight. IEEE Trans. on Robotics, 2021, 37(6): 1992–2009. [doi: 10.1109/TRO.2021.3071527]

[34] Yang DJ, Rao JM, Chen KZ, Guo XY, Zhang YW, Yang J, Zhang Y. IM-RAG: Multi-round retrieval-augmented generation through learning inner monologues. In: Proc. of the 47th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Washington: ACM, 2024. 730–740. [doi: 10.1145/3626772.3657760]

[35] Wang ZH, Liu AJ, Lin HW, Li JQ, Ma XJ, Liang YT. RAT: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. arXiv:2403.05313, 2024.

[36] Asai A, Wu ZQ, Wang YZ, Sil A, Hajishirzi H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net. 2024.

[37] Gutiérrez BJ, Shu YH, Gu Y, Yasunaga M, Su Y. HippoRAG: Neurobiologically inspired long-term memory for large language models. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 1902.

[38] Chen WH, Hu HX, Chen X, Verga P, Cohen W. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 5558–5570. [doi: 10.18653/v1/2022.emnlp-main.375]

[39] Besta M, Kubicek A, Gerstenberger G, Chrapek M, Niggli R, Okanovic P, Zhu Y, Iff P, Podstawski M, Weitzendorf L, Chi MY, Gajda J, Nyczyk P, Müller J, Niewiadomski H, Hoefler T. Multi-head RAG: Solving multi-aspect problems with LLMs. arXiv:2406.05085, 2024.

[40] Chang RC, Zhang JW. CommunityKG-RAG: Leveraging community structures in knowledge graphs for advanced retrieval-augmented generation in fact-checking. arXiv:2408.08535, 2024.

[41] Li SL, He YC, Guo HY, Bu XY, Bai G, Liu J, Liu JH, Qu QW, Li YG, Ouyang WL, Su WB, Zheng B. GraphReader: Building graph-based agent to enhance long-context abilities of large language models. In: Findings of the Association for Computational Linguistics: ACL 2024. Miami: ACL 2024. 12758–12786. [doi: 10.18653/v1/2024.findings-emnlp.746]

[42] Fatehkia M, Lucas JK, Chawla S. T-RAG: Lessons from the LLM trenches. arXiv:2402.07483, 2024.

[43] Wu J, Zhang S, Che F, Feng MK, Shao PP, Tao JH. Pandora's Box or Aladdin's Lamp: A comprehensive analysis revealing the role of rag noise in large language models. In: Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vienna: ACL, 2025. 5019–5039. [doi: 10.18653/v1/2025.acl-long.250]

[44] Wei J, Wang XZ, Schuurmans D, Bosma M, Ichter B, Xia F, Chi H, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1800.

[45] Pfeiffer J, Kamath A, Rücklé A, Cho K, Gurevych I. AdapterFusion: Non-destructive task composition for transfer learning. In: Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics. ACL, 2021. 487–503. [doi: 10.18653/v1/2021.eacl-main.39]

[46] Hu EJ, Shen YL, Wallis P, Allen-Zhu Z, Li YZ, Wang SA, Wang L, Chen WZ. LoRA: Low-rank adaptation of large language models. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[47] Lialin V, Deshpande V, Yao XW, Rumshisky A. Scaling down to scale up: A guide to parameter-efficient fine-tuning. arXiv:2303.15647, 2023.

[48] Renze M. The effect of sampling temperature on problem solving in large language models. In: Findings of the Association for Computational Linguistics: ACL 2024. Miami: ACL, 2024. 7346–7356. [doi: 10.18653/v1/2024.findings-emnlp.432]

[49] Holtzman A, Buys J, Du L, Forbes M, Choi Y. The curious case of neural text degeneration. In: Proc. of the 2020 Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[50] Sun JS, Xu CJ, Tang LMY, Wang SZ, Lin C, Gong YY, Ni L, Shum HY, Guo J. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2024.

[51] Yue SB, Chen W, Wang SY, Li BX, Shen CC, Liu SJ, Zhou YX, Xiao Y, Yun S, Huang XJ, Wei ZY. DISC-LawLLM: Fine-tuning

large language models for intelligent legal services. arXiv:2309.11325, 2023.

[52] Févry T, Soares LB, FitzGerald N, Choi E, Kwiatkowski T. Entities as experts: Sparse memory access with entity supervision. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 4937–4951. [doi: 10.18653/v1/2020.emnlp-main.400]

[53] Guu K, Lee K, Tung Z, Pasupat P, Chang MW. REALM: Retrieval-augmented language model pre-training. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 368.

[54] Izacard G, Lewis P, Lomeli M, Hosseini H, Petroni F, Schick T, Dwivedi-Yu J, Joulin A, Riedel S, Grave E. Atlas: Few-shot learning with retrieval augmented language models. The Journal of Machine Learning Research, 2023, 24(1): 251.

[55] Wang Y, Li P, Sun MS, Liu Y. Self-knowledge guided retrieval augmentation for large language models. In: Findings of the Association for Computational Linguistics: ACL 2023. Singapore: ACL, 2023. 10303–10315. [doi: 10.18653/v1/2023.findings-emnlp.691]

[56] Baek J, Aji A F, Saffari A. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In: Proc. of the 1st Workshop on Natural Language Reasoning and Structured Explanations. Toronto: ACL, 2023. 78–106. [doi: 10.18653/v1/2023.nlrse-1.7]

[57] Limkonchotiwat P, Ponwitayarat W, Udomcharoenchaikit C, Chuangsuwanich E, Nutanong S. CL-ReLKT: Cross-lingual language knowledge transfer for multilingual retrieval question answering. In: Findings of the Association for Computational Linguistics: ACL 2022. Seattle: ACL, 2022. 2141–2155. [doi: 10.18653/v1/2022.findings-naacl.165]

[58] Liu Y, Wan Y, He LF, Peng H, Yu PS. KG-BART: Knowledge graph-augmented BART for generative commonsense reasoning. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 6418–6425. [doi: 10.1609/aaai.v35i7.16796]

[59] Zhang HY, Liu ZH, Xiong CY, Liu ZY. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 2031–2043. [doi: 10.18653/v1/2020.acl-main.184]

[60] Cai D, Wang Y, Bi W, Tu ZP, Liu XJ, Lam W, Shi SM. Skeleton-to-response: Dialogue generation guided by retrieval memory. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 1219–1228. [doi: 10.18653/v1/N19-1124]

[61] Shuster K, Xu J, Komeili M, Ju D, Smith EM, Roller S, Ung M, Chen MY, Arora K, Lane J, Behrooz M, Ngan W, Poff S, Goyal N, Szlam A, Boureau YL, Kambadur M, Weston J. BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage. arXiv:2208.03188, 2022.

[62] Li WT, Li JK, Ma WZ, Liu Y. Citation-enhanced generation for LLM-based chatbots. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: ACL, 2024. 1451–1466. [doi: 10.18653/v1/2024.acl-long.79]

[63] Dhole K. KAUCUS—Knowledgeable user simulators for training large language models. In: Proc. of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024). St. Julians: ACL, 2024. 53–65. [doi: 10.18653/v1/2024.scichat-1.5]

[64] Bertsch A, Alon U, Neubig G, Gormley MR. Unlimiformer: Long-range Transformers with unlimited length input. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 1543.

[65] Fan RZ, Fan YX, Chen JG, Guo JF, Zhang RQ, Cheng XQ. RIGHT: Retrieval-augmented generation for mainstream hashtag recommendation. In: Proc. of the 46th European Conf. on Information Retrieval. Glasgow: Springer, 2024. 39–55. [doi: 10.1007/978-3-031-56027-9_3]

[66] Wang Z, Teo S, Ouyang JE, Xu YJ, Shi W. M-RAG: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: ACL, 2024. 1966–1978. [doi: 10.18653/v1/2024.acl-long.108]

[67] Parvez R, Ahmad W, Chakraborty S, Ray B, Chang KW. Retrieval augmented code generation and summarization. In: Findings of the Association for Computational Linguistics: ACL 2021. Punta Cana: ACL, 2021. 2719–2734. [doi: 10.18653/v1/2021.findings-emnlp.232]

[68] Wang Y, Le H, Gotmare A, Bui N, Li JN, Hoi S. CodeT5+: Open code large language models for code understanding and generation. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 1069–1088. [doi: 10.18653/v1/2023.emnlp-main.68]

[69] Ding YRB, Wang ZJ, Ahmad W, Ramanathan MK, Nallapati R, Bhatia P, Roth D, Xiang B. CoCoMIC: Code completion by jointly modeling in-file and cross-file context. In: Proc. of the 2024 Joint Int'l Conf. on Computational Linguistics, Language Resources and Evaluation. Torino: ELRA and ICCL, 2024. 3433–3445.

[70] Liang M, Xie XH, Zhang GH, Zheng XJ, Di P, Jiang W, Chen HW, Wang CP, Fan G. REPOFUSE: Repository-level code completion with fused dual context. arXiv:2402.14323, 2024.

[71] Shrivastava D, Kocetkov D, de Vries H, Bahdanau D, Scholak T. RepoFusion: Training code models to understand your repository. In:

Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2024.

[72] Hayati SA, Olivier R, Avvaru P, Yin PC, Tomasic A, Neubig G. Retrieval-based neural code generation. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 925–930. [doi: 10.18653/v1/D18-1111]

[73] Zhou SY, Alon U, Xu FF, Jiang ZB, Neubig G. DocPrompting: Generating code by retrieving the docs. In: Proc. of the 11th Int'l Conf. on Learning Representations. OpenReview.net, 2023.

[74] Li J, Zhao YF, Li YM, Li G, Jin Z. AceCoder: Utilizing existing code to enhance code generation. arXiv:2303.17780, 2023.

[75] Zhang XY, Zhou Y, Yang G, Chen TL. Syntax-aware retrieval augmented code generation. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: ACL, 2023. 1291–1302. [doi: 10.18653/v1/2023.findings-emnlp.90]

[76] Liao DS, Pan SD, Sun XY, Ren XX, Huang Q, Xing ZC, Jin H, Li QY. A^3-CodGen: A repository-level code generation framework for code reuse with local-aware, global-aware, and third-party-library-aware. IEEE Trans. on Software Engineering, 2024, 50(12): 3369–3384. [doi: 10.1109/TSE.2024.3486195]

[77] Li J, Li YM, Li G, Jin Z, Hao YY, Hu X. SkCoder: A sketch-based approach for automatic code generation. In: Proc. of the 45th IEEE/ACM Int'l Conf. on Software Engineering (ICSE). Melbourne: IEEE, 2023. 2124–2135. [doi: 10.1109/ICSE48619.2023.00179]

[78] Liu MW, Yang TY, Lou YL, Du XY, Wang Y, Peng X. CodeGen4Libs: A two-stage approach for library-oriented code generation. In: Proc. of the 38th IEEE/ACM Int'l Conf. on Automated Software Engineering (ASE). Luxembourg: IEEE, 2023. 434–445. [doi: 10.1109/ASE56229.2023.00159]

[79] Su HJ, Jiang SY, Lai YH, Wu HY, Shi BA, Liu C, Liu Q, Yu T. EvoR: Evolving retrieval for code generation. In: Findings of the Association for Computational Linguistics. Miami: ACL, 2024. 2538–2554. [doi: 10.18653/v1/2024.findings-emnlp.143]

[80] Lu S, Duan N, Han H, Guo DY, Hwang SW, Svyatkovskiy A. ReACC: A retrieval-augmented code completion framework. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022. 6227–6240. [doi: 10.18653/v1/2022.acl-long.431]

[81] Zhang FJ, Chen B, Zhang Y, Keung J, Liu J, Zan DG, Mao Y, Lou JG, Chen WZ. RepoCoder: Repository-level code completion through iterative retrieval and generation. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 2471–2484. [doi: 10.18653/v1/2023.emnlp-main.151]

[82] Eghbali A, Pradel M. De-Hallucinator: Mitigating LLM hallucinations in code generation tasks via iterative grounding. arXiv:2401.01701, 2024.

[83] Li JA, Li YM, Li G, Hu X, Xia X, Jin Z. EditSum: A retrieve-and-edit framework for source code summarization. In: Proc. of the 36th IEEE/ACM Int'l Conf. on Automated Software Engineering (ASE). Melbourne: IEEE, 2021. 155–166. [doi: 10.1109/ASE51524.2021.9678724]

[84] Wang HY, Xia X, Lo D, He Q, Wang XY, Grundy J. Context-aware retrieval-based deep commit message generation. ACM Trans. on Software Engineering and Methodology (TOSEM), 2021, 30(4): 56. [doi: 10.1145/3464689]

[85] Shi ES, Wang YL, Tao W, Du L, Zhang HY, Han S, Zhang DM, Sun HB. RACE: Retrieval-augmented commit message generation. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 5520–5530. [doi: 10.18653/v1/2022.emnlp-main.372]

[86] Yu C, Yang G, Chen X, Liu K, Zhou YL. BashExplainer: Retrieval-augmented bash code comment generation based on fine-tuned CodeBERT. In: Proc. of the 2022 IEEE Int'l Conf. on Software Maintenance and Evolution (ICSME). Limassol: IEEE, 2022. 82–93. [doi: 10.1109/ICSME55016.2022.00016]

[87] Gao MH, Li JC, Fei H, Pang L, Ji W, Wang GM, Lv ZQ, Zhang WQ, Tang SL, Zhuang YT. De-fine: Decomposing and refining visual programs with auto-feedback. In: Proc. of the 32nd ACM Int'l Conf. on Multimedia. Melbourne: ACM, 2024. 7649–7657. [doi: 10.1145/3664647.3681082]

[88] Guo YC, Li ZX, Jin XL, Liu YT, Zeng YT, Liu WX, Li X, Yang P, Bai L, Guo JF, Cheng XQ. Retrieval-augmented code generation for universal information extraction. In: Proc. of the 13th Int'l CCF Conf. on Natural Language Processing and Chinese Computing. Hangzhou: Springer, 2024. 30–42. [doi: 10.1007/978-981-97-9434-8_3]

[89] Chen WH, Hu HX, Saharia C, Cohen WW. Re-Imagen: Retrieval-augmented text-to-image generator. In: Proc. of the 11th Int'l Conf. on Learning Representations. OpenReview.net, 2023.

[90] Sarto S, Cornia M, Baraldi L, Cucchiara R. Retrieval-augmented Transformer for image captioning. In: Proc. of the 19th Int'l Conf. on Content-based Multimedia Indexing. New York: ACM, 2022. [doi: 10.1145/3549555.354958]

[91] Tseng HY, Lee HY, Jiang L, Yang MH, Yang WL. RetrieveGAN: Image synthesis via differentiable patch retrieval. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 242–257. [doi: 10.1007/978-3-030-58598-3_15]

[92] Sheynin S, Ashual O, Polyak A, Singer U, Gafni O, Nachmani E, Taigman Y. KNN-Diffusion: Image generation via large-scale

retrieval. In: Proc. of the 11th Int'l Conf. on Learning Representations. OpenReview.net, 2023.

[93] Hu ZN, Iscen A, Sun C, Wang ZR, Chang KW, Sun YZ, Schmid C, Ross DA, Fathi A. REVEAL: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 23369–23379. [doi: 10.1109/CVPR52729.2023.02238]

[94] Ramos R, Martins B, Elliott D, Kementchedjhieva Y. SmallCap: Lightweight image captioning prompted with retrieval augmentation. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 2840–2849. [doi: 10.1109/CVPR52729.2023.00278]

[95] Lin WZ, Byrne B. Retrieval augmented visual question answering with outside knowledge. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 11238–11254. [doi: 10.18653/v1/2022.emnlp-main.772]

[96] Fang QK, Feng Y. Neural machine translation with phrase-level universal visual representations. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Dublin: ACL, 2022. 5687–5698. [doi: 10.18653/v1/2022.acl-long.390]

[97] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.

[98] Yang B, Cao M, Zou YX. Concept-aware video captioning: Describing videos with effective prior information. IEEE Trans. on Image Processing, 2023, 32: 5366–5378. [doi: 10.1109/TIP.2023.3307969]

[99] Xu JL, Huang YF, Hou JL, Chen G, Zhang YJ, Feng R, Xie WD. Retrieval-augmented egocentric video captioning. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 13525–13536. [doi: 10.1109/CVPR52733.2024.01284]

[100] Yin CX, Tang J, Xu ZY, Wang YZ. Memory augmented deep recurrent neural network for video question answering. IEEE Trans. on Neural Networks and Learning Systems, 2020, 31(9): 3159–3167. [doi: 10.1109/tnnls.2019.2938015]

[101] Lei J, Yu YC, Berg T, Bansal B. 2020. TVQA+: Spatio-temporal grounding for video question answering. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 8211–8225. [doi: 10.18653/v1/2020.acl-main.730]

[102] Le H, Chen N, Hoi S. VGNMN: Video-grounded neural module networks for video-grounded dialogue systems. In: Proc. of the 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: ACL, 2022. 3377–3393. [doi: 10.18653/v1/2022.naacl-main.247]

[103] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 2009, 3(4): 333–389. [doi: 10.1561/1500000019]

[104] Wang WH, Wei FR, Dong L, Bao HB, Yang N, Zhou M. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained Transformers. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 485.

[105] Chen JL, Xiao ST, Zhang PT, Luo K, Lian DF, Liu Z. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In: Findings of the Association for Computational Linguistics: ACL 2024. Bangkok: ACL, 2024. 2318–2335. [doi: 10.18653/v1/2024.findings-acl.137]

[106] Khattab O, Zaharia M. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In: Proc. of the 43rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2020. 39–48. [doi: 10.1145/3397271.3401075]

[107] Lin WZ, Mei JB, Chen JH, Byrne B. PreFLMR: Scaling up fine-grained late-interaction multi-modal retrievers. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: ACL, 2024. 5294–5316. [doi: 10.18653/v1/2024.acl-long.289]

[108] Zhang T, Patil SG, Jain N, Shen S, Zaharia M, Stoica I, Gonzalez JE. RAFT: Adapting language model to domain specific RAG. arXiv:2403.10131, 2024.

[109] Yan SQ, Gu JC, Zhu Y, Ling ZH. Corrective retrieval augmented generation. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net. 2024.

[110] Yan YB, Xie WD. EchoSight: Advancing visual-language models with wiki knowledge. In: Findings of the Association for Computational Linguistics: ACL 2024. Miami: ACL, 2024. 1538–1551. [doi: 10.18653/v1/2024.findings-emnlp.83]

[111] Shohan FT, Nayeem MT, Islam S, Akash AU, Joty S. XL-HeadTags: Leveraging multimodal retrieval augmentation for the multilingual generation of news headlines and tags. In: Findings of the Association for Computational Linguistics: ACL 2024. Bangkok: ACL, 2024. 12991–13024. [doi: 10.18653/v1/2024.findings-acl.771]

[112] Saito K, Sohn K, Zhang X, Li CL, Lee CY, Saenko K, Pfister T. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 19305–19314. [doi: 10.1109/CVPR52729.2023.01850]

[113] Arefeen A, Debnath B, Uddin YS, Chakradhar S. iRAG: Advancing RAG for videos with an incremental approach. In: Proc. of the 33rd ACM Int'l Conf. on Information and Knowledge Management. Boise: ACM, 2024. 4341–4348. [doi: 10.1145/3627673.3680088]

[114] Jin XJ, Zhang BW, Gong WB, Xu K, Deng XQ, Wang P, Zhang Z, Shen XH, Feng JS. MV-Adapter: Multimodal video transfer learning for video text retrieval. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 27134–27143. [doi: 10.1109/CVPR52733.2024.02563]

[115] Luo YD, Zheng XW, Li GL, Yin SK, Lin HJ, Fu CY, Huang JF, Ji JY, Chao F, Luo JB, Ji RR. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. arXiv:2411.13093, 2024.

[116] Wang JM, Wang PC, Sun GH, Liu DF, Dianat S, Rao R, Rabbani M, Tao ZQ. Text is MASS: Modeling as stochastic embedding for text-video retrieval. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 16551–16560. [doi: 10.1109/CVPR52733.2024.01566]

[117] Ren XB, Xu LR, Xia L, Wang SQ, Yin DW, Huang C. VideoRAG: Retrieval-augmented generation with extreme long-context videos. In: Proc. of the 32nd ACM SIGKDD Conf. on Knowledge Discovery and Data Mining V.1. Jeju: ACM, 2026. 2390–2401. [doi: 10.1145/3770854.3783944]

[118] Shen XB, Huang QX, Lan L, Zheng YH. Contrastive Transformer cross-modal hashing for video-text retrieval. In: Proc. of the 33rd Int'l Joint Conf. on Artificial Intelligence. Int'l Joint Conf. on Artificial Intelligence Organization. Jeju: ACM, 2024. 136. [doi: 10.24963/ijcai.2024/136]

[119] Du Y, Liu YQ, Jin Q. Reversed in time: A novel temporal-emphasized benchmark for cross-modal video-text retrieval. In: Proc. of the 32nd ACM Int'l Conf. on Multimedia. Melbourne: ACM, 2024. 5260–5269. [doi: 10.1145/3664647.3680731]

[120] Zhang L, Zhao TC, Ying HT, Ma YB, Lee K. OmAgent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing. Miami: ACL, 2024. 10031–10045. [doi: 10.18653/v1/2024.emnlp-main.559]

[121] Ma ZY, Gou CH, Shi HC, Sun B, Li ST, Rezatofighi H, Cai JF. DrVideo: Document retrieval based long video understanding. In: Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025. 18936–18946.

[122] Kwiatkowski T, Palomaki J, Redfield O, Collins M, Parikh A, Alberti C, Epstein D, Polosukhin I, Devlin J, Lee K, Toutanova K, Jones L, Kelcey M, Chang MW, Dai AM, Uszkoreit J, Le Q, Petrov S. Natural Questions: A benchmark for question answering research. Trans. of the Association for Computational Linguistics, 2019, 7: 452–466. [doi: 10.1162/tacl_a_00276]

[123] Joshi M, Choi E, Weld D, Zettlemoyer L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vancouver: ACL, 2017. 1601–1611. [doi: 10.18653/v1/P17-1147]

[124] Bajaj P, Campos D, Craswell N, Deng L, Gao JF, Liu XD, Majumder R, McNamara A, Mitra B, Nguyen T, Rosenberg M, Song X, Stoica A, Tiwary S, Wang T. MS MARCO: A human generated machine reading comprehension dataset. arXiv:1611.09268, 2016.

[125] Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: A large-scale dataset for fact extraction and verification. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long Papers). New Orleans: ACL, 2018. 809–819. [doi: 10.18653/v1/N18-1074]

[126] Min S, Michael J, Hajishirzi H, Zettlemoyer L. AmbigQA: Answering ambiguous open-domain questions. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2020. 5783–5797. [doi: 10.18653/v1/2020.emnlp-main.466]

[127] Xu L, Hu H, Zhang XW, *et al*. CLUE: A Chinese language understanding evaluation benchmark. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 4762–4772. [doi: 10.18653/v1/2020.coling-main.419]

[128] Yang ZL, Qi P, Zhang SZ, Bengio Y, Cohen W, Salakhutdinov R, Manning CD. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 2369–2380. [doi: 10.18653/v1/D18-1259]

[129] Ho X, Nguyen AKD, Sugawara S, Aizawa A. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 6609–6625. [doi: 10.18653/v1/2020.coling-main.580]

[130] Qiu YF, Li HY, Qu YQ, Chen Y, She QQ, Liu J, Wu H, Wang HF. DuReader-Retrieval: A large-scale chinese benchmark for passage retrieval from Web search engine. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL,

2022. 5326–5338. [doi: 10.18653/v1/2022.emnlp-main.357]

[131] Izacard G, Grave E. Leveraging passage retrieval with generative models for open domain question answering. In: Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics: Main Volume. ACL, 2021. 874–880. [doi: 10.18653/v1/2021.eacl-main.74]

[132] Shi WJ, Min S, Yasunaga M, Seo M, James R, Lewis M, Zettlemoyer L, Yih WT. REPLUG: Retrieval-augmented black-box language models. In: Proc. of the 2024 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long Papers). Mexico City: ACL, 2024. 8371–8384. [doi: 10.18653/v1/2024.naacl-long.463]

[133] Berant J, Chou A, Frostig R, Liang P. Semantic parsing on freebase from question-answer pairs. In: Proc. of the 2013 conf. on empirical methods in natural language processing. Seattle: ACL, 2013. 1533–1544.

[134] Stelmakh I, Luan Y, Dhingra B, Chang MW. ASQA: Factoid questions meet long-form answers. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 8273–8288. [doi: 10.18653/v1/2022.emnlp-main.566]

[135] Shao ZH, Gong YY, Shen YL, Huang ML, Duan N, Chen WZ. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: ACL, 2023. 9248–9274. [doi: 10.18653/v1/2023.findings-emnlp.620]

[136] Wang HY, Li RR, Jiang HM, Tian JJ, Wang ZY, Luo C, Tang XF, Cheng MX, Zhao T, Gao J. BlendFilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing. Miami: ACL, 2024. 1009–1025. [doi: 10.18653/v1/2024.emnlp-main.58]

[137] Yao SY, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, Cao Y. ReAct: Synergizing reasoning and acting in language models. In: Proc. of the 11th Int'l Conf. on Learning Representations. OpenReview.net, 2023.

[138] Press O, Zhang M, Min S, Schmidt L, Smith N, Lewis M. Measuring and narrowing the compositionality gap in language models. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: ACL, 2023. 5687–5711. [doi: 10.18653/v1/2023.findings-emnlp.378]

[139] Khandelwal U, Levy O, Jurafsky D, Zettlemoyer L, Lewis M. Generalization through memorization: Nearest neighbor language models. In: Proc. of the 2020 Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[140] Borgeaud S, Mensch A, Hoffmann J, *et al*. Improving language models by retrieving from trillions of tokens. In: Proc. of the 39th Int'l Conf. on Machine Learning. Baltimore: PMLR, 2022. 2206–2240.

[141] White J. PubMed 2. 0. Medical Reference Services Quarterly, 2020, 39(4): 382–387. [doi: 10.1080/02763869.2020.1826228]

[142] Bollerslev T, Engle RF, Nelson DB. ARCH models. Handbook of Econometrics, 1994, 4: 2959–3038. [doi: 10.1016/S1573-4412(05)80018-2]

[143] Wang CT, Li MD, Sun MX, Wang J, Yang BX, Niu JH, He ZG, Zhang WS. Cross-language bilayer knowledge graph with large-scale medical facts. Ruan Jian Xue Bao/Journal of Software, 2025, 36(3): 1240–1253 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7173.htm [doi: 10.13328/j.cnki.jos.007173]

[144] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The muppets straight out of law school. In: Findings of the Association for Computational Linguistics: ACL 2020. ACL, 2020. 2898–2904. [doi: 10.18653/v1/2020.findings-emnlp.261]

[145] Huang AH, Wang H, Yang Y. FinBERT: A large language model for extracting information from financial text. Contemporary Accounting Research, 2023, 40(2): 806–841. [doi: 10.1111/1911-3846.12832]

附中文参考文献

[22] 刘泽垣, 王鹏江, 宋晓斌, 张欣, 江奔奔. 大语言模型的幻觉问题研究综述. 软件学报, 2025, 36(3): 1152–1185. http://www.jos.org.cn/1000-9825/7242.htm [doi: 10.13328/j.cnki.jos.007242]

[143] 王楚童, 李明达, 孙孟轩, 王静, 杨雪冰, 牛景昊, 贺志阳, 张文生. 融合大规模医学事实的跨语言双层知识图谱. 软件学报, 2025, 36(3): 1240–1253. http://www.jos.org.cn/1000-9825/7173.htm [doi: 10.13328/j.cnki.jos.007173]

作者简介

刘澳迪, 硕士生, CCF 学生会员, 主要研究领域为自然语言处理.

奚雪峰, 博士, 教授, CCF 高级会员, 主要研究领域为自然语言处理, 多模态机器学习, 软件工程.

周国栋, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为自然语言理解, 中文计算, 信息抽取.