

从 RAG 到 SAGE: 现状与展望

田永林^{1,2} 王雨桐^{1,2} 王兴霞^{1,2,3} 杨静^{1,2,3} 沈甜雨⁴ 王建功⁵
范丽丽⁶ 郭超^{1,2} 王寿文⁷ 赵勇⁸ 武万森⁸ 王飞跃^{2,3,7}

摘要 大模型技术的兴起显著提升了人们获取和利用知识的效率,但在实际应用中仍然面临着知识受限、迁移障碍和幻觉等挑战,阻碍了可信可靠人工智能系统的构建.检索增强生成(RAG)通过利用外接知识库和查询关联的检索有效增强大模型的能力水平,为大模型掌握实时型、行业型及私有型知识提供有力支撑,进而促进大模型技术向多样场景的快速推广和实施.围绕 RAG,阐述其基本原理、发展现状及典型应用,并分析其优势和面临的挑战.在 RAG 的基础上,通过结合搜索模块和多级缓存管理模块,提出 RAG 的拓展框架 SAGE,以建立更加灵活和高效的大模型知识外挂工具链.

关键词 大模型,检索增强生成,基础智能,知识自动化

引用格式 田永林,王雨桐,王兴霞,杨静,沈甜雨,王建功,范丽丽,郭超,王寿文,赵勇,武万森,王飞跃.从 RAG 到 SAGE:现状与展望.自动化学报,2025,51(6):1145-1169

DOI 10.16383/j.aas.c240163 **CSTR** 32138.14.j.aas.c240163

From Retrieval-augmented Generation to SAGE: The State of the Art and Prospects

TIAN Yong-Lin^{1,2} WANG Yu-Tong^{1,2} WANG Xing-Xia^{1,2,3} YANG Jing^{1,2,3} SHEN Tian-Yu⁴
WANG Jian-Gong⁵ FAN Li-Li⁶ GUO Chao^{1,2} WANG Shou-Wen⁷
ZHAO Yong⁸ WU Wan-Sen⁸ WANG Fei-Yue^{2,3,7}

Abstract The emergence of large model technologies has significantly enhanced the efficiency with which humans acquire and utilize knowledge. However, in practical applications, they still confront challenges such as constrained knowledge, transfer obstacles, and hallucinations, which impede the construction of trustworthy and reliable artificial intelligence systems. Retrieval-augmented generation (RAG), by leveraging external knowledge bases and query-related retrieval, has effectively strengthened capability of large models and offers strong support for large models to master real-time, industry-specific, and private knowledge, thereby facilitating the rapid promotion and implementation of large model technologies across diverse scenarios. This paper focuses on RAG, detailing its basic principles, current development status, as well as exemplary applications, and analyzing its advantages and the challenges it faces. Based on RAG, we propose the extended framework of search-augmented generation and extension by incorporating the search module and multi-level cache management module, aiming to create a more flexible and efficient knowledge toolchain for large models.

Key words Large model, retrieval-augmented generation, foundation intelligence, knowledge automation

Citation Tian Yong-Lin, Wang Yu-Tong, Wang Xing-Xia, Yang Jing, Shen Tian-Yu, Wang Jian-Gong, Fan Li-Li, Guo Chao, Wang Shou-Wen, Zhao Yong, Wu Wan-Sen, Wang Fei-Yue. From retrieval-augmented generation to SAGE: The state of the art and prospects. *Acta Automatica Sinica*, 2025, 51(6): 1145-1169

收稿日期 2024-03-29 录用日期 2024-10-05

Manuscript received March 29, 2024; accepted October 5, 2024
国家自然科学基金青年基金(62303460),澳门特别行政区科学技术发展基金(0145/2023/RIA3),中国科协青年人才托举工程(YESS20220372)资助

Supported by National Natural Science Foundation of China (62303460), Science and Technology Development Fund of Macau SAR (0145/2023/RIA3), and Young Elite Scientists Sponsorship Program of China Association of Science and Technology (YESS20220372)

本文责任编辑 赫然

Recommended by Associate Editor HE Ran

1. 中国科学院自动化研究所多模态人工智能系统全国重点实验室 北京 100190 2. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室 北京 100190 3. 中国科学院大学人工智能学院 北京 100049 4. 北京化工大学信息科学与技术学院 北京 100029 5. 中国航空系统工程研究所 北京 100029 6. 北京理工

大学信息与电子学院 北京 100081 7. 澳门科技大学创新工程学院 澳门 999078 8. 国防科技大学系统工程学院 长沙 410073

1. State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 3. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049 4. College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029 5. Aviation System Engineering Institute of China, Beijing 100029 6. School of Information and Electronics, Beijing Institute of Technology, Beijing 100081 7. Faculty of Innovation Engineering, Macau University of Science and Technology, Macau 999078 8. College of Systems Engineering, National University of Defense Technology, Changsha 410073

在 Transformer^[1]、深度强化学习 (Deep reinforcement learning)^[2]、扩散模型 (Diffusion model)^[3] 等算法技术、分布式训练等大算力技术以及规模法则 (Scaling law^[4]) 驱动的大数据技术的协同促进下, 大语言模型 (Large language model, LLM) 及基础模型 (Foundation models) 已经在内容创作^[5]、自动驾驶^[6]、机器人^[7]、医疗^[8] 等领域展现出广阔的应用潜力^[9]。大模型所具备的高理解力和强泛化性极大地促进生产方式的变革, 以大模型为基础的科研和产业创新生态得到密切关注。然而, 随着通用大模型在实际场景的应用加深, 其不足之处也逐渐显现, 具体而言, 主要包括以下几点:

1) 知识受限。大模型的能力受限于其训练过程使用的数据集, 在训练完成后知识将固定下来, 无法对实时知识和新专业领域知识进行有效掌握。例如, 基于 GPT3.5 的 ChatGPT 的知识更新截止至 2021 年 9 月^[10], 虽然可以通过后期升级更新知识库 (如 GPT4 Turbo 的知识库更新至 2023 年 4 月^[11]), 但即便如此, 仍然难以满足用户对最新知识获取的需求^[3, 12]。此外, 诸如 ChatGPT 之类的通用大模型对垂直场景和私有领域知识的掌握能力不足, 导致其难以直接在专业应用和个性化任务中取得理想效果^[13]。

2) 迁移障碍。通用大模型向专用领域迁移时面临成本和技术障碍^[14], 并存在灾难性遗忘等问题^[15]。通用大模型可以借助微调和提示工程实现能力迁移, 但前者需要使用者构建数据和算力平台, 从而带来一定的技术和资源障碍, 而且存在较大的隐私泄露风险; 后者虽然可以通过零样本学习的方式实现能力快速迁移, 但存在着上下文窗口长度受限的问题。虽然目前部分模型已经支持数十甚至上百万 token 的超长上下文, 却仍然难以满足大规模的信息注入需求, 而且在面对“大海捞针” (Needle in the haystack¹) 的压力测试任务时, 基于长上下文的大模型在提取相关信息方面也面临着准确性和效率的挑战^[16]。

3) 幻觉。大模型产生与事实、用户指令、上下文等不一致的答案的现象称为大模型幻觉。这种情况时有发生, 并给大模型走向实际应用, 尤其是一些安全敏感场景, 带来较大阻碍^[17]。

检索增强生成 (Retrieval-augmented generation, RAG) 的提出为解决通用大模型在实际应用中存在的上述问题提供了思路^[18]。它通过接入外部知识库为大模型提供必要的信息支撑, 并借助信息检索技术获取与用户指令或查询相关的上下文信

息, 实现对原始查询及提示的扩充, 进而增强大模型的生成过程。RAG 技术显著改善大模型能力, 为推动大模型走向实际应用产生积极作用。为此, 本文从关键技术、典型应用、开源平台以及新架构四个主要方面对 RAG 技术进行全面综述。作为对比, 本文在表 1 中总结现有的 RAG 综述文章^[12, 19-23], 从中可以看出, 现有的综述文章虽然涵盖 RAG 技术中的检索和生成部分, 但对知识库构建方法的介绍仍然不足。此外, 当前综述侧重 RAG 在自然语言处理 (Natural language processing, NLP) 领域的应用介绍, 对计算机视觉 (Computation vision, CV) 以及垂直领域 (医学、法律、自动驾驶) 中的应用尚未进行充分讨论, 同时缺乏对开源 RAG 平台的总结。为此, 本文在上述方面进行针对性介绍, 以对现有文章形成补充。本文涵盖数据库构建、检索优化方法以及内容生成方法, 形成针对 RAG 技术的完整技术介绍。同时, 本文总结 RAG 在 NLP 和 CV 以及垂直领域 (医学、法律、自动驾驶) 中的应用, 并首次汇总热门 RAG 开源平台。此外, 本文给出用于提升检索效率和开放性的新型架构。本文组织结构如下: 第 1 节介绍基础型 RAG 系统 (Navie RAG^[20]) 的架构; 第 2 节对其关键技术进行分析和综述; 第 3 节介绍 RAG 在语言、视觉等通用领域以及医疗、自动驾驶等专用领域的应用; 第 4 节给出基于搜索和多级缓存架构的拓展式检索增强技术的架构; 第 5 节总结全文并对 RAG 优势和挑战进行展望。

1 基础型 RAG 的一般架构

基础型 RAG 的架构如图 1 所示, 包括三个部分: 知识层、检索层和生成层。知识层旨在将外部信息转换为向量型数据库, 从而给大模型提供必要的信息支撑。相比于通用型大模型在训练过程中已经使用的海量信息, RAG 额外引入的知识库主要包括三方面内容: 实时数据、行业数据以及私有数据, 分别用于提供对新知识、专业知识和私域知识的连接^[24]。向量知识库的构建过程包括数据分块和嵌入编码两个部分。在数据分块过程中, 以不同格式和不同模态存储的外部数据将被拆分成一定长度的数据段, 以便于后续的编码过程, 同时为检索过程提供精确且高相关性的信息片段。在此基础上, RAG 采用嵌入机制将数据段编码成向量形式。经过编码的数据段具有更为高级的语义表示, 能够提升检索的准确性和相关性。经过数据分块和嵌入编码之后, 可以对外部数据构建向量数据库。检索层主要通过

¹ https://github.com/gkamradt/LLMTest_NeedleInAHaystack

表 1 RAG 综述文章总结与对比
Table 1 Summary and comparison of surveys on RAG

文献	年份	RAG 技术点	RAG 应用领域	RAG 平台	新架构
文献 [19]	2024	检索、生成	NLP	×	×
文献 [12]	2024	架构、学习、检索	NLP 及下游应用	×	×
文献 [20]	2023	检索、生成、搜索	NLP	×	×
文献 [21]	2023	架构、检索、生成	NLP 及下游应用	×	Module RAG
文献 [22]	2022	检索、生成	NLP 及下游应用	×	×
文献 [23]	2024	检索、生成	NLP 及下游应用	×	×
本文	2024	知识库、检索、生成	NLP、CV、垂直应用	✓	SAGE

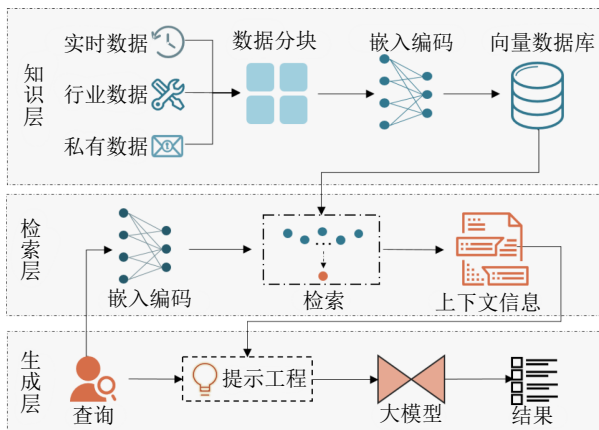


图 1 基础型 RAG 的框架

Fig. 1 The framework of naive RAG

对查询嵌入与向量库文本的相似性度量, 获取查询相关的上下文信息. 检索过程首先对输入查询进行嵌入编码以获取向量类型的查询表示; 其次, 通过语义空间的各类相似性计算方法, 从数据库中选择相似性最高的若干上下文, 并根据其索引得到原文信息.

RAG 与单纯基于大模型的生成式问答的不同之处在于 RAG 对提示工程部分进行增强. 具体而言, 检索得到的上下文信息, 将与原始查询信息一起嵌入提示模板, 进而送入大模型中. 提示工程的设计中通常包含对上下文信息的特别说明, 旨在促进大模型对上下文和原始查询的区分能力, 从而更好地理解任务需求.

2 RAG 关键技术

本节将从关键技术角度对进阶式 RAG 系统展开介绍, 讨论知识库构建技术、信息检索技术和内容生成等技术 (如图 2 所示) 及典型方法. 其中, 知识库构建和信息检索技术作为 RAG 系统最有代表性的环节, 提供 LLM 所需要的知识源以及与用户查询的交互. 为此, 本节将对其进行重点介绍.

2.1 知识库构建

知识库的构建在 RAG 任务中具有重要意义^[25]. 在对话式人工智能 (Artificial intelligence, AI) 中, 知识库使大模型有能力回答复杂的专业问题, 并增强大模型的连续对话能力以及根据用户需求进行答案修改的能力, 从而提升系统的智能性. 此外, 基于知识库的大模型问答系统还具有更强的可解释性与透明度. 大模型在生成回答的同时能够给出回答的依据, 从而帮助用户理解为何某个结果被返回, 实现系统可解释性和可信度的提升. 按照知识存储的特点, 可以将 RAG 系统中的知识库划分成三种主要类型: 实时型知识库、行业型知识库和私有型知识库.

1) 实时型知识库. 其突出特点是实时更新, 这使得大模型能够获取最新的信息, 从而帮助大模型理解快速变化的世界 (如新闻事件、科技发展等)^[26].

2) 行业型数据库. 其作用是为大模型提供专业级知识, 使大模型能更好地理解解析特定领域和任务的查询需求, 并利用知识库中存储的知识进行推理, 从而提供符合行业约束和任务需求的回答结果. 在跨领域应用中, 通过挂载不同领域的行业知识库可以以较小的代价增强通用大模型在医疗^[27]、法律^[28]、金融^[29]等垂直领域提供专业知识问答服务的能力.

3) 私有型知识库. 用来存储涉及个人或组织隐私的信息, 例如用户的历史行为和偏好等^[30]. 由于数据的敏感性, 这部分信息很难直接用于模型的预训练或微调过程. 通过基于利用私有知识库的 RAG 系统, 通用大模型可以提供个性化的建议和结果, 从而提升用户体验.

在基于 RAG 的大模型系统中, 可以使用单一知识库, 也可以使用多种不同类型的知识库以提供丰富多样的信息. 此外, 还可以通过对知识库进行多层级的组织来形成不同信息粒度的知识库 (例如摘要数据库和原始信息数据库), 从而为高效信息检索提供基础. 知识库构建主要包括基于向量数据

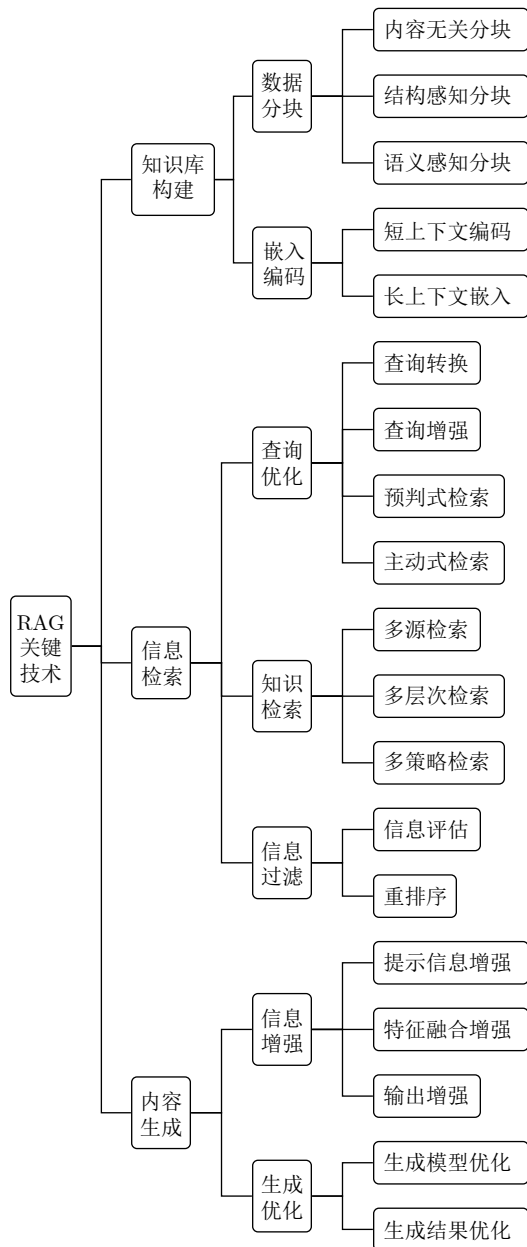


图 2 RAG 关键技术

Fig.2 Key technologies in RAG

库的方法和基于传统数据库的方法,考虑到向量数据库在复杂多元化检索需求中的语义感知精准性,本节主要总结基于向量库的方法,并对其中涉及的两个关键步骤:数据分块和嵌入编码进行介绍。基于传统数据库的方法将在第 2.1.3 节中给出简要总结。

2.1.1 数据分块

数据分块 (Data chunking) 的目的在于构建细粒度的上下文信息,以便提高检索的精准性,从而高效地进行大模型能力增强。虽然大模型可以支持的上下文窗口长度不断得到增长,尤其是谷歌发布

的 Gemini 1.5 目前已经实现支持 1 M token 的超长上下文语义理解^[31],但对于大多数应用场景而言,几句或者一段话表达的语义信息能更加明确且简洁地表达语义关联,从而提高内容生成的效率^[32]。数据分块的核心在于平衡语义的完整性与分块效率之间的矛盾。原始数据中通常存在复杂且多层次的语义关联,如何在分块长度和分块复杂度的限制下,实现对完整语义的挖掘是一项具有挑战性的工作。

具体而言,数据分块是指将原始数据按照一定的规则分割成若干个不同的数据块。在这一过程中,应尽量保证具有强相关性的语料不会被分割到不同的数据块中,而且每个数据块内部均具有相对完整的语义信息。数据分块的两个要素分别为块大小与语义关联程度。其中,块大小的上限主要取决于所使用嵌入模型的 token 容量。随着技术的发展,嵌入模型的 token 容量也有着不断提升的趋势,例如基于 BERT^[33] 的嵌入模型仅有 512 个 token 容量,而更新的 M3E^[34] 和 OpenAI-ADA-002^[35] 则分别提高到 768 和 1536,最新的大模型如 GPT4 Turbo 以及 Gemini 1.5 Pro 模型甚至达到 128 K 和 1 M 的上下文窗口规模。嵌入模型 token 容量的提升也使得语义信息分割受到的限制减少。在通过语义关联度索引相关资料时是以块为最小单位的,有效的分块策略能够将特定主题和相关性的信息分到同一块中,从而确保检索结果的准确性^[36]。分块的大小不合适将会导致索引结果不精确,从而使大模型获取到错误的提示资料^[37]。因此,在一定块大小的限制下,根据语义信息对数据进行最佳分块对于确保索引准确性和大模型回答可靠性至关重要。以此为目标,当前已经开发有多种相关的数据分块技术。按照对原始文档内容的感知程度,可以分为内容无关的分块方法和内容相关的分块方法(结构感知的分块方法、语义感知的分块方法)。

1) 内容无关分块。内容无关分块基于固定的分割策略处理数据,该方法简单高效,但存在分块语义提取不佳的问题。采用固定尺寸的分块策略是一种直接自然的方法,该方法以单一或多个事先确定的块尺寸对文档进行划分,而不考虑文档的具体内容^[38]。为减少分割导致的上下文语义信息丢失,通常会在相邻块之间设置一定的重叠值,以增强内容的连续性。该方法简单易用,是实践过程中常用的分块算法之一。

2) 结构感知分块。结构感知的方法通过制定一系列的文档分割符将位于不同结构段的文档块进行拆分^[39]。分割可以根据文档数据的类型进行设计,常见的分割符包括标点符号、空格、换行符、标题级

别符以及特殊符号(如代码中的方法符号和类名符号)等. 除利用分割符进行直接的文档划分外, 还可以将分割符组成递归式划分策略, 从而形成不同尺寸的分割块以灵活适应查询需求. 结构感知的分块方法可保障文档块在格式上的完整性, 但仍然缺乏对文档内容的语义理解.

3) 语义感知分块. 相比于结构感知, 语义感知进一步考虑文档的高层级信息, 并依据文档语义产生不同的分块结果. 基于聚类的分块策略是常用的方法, 其利用深度学习模型或其他算法提前提取文档的语义信息, 并将语义相似性较高的内容和信息聚合在一起, 从而增强最终检索的相关性和准确性^[40]. 该类方法减少分块导致的语义缺失和噪声引入, 但由于需要使用深度学习或其他自然语言处理算法进行语义提取, 因此过程相对繁琐. 一般而言, 对于给定文本 $T = \{s_1, s_2, \dots, s_n\}$, 其中, s_i 表示第 i 个句子. 基于语义感知的分块方法需要首先使用语言模型将每个句子转换为嵌入向量 e_i , 即

$$e_i = \text{embed}(s_i)$$

然后, 通过计算对应向量 e_i 和 e_{i+1} 的余弦相似度获得相邻的句子 s_i 和 s_{i+1} 之间的语义相似度. 其中, 余弦相似度计算式为

$$\text{cosine_sim}(e_i, e_{i+1}) = \frac{e_i \cdot e_{i+1}}{\|e_i\| \|e_{i+1}\|}$$

接下来, 通过动态规划进行分块, 定义代价函数 $C(i, j)$ 为句子 s_i 和 s_j 形成的一个块的代价, 即

$$C(i, j) = \sum_{k=i}^{j-1} (1 - \text{cosine_sim}(e_k, e_{k+1}))$$

设 $D(j)$ 表示从句子 s_i 到 s_j 的最小代价, 则有

$$D(j) = \min_{1 \leq i < j} \{D(i) + C(i, j)\}$$

随着大模型的应用, 以大模型作为语义感知器取得良好的效果, 成为当前的热点语义感知方法. 大模型方法能够支持自然类人化的文档理解, 因此也有望实现强语义关联的文档划分. 分块过程中, 可以将文档依次输入大模型, 并通过提示工程等方式促使大模型进行文档内容的理解和自动拆分^[41]. 大模型的优势一方面体现在能够建立对文档内容的深度理解以及灵活个性化的划分; 另一方面, 由于大模型出色的生成能力, 因此还可以对文档内容进行适当改写, 以在格式和表达上进行优化, 从而提升编码及检索过程的效率. 图 3 展示了 DenseX Retrieval 方法^[41]的原理图. 该方法使用语言大模型^[42]对段落级别的语料 (Passage-level corpus) 进行处理, 并根据用户的提示信息进行细粒度语料, 即命

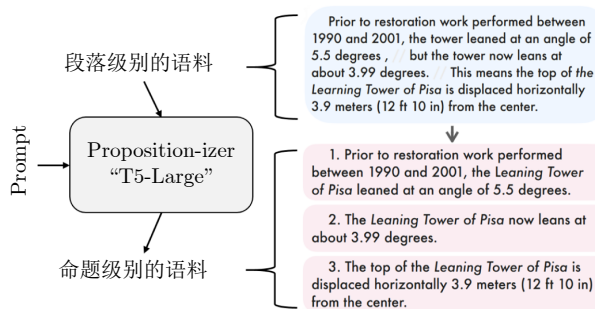


图 3 DenseX Retrieval 方法的框架

Fig. 3 The framework of DenseX Retrieval

题级别的语料 (Propositions-level corpus) 内容的生成. DenseX Retrieval 方法实现对语义信息的完整保留, 同时其细粒度语料的特点有效减少了噪声信息的引入.

4) 各类分块方法的对比分析. 内容无关的分块方法通常依赖于文本的结构或固定的规则对文本进行分块, 例如句子、段落或固定长度的文本块, 而不考虑文本的语义内容^[43-44], 这类方法简单直接, 无需额外的语义理解或模型, 但忽略文本的语义信息, 可能导致分块不连贯, 影响后续的信息检索和生成任务. 结构感知的分块方法考虑文本的结构信息, 例如标题、段落等, 以及一些固定的格式要求, 根据文本的标题、段落、列表等结构进行分块^[45-46], 生成的分块更符合文档的组织结构, 但仍然忽略文本的语义信息, 可能导致分块不够准确. 而语义感知的分块方法通过理解文本的语义信息来进行分块, 基于文本的语义相似性来确定最优的分块位置, 以保持语义的连贯性和完整性^[47-48], 这些方法考虑文本的语义信息, 生成的分块更加连贯和准确, 但也面临着计算量较大, 需要较多的计算资源, 可能会增加系统的复杂性的问题.

2.1.2 嵌入编码

传统的数据库不需要对其中存储的信息与知识进行语义层面的理解, 因此可以直接存储文本等类型的原始数据, 但在 RAG 中, 信息检索结果的语义关联程度将直接影响大模型的生成效果, 因此对文档中的语义信息理解提出较高要求. 为此, 通过嵌入编码对分块后的文档进行映射以形成高语义向量变得十分必要. 文档块的向量化为度量信息之间的语义相关性提供基础, 能够有效增强检索模型对近似语义和混淆语义的感知能力, 同时特征空间的引入有助于缩小跨域输入 (如不同语言或不同模态) 之间的差异, 进而提升信息检索的准确性. 在多模态 RAG 系统中, 理解各类模态数据的语义信息同样至关重要, 检索结果的语义关联性将直接影响生

成模型的效果. 接下来, 本文将介绍两种主要的文本嵌入编码方法: 基于短上下文的嵌入方法和基于长上下文的嵌入方法, 并探讨其他模态的数据嵌入技术.

1) 短上下文嵌入. 这类方法通常以字或词语作为最小处理单元, 通过对字词的嵌入进行组合得到句子或更长文本的表示. 自然语言处理中传统的词向量表示方法有 One-Hot 编码^[49]、词袋模型 (Bag of words, BOW)^[50]、TF-IDF 方法^[51]、 N -Gram^[52]等, 这些方法简单直观, 但通常无法捕捉词与词之间的复杂关系. 相比之下, 基于上下文的词嵌入模型, 例如 Word2Vec^[53] 和 GloVe^[54], 通过分析词在上下文中的分布来学习词向量, 能够更有效地捕捉词的语义和语法信息. 尽管这些模型在处理大规模数据集时表现出色, 但它们在长距离依赖关系建模上存在局限.

2) 长上下文嵌入. 随着 Transformer 和大模型技术的迅速发展, 词嵌入的方式逐步进入基于长上下文的编码时代. BERT^[33]、GPT 等嵌入模型成为文本向量的主流方法. 这类方法使用自注意力机制对输入的长序列进行端到端的文本向量化, 可以将任意文本映射为低维稠密的向量, 在语义信息提取、上下文关联分析等方面相比传统方法具有较大优势. 由于各类嵌入模型层出不穷, 为对比各类模型的优劣, 海量文本语义向量基准测试 (Massive text embedding benchmark, MTEB)^[55] 得到广泛使用, 它包含双语文本挖掘、分类、聚类、重排、检索等 8 个语义向量任务, 涵盖句子对句子 (S2S)、段落到段落 (P2P) 和句子到段落 (S2P) 等三类共 58 个数据集和 112 种语言^[55]. 在该基准的英文测试结果较好的嵌入模型主要有: SFR^[56]、GritLM^[57]、e5^[58]等, 中文测试结果较好的嵌入模型主要有: Baichuan^[59]、BGE^[60]、C-pack^[61]等. 在 RAG 中, 基于长上下文的嵌入方法通常采用预训练方法以建立通用可泛化的嵌入表示, 同时, 针对特定场景和任务, 为提升嵌入效果, 可以进一步通过微调方法对嵌入模型进行少量训练^[62], 以实现更符合实际需求的语义特征提取. 本节以 BGE-LE^[63] 为例, 介绍长上下文编码的案例. 如图 4 所示, BGE-LE 对输入的语料信息构建上下文窗口, 将长度为 l 的句子段输入 LLM (如 LLaMA^[64]) 中进行编码, 并基于标记词符 (Landmark token, LMK) 聚合句子段的信息, 形成对句子段的编码表示 LE (Landmark embedding). 其编码过程可表示为

$$LE_i \leftarrow LLM(C_{i-1}, \dots, C_i; LMK) \cdot embed[-1]$$

3) 其他模态的数据嵌入方法. 针对较为常见的

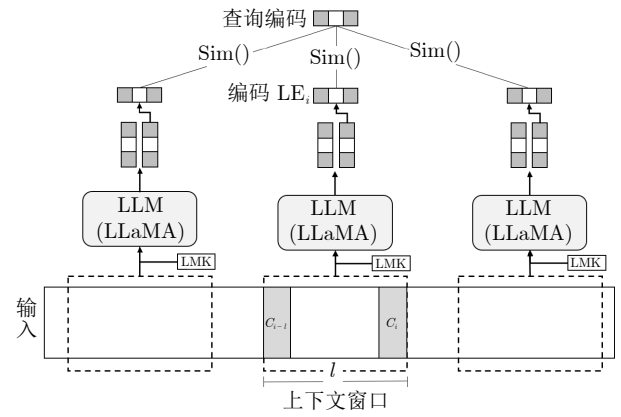


图 4 BGE-LE 方法的框架

Fig.4 The framework of BGE-LE

图像模态数据, ResNet^[65]、EfficientNet^[66] 等模型能够从图像中提取特征, 实现图像数据向特征向量的转化, CLIP^[67] 等对比学习方法通过同时获取图像和文本的特征向量, 还可以实现图像与文本之间跨模态对齐; 针对音频数据, VGGish^[68]、Wave2Vec^[69] 等网络模型能够从音频中通过有监督或无监督的方式提取特征向量; 针对视频数据, 3D-CNN^[70] 能够在空间和时间维度上进行卷积操作, 提取视频的时空特征, 基于 Transformer 的 TimeSformer^[71] 模型可以通过注意力机制捕捉视频中的时序信息, 形成特征向量. 为提升语料库条目的关联分析能力, 知识图谱也可以作为一种有效的数据组织方式应用于 RAG 系统中, 例如, 微软推出的 GraphRAG 平台, 在全局信息和长关联内容分析上表现出显著优势.

2.1.3 非向量式检索方法的数据库构建

非向量式检索方法通常采用基于关键词的检索技术, 其数据库构建过程主要采用基于文本内容的直接匹配和索引机制. 这种检索方法依赖于文本预处理、分词、倒排索引的创建、相关性评分、同义词扩展以及索引优化等步骤. 其中, 文本预处理包括去除停用词、词干提取和词形还原, 以简化文本数据并提高检索效率. 分词是将文本分解成更小的单元, 为后续索引和检索提供基础. 倒排索引记录每个词元在文档中的出现情况, 允许快速检索包含特定词元的文档^[51]. 相关性评分算法用于评估文档与查询的相关性, 并进行结果排序. 同义词扩展通过同义词词典或自动学习来提高检索覆盖率. 索引优化技术如压缩和剪枝可以减小索引大小, 同时保持高效检索. 在非向量式检索方法中, 典型的方法包括 TF-IDF^[51] 和 BM25^[72], 它们分别通过词频和逆文档频率以及排名函数来评估文档与查询的相关

性. 这些方法在 Lucene^[73]、Elasticsearch^[74]、Meilisearch 以及 Redisearch^[75] 等搜索引擎中得到大量应用.

非向量式检索方法以其简单性和高效性在大规模数据处理中具有显著优势, 但在语义理解和复杂查询处理方面存在局限. 实际应用中, 根据需求选择合适的检索方法或将非向量式与向量式检索方法结合使用, 可以提供更全面的搜索解决方案.

2.2 信息检索

通过获取与用户查询相关的文档上下文, 信息检索为答案生成提供必要的背景和上下文, 为大模型生成更加贴近用户需求的响应提供重要支持^[62]. 在高级别的 RAG 系统中, 为提高查询的准确性和返回结果的相关性, 通常还需要在查询前后进行查询指令的转换增强和召回结果的筛选过滤. 为此, 本节将围绕查询构建、知识检索和知识过滤三个部分对相关技术进行介绍^[76]. 信息检索反映用户与知识库的交互过程, 其核心在于如何准确地反映用户的需求并将其与知识库的表征方式进行对齐.

2.2.1 查询优化

考虑到系统用户水平的专业差异, 原始查询存在着潜在的语义缺失、语义模糊和语义冗余等问题, 易对检索过程造成干扰. 查询优化模块以用户原始查询为输入, 借助语法分析、语义理解、内容生成等工具, 进行高质量查询构建, 从而提升检索的效率和准确性. 本节将主要介绍查询转换和查询增强两大类方法.

1) 查询转换. 查询转换是将原始查询转换成更加明确的形式, 以便更准确地表达用户的需求. 这种转换包括修正语法错误、替换模糊信息、补充缺失关键字等初级转换方法, 也包括对查询进行重写及分解等高级转换方法. 初级转换方法往往依赖于规则的策略、语法库、知识图谱等进行查询优化, 如基于领域本体的语义查询可以通过扩展同义词、上位词和类似词, 从而提高查询的准确性和信息召回质量^[77]; 基于图的信息检索系统将查询与句法和语义扩展相结合, 以增强检索所需的信息^[78]. 高级转换方法通常采用基于大模型的生成式方法进行查询的转换, 优化重点在于语义级别的优化, 包括语义分解、语义抽象、语义补充等. 其中, 考虑到实际查询的复杂性和多重需求, 单一查询时常难以全面反映用户的实际意图. 利用大模型自身的任务理解能力能对原始查询进行适当分解从而产生多级子查询, 是一种有效的提高检索相关性并进行层次化答案生成的策略^[79]. 此外, 结合大模型自身的推理和

生成能力, 对原始查询进行重新表述, 从而消除歧义或提高查询合理性也能够提高查询准确性^[80], 例如, 退后提示 (Step back prompting) 方法将原始查询转换为更为抽象和一般性的“退后问题”来丰富检索的上下文^[81]. 为增强大模型回答的连续性和一致性, 在查询重新表述过程中, 结合历史对话信息进行查询优化同样十分必要. 作为案例, 图 5 展示了查询重写方案 Query-Rewriter^[80]. 该方法以可训练的 LLM 为查询重写器, 通过对原始查询进行提炼和分解产生一个或多个新的查询, 并依次作为后续检索过程的输入. 为对重写器进行优化, 该方法提出对语言大模型最终的生成结果进行评估, 并构建奖励函数, 从而监督重写器的训练过程. 这种方式有助于提高对特定任务的针对性和查询的准确性.

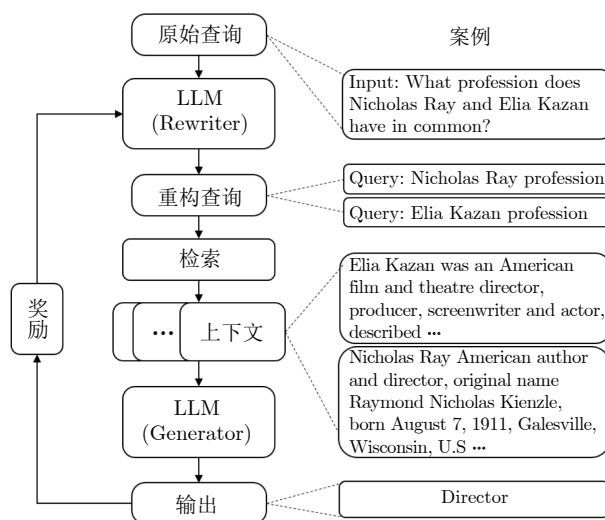


图 5 Query-Rewriter 方法的框架
Fig.5 The framework of Query-Rewriter

2) 查询增强. 查询增强方法旨在为原始查询添加额外信息, 如前置信息以及引导性信息等, 从而提高检索过程的信息覆盖全面性. 其中, 典型的方法包括多路召回 (Multi query retrieval)^[82] 和假想文档嵌入 (Hypothetical document embeddings, HyDE)^[83-84] 等. 多路召回方法通过 LLM 等工具利用多个变体查询或相关查询来增强原始查询, 从而提高搜索结果的多样性和覆盖率. 这种方法认为, 不同的查询可能会触及信息检索系统中的不同部分, 使得用户能够获得更全面的信息集合. 例如, 对于一个关于“可再生能源”的查询, 多路召回方法可能会生成包括“太阳能”、“风能”和“生物质能”等相关查询, 每个查询都独立进行搜索, 最后将所有结果合并, 从而提供一个全面的信息视图. 而假想文

档嵌入方法则首先利用大模型直接生成针对查询的回复结果(假想文档),并将假想文档嵌入原始查询进行查询增强.例如,LLM根据查询生成一个或多个假想文档.这些文档可能包含对“可再生能源”的优势的描述,例如太阳能的可持续性、风能的清洁性等.值得注意的是,假想文档所反映的事实未必完全正确,而是只需要能够在形式或者语义表述上与查询相关就有望提升检索内容的精准性.图6展示了Query2doc方法^[84]的原理图,该方法结合用户查询以及少数的样例信息等内容,动态地生成假设文档,并将其与用户查询等数据进行融合后,再进行信息的检索.

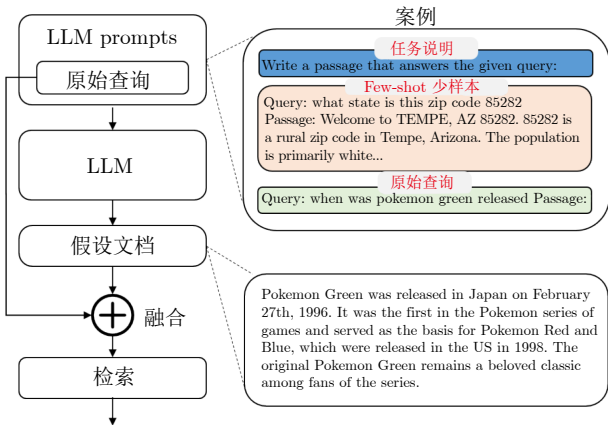


图6 Query2doc方法的框架

Fig.6 The framework of Query2doc

2.2.2 知识检索

知识检索旨在从数据库中挖掘与用户查询相关

的有效信息,这一过程不仅需要考虑文档的相关性,还需要考虑其中所包含的知识和信息的意义和价值.本节将围绕检索过程的必要性、检索内容选择和检索策略几个角度,简述进行RAG系统检索功能升级的典型方法.值得注意的是,出于论述的清晰性需要,本节将不同方法进行单独阐述,但在实际系统和应用中,下述各类方法通常组合使用.

1) 预判式检索.当用户输入的查询已经包括清晰且明确的信息时,引入检索过程的必要性将大大降低.例如,在一般的基于大模型的问答系统中,用户通过提示工程等手段已经为信息生成构建清晰的需求描述和必要的指示信息,此时如果再进行检索操作将引入不必要的计算和时间开销,而且还可能引入干扰信息,从而降低用户的问答体验. Self-RAG^[85]提出在检索过程进行前首先对检索必要性进行判断,根据判别网络的预测结果,决定是否进行检索操作.此外,预判式检索机制的引入还可以实现自适应检索,通过对已检索得到的内容的判断,来实时确定继续检索的必要性,从而提高计算资源的分配性.如图7所示, Self-RAG^[85]利用LLM对查询 x 和历史迭代产生的答案 $y_{<t}$ 等信息进行判断,从而判断是否需要进一步检索 \widehat{ret} :

$$(x, y_{<t}) \rightarrow \widehat{ret} \in \{yes, no, continue\}$$

在需要进一步检索时,根据检索得到的上下文 d ,判断每一条检索结果与查询的相关性 \widehat{rel} :

$$(x, d) \rightarrow \widehat{rel} \in \{relevant, irrelevant\}$$

之后,将检索的上下文 d 与原始查询以及历史

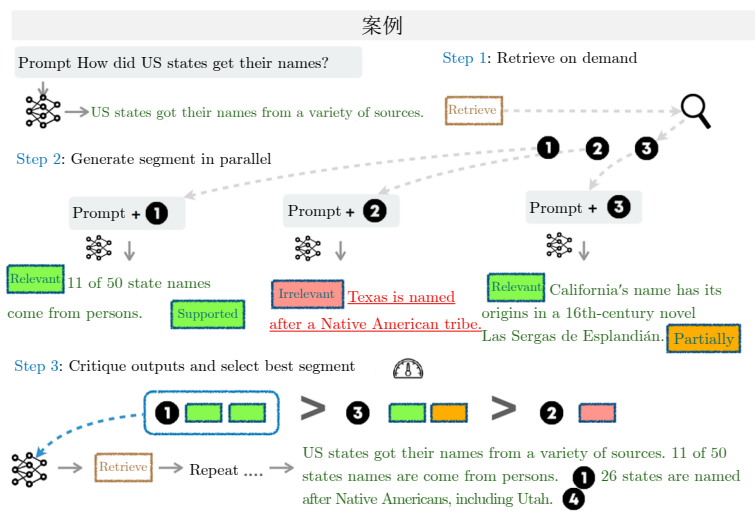
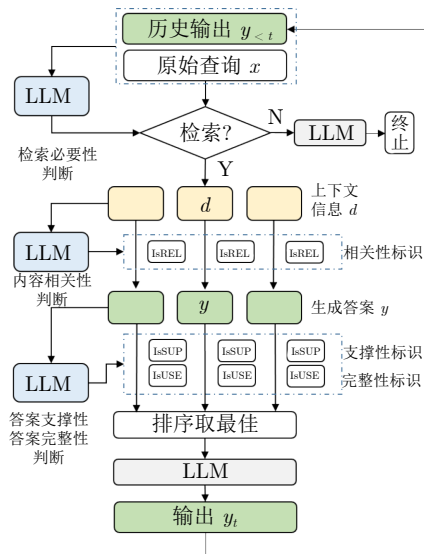


图7 Self-RAG方法的框架

Fig.7 The framework of Self-RAG

输出结合后输入 LLM 中进行内容生成, 从而产生当前结果 y_t :

$$(x, d, y_{<t}) \rightarrow y_t$$

为评估生成结果对查询的回答效果, 进一步利用 LLM 对输出的有用性 \widehat{use} 进行判断, 分数越高有用性越高:

$$(x, y_t) \rightarrow \widehat{use} \in \{1, 2, 3, 4, 5\}$$

上述过程不断循环, 直至 \widehat{ret} 为 no 为止。

2) 主动式检索. 在诸如长文本的生成任务中, 通过单次检索提供上下文信息往往难以适应生成过程的动态需求, 从而导致生成结果的不理想. 为此, 主动式检索方法, 如 Self-RAG^[85] 和 FLARE^[86], 提出利用大模型本身的任务理解能力, 通过对用户查询和实时生成内容的评估, 以确定检索的时机和检索的内容. 与预判式检索类似, 这种方法同样使用大模型来判断检索过程的必要性, 但除此之外, 主动式检索方法还通过已有生成结果来进一步生成检索的查询, 从而对知识进行针对性探索。

3) 多源检索. 考虑到不同数据库存储信息的差异, 为建立对用户查询的全面准确回答, 使用多个信息源 (如农业、工业、医学等不同内容的数据库以及向量、图和关系数据库等不同类型的数据库) 进行上下文获取是一种有效的改善回答质量的方法^[87-88]. 为在检索过程中实现对多数据库的管控, 常用的方法是利用 LLM 等进行相关信息的选择和检索结果组合^[87].

4) 多层次检索. 在多文档检索的基础上, 为提高检索过程中对无关信息的过滤效率, 在检索过程中, 可以通过知识库环节构建的多层级知识库实现快速的信息检索^[89-90]. 例如, 可以通过对不同的原始数据库建立摘要和关键字信息并作为元数据从而建立元数据知识库. 在此技术上, 针对用户查询可以先通过对元数据知识库的检索, 定位相关文档数据库, 并进行进一步的原始信息检索. 这种方式可以快速缩小检索范围, 从而将计算资源集中在具有高相关性的若干文档上, 实现动态路由以提高效率。

5) 多策略检索. 也称混合检索 (Mixed retrieval), 该类方法的典型特征是采用不同的检索技术进行文档查询, 其中, 语义向量检索和关键字检索混合的方案 (也称融合稀疏向量检索方法和稠密向量检索的方法) 是常用的策略. 多策略检索能够从不同的角度实现对相关信息的挖掘, 从而提高信息召回率. 单纯式 RAG 系统采用基于向量的检索方法, 这种方式能够实现高层级的语义理解, 但是缺乏对表述不充分场景的有效应对能力. 例如, 用户输入

的查询指令中常会出现仅包含几个离散关键词的场景, 这将导致难以通过嵌入编码正确提取语义信息进而影响检索效果. 而传统的基于关键字的检索方法则能够有效处理这种情况. 通过将不同检索系统组合以进行检索优势互补, 能够提升查询与知识库内容的多角度关联, 是在多样化输入场景下保障检索质量的有效手段. Adaptive-RAG 提出一种通过对查询所对应任务的复杂度进行预测, 从而使用零步、单步或多步检索的方案, 实现对检索效率和检索结果的平衡^[91].

2.2.3 检索信息后处理

在进行生成过程前, 是否有必要对检索得到的信息进行进一步处理是一个值得讨论的问题. 理论上, 在信息检索阶段, RAG 已经基于相似性计算获取数据库中与查询最为相关的上下文片段信息, 但实际应用中, 出于对计算效率的考量, 这种相似性检索往往会通过引入不确定因素以提高检索的速度^[92], 这也造成检索得到的结果与查询之间有时并不具有高度相关性. 另外, 基础 RAG 的相似性计算过程采用的是双编码器 (Bi-encoder) 架构, 该方法通过两路编码器分别计算查询和上下文的嵌入表征, 之后通过余弦相似度等方法计算两个嵌入的相似性. 这种方式的优势在于在相似性计算的过程中能够独立地获取输入表征, 从而为嵌入的事先计算提供便利, 但也由于未进行查询和上下文间的深度交互, 从而易导致高层级语义相似度的匹配误差。

1) 信息评估. 检索结果的质量对大模型内容生成效果具有重要影响, 因此, 高级别的 RAG 系统中通常会引入信息评估模块. 常用的评估维度包括: 相关性、支持性以及可用性^[85]. 其中, 相关性指标用于评估检索到的文本段落 (或证据) 与用户查询和历史信息的相关程度; 支持性指标用于衡量检索到的文本段落对生成答案的支持程度; 可用性衡量生成的答案与查询的契合程度, 例如, 是否回答用户的问题. 相关性和支持性指标可视为对检索结果的直接评估, 而可用性指标则是对检索结果的间接评估. Self-RAG^[85] 通过检索和生成结果的评估, 建立自我反思, 动态调整和优化 RAG 的检索过程, 从而满足不同任务的需求和用户偏好. CEG^[93] 方法对回答内容生成引用, 利用自评估机制来根据引用评估答案的准确性, 并对不可靠的回答进行重新生成。

2) 重排序. 该技术旨在对初步检索得到的上下文条目进行重要程度的重新排序, 以挑选出更符合用户查询和历史对话信息的内容^[94]. 常用的重新排序方法是基于交叉编码器的排序方法和基于大模型

的重排序方法. 基于交叉编码器的方法将检索得到的上下文信息和查询或历史对话进行组合后, 送入网络中进行相关程度的判断. 由于引入查询和上下文之间的信息交互以及基于神经网络的相似性计算方法, 该方法能够提取更深层次的语义关联, 从而优化相似性评估效果. 基于大模型的方法则利用大模型本身具备的强语义理解能力, 对检索得到的上下文信息进行自动排序, 优先选择更为相关的检索结果. 图 8 展示了 RichRAG^[95] 模型中使用的检索结果排序方法, 该方法首先将原始查询分解为若干个子查询, 并基于子查询进行检索过程. 针对检索得到的上下文信息, RichRAG 将其与原始查询一起组成新句子, 输入排序网络 (Ranker) 中进行排序. 排序网络由编码器解码器组成, 其中编码器对输入句子内容进行编码, 解码器则利用特殊标记对句子信息进行聚合以建立全局特征, 并据此进行排序结果生成.

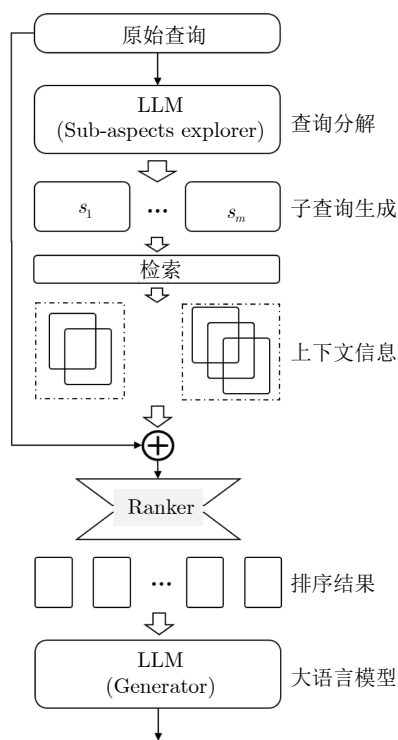


图 8 RichRAG 方法的框架
Fig.8 The framework of RichRAG

2.3 内容生成

RAG 系统中, 通过向量数据库检索到的相关上下文信息作为额外的补充知识, 对原始任务信息进行增强, 从而提升模型最终的生成质量和效果. 该信息生成过程主要分为两类关键步骤: 信息增强和结果生成. 值得注意的是, 在很多 RAG 系统中,

信息增强和结果生成过程是高度耦合的或迭代进行的, 本文为陈述清晰, 将对其进行分别阐述.

2.3.1 信息增强

信息增强的关键在于基于检索到的内容, 面向查询任务提供更加全面的上下文信息, 从而为进一步产生更完善、相关、流畅的结果奠定基础. 根据信息增强在 RAG 系统中所处的环节, 可以分为提示信息增强、特征融合信息增强和输出信息增强^[12, 23], 信息融合增强分别发生在内容生成模型之前、生成模型中和生成模型之后, 其增强的方法和特性也有所不同.

1) 提示信息增强. 基于检索到的结果对初始用户查询的提示进行扩充、重组和修改, 并将更新后的提示输入 LLM 等生成器, 通过改善输入预训练生成模型中的上下文信息, 从而提升最终检索增强生成的效果. 该类信息增强方法逻辑简单, 目前得到广泛应用. 在提示信息增强中, 最为关键的是如何进行上下文信息的集成, 而不发生信息冲突、重复冗余、过度依赖检索内容等情形导致生成结果不连贯或者产生错误内容. 通常采用的增强方式有基于相关性的检索信息重组和提示模板引导等. 例如, Self-RAG^[85] 在获取到检索内容后, 对其与任务查询的相关性和支持度进行评估, 由此选择能提升生成质量的片段重组到原始的提示中. 此外, 通过提示模板中显式的信息提示, 来指示生成模型如何使用这些检索到的信息, 同样可以提升其对检索信息的利用效率. ActiveRAG^[96] 针对检索信息主动构建多角度知识, 结合思维链 (Chain of thought, CoT) 补充未知信息和识别混淆信息, 克服标准 RAG 被动利用知识导致的信息松散问题, 提升答案的准确性.

2) 特征融合信息增强. 将检索到的知识信息通过隐空间潜在特征的方式, 在生成模型中与原有任务信息特征表示进行融合, 实现在隐空间中的信息增强, 从而提升生成模型的关联理解能力和生成效果. 该方法更加高效, 可以有效地减少计算复杂度, 并克服 LLM 输入窗口大小限制的问题. 比较典型的方式有解码器融合^[97], 其基于检索获取到的能够产生支撑作用的内容和知识进行编码, 进一步在解码器中融合多个检索内容片段, 实现高效的上下文信息关联. 该方法在文本处理和知识问答领域得到广泛的应用, 并展现出其在提升模型效率和准确性上的优势. 该方法同样应用于代码摘要生成中, 从训练集中检索到最相似代码片段, 将输入代码片段和检索到的代码片段进行编码, 并在解码过程中

进行融合以预测摘要^[98], 从而提升代码到文本转换的整体效果。

3) 输出增强. 主要在生成模型的输出阶段与检索内容进行融合, 比较通用的做法是通过衡量检索内容与原始任务的匹配程度, 用检索到的信息对生成内容进行融合增强. 例如, KNN-LM 检索与目标任务相近的 k 个答案并建立其分布, 与原始模型输出的分布进行插值融合, 从而利用训练数据集中的相关知识增强输出结果^[99].

2.3.2 生成优化

结果生成环节主要针对基于检索信息增强产生最终输出结果的过程进行优化, 关键影响因素包括上下文信息、生成模型、检索增强生成模式、后处理方式等. 其中, 上下文信息要素在信息检索和信息增强部分进行了分析, 本节主要面向生成模型、RAG 的生成模式和生成结果优化进行阐述。

1) 生成模型优化. 生成模型是检索增强生成的核心模块, 它所具备的上下文关联理解能力、推理能力和生成能力对模型生成结果有着决定性的影响. 因此, 针对生成模型进行面向特定任务的优化是提升 RAG 性能的重要方式. 典型的方法包括生成模型微调和超参数调整. 对于前者, 通过在特定领域文档和知识的基础上进行模型微调, 提升模型整体对该领域数据库的适应性和生成能力; 对于后者, 在原始生成模型的基础上, 添加额外的控制模块, 通过调整模型的超参数来调节模型的不同倾向, 例如生成结果的多样性、随机性和内容限制等^[100].

2) 生成模式优化. 对于需要复杂推理过程的问题而言, 单次的检索增强生成可能无法直接生成准确、自洽、连贯的回答, 难以满足最终的任务需求, 如何改进整个检索与生成的配合流程是重要的研究内容. 较为典型的方法有检索选择和迭代检索增强生成, 其中所使用的检索模式采用前文所述的预判式检索^[85]和主动式检索^[86]. 检索选择方法的假设是, 当模型本身知识足以覆盖相关任务领域问题时, 过度的检索可能会带来结果的不一致. 往往可以通过设定选择检索比例的超参数或评估检索所带来的性能提升, 来对是否进行检索增强进行选择 and 平衡. 对于迭代检索方法, 则通过迭代地进行整个检索增强生成的过程, 逐步提升最终的内容生成结果. 例如基于模型当前的生成结果判断所缺失的知识, 在下一轮检索重点补充相关的上下文信息, 这与前文所述的主动式检索是一致的。

3) 生成结果优化. 尽管 RAG 基于向量数据库融合了大量专业、实时的数据和知识, 但其生成内容仍然无法避免幻觉和矛盾的结果. 因此, 大量方

法基于面向任务的特定规则, 对 RAG 系统的生成结果进一步修改和优化, 以满足任务要求. 例如, 对特定的生成片段进行分类和替换, 以满足生成内容上下文的行文规范和语言风格. 根据生成内容的概率, 对候选项进行重新排序, 从而获得多样化的代码修复生成结果^[101]. 此外, 根据不同的上下文知识语境和原始任务提示, 生成多个答案, 并对答案进行逻辑上的整合, 从而输出更加全面深入的结果. 这些直接针对生成结果的优化因任务性质的不同而存在较大差异。

3 RAG 应用

本节围绕自然语言处理、多模态任务以及垂直场景, 介绍了 RAG 方法的应用现状与实际作用, 并在表 2 中对其典型应用进行了归纳总结^[87-119].

3.1 语言应用场景

在 LLM 中, RAG 技术的应用主要体现在两个方面: 1) 通过检索相关信息, 将外部数据检索整合到生成过程, 提高生成文本的质量和准确性; 2) 利用检索模型对生成的文本进行评估和优化. 通过与检索模型的结合, LLM 可以更好地理解上下文信息, 生成更加合理和连贯的文本. 此外, RAG 技术还可以帮助 LLM 应对一些特定领域的挑战, 避免生成不准确或不相关的内容, 提高生成文本的多样性和创造性. 在推断阶段动态检索来自知识库的信息使 RAG 能够解决 LLM 生成事实错误内容或“幻觉”的问题, 并成为完善聊天机器人、代码生成等使 LLM 更适用于实际应用的关键技术。

3.1.1 聊天机器人

近年来, 基于 LLM 的聊天机器人展现出令人难以置信的能力, 可以即时创建回复并与用户长时交流. 然而, 现有的聊天机器人模型大多只能根据预先训练的数据生成回答^[102], 缺乏对实时信息和上下文的理解能力, 存在显著的不足. 为应对上述挑战, RAG 技术被引入到 LLM 聊天机器人的相关研究和应用中, 其旨在从外部知识库中获取准确和最新的信息内容, 通过检索、增强和生成为 LLM 提供更强大的上下文信息和理解能力, 帮助聊天机器人更好地理解用户的意图, 并以更灵活和智能的方式针对不同应用场景提供准确和个性化的问答服务^[21, 103]. 例如, 在企业的智能化客服服务中, 利用 RAG 技术检索公司的知识库和常见问题解答数据库, 从而构建更智能的客服机器人, 为客户提供更专业、快速、个性化的解答和帮助; 在特定应用的智能助手, 利用 RAG 技术帮助用户获取各种允许查询的个人

表 2 基于 RAG 的应用案例
Table 2 RAG-based application cases

方法	应用领域	RAG 作用	方法介绍
UniMS-RAG ^[87]	通用对话	个性化	知识库阶段, 构建人物角色库与上下文语料库
ERAGent ^[104]	通用对话	个性化	生成阶段, 使用人物角色资料作为提示构建的输入
HyKGE ^[105]	医疗问答	专业化	检索阶段, 基于医学知识图谱增强医学知识理解
CBR-RAG ^[106]	法律问答	专业化	数据库阶段, 基于法律案例库增强法学知识理解
uRAG ^[107] , SEA ^[108]	通用对话	实时化	检索阶段, 基于搜索引擎的 RAG 系统
RA-VQA ^[109] , KAT ^[110]	视觉问答	知识增强	生成阶段, 基于检索的知识增强视觉推理能力
Plug-and-Play ^[111] , MuRAG ^[112]	图像描述	知识增强	生成阶段, 基于检索的知识增强视觉推理能力
RA-CM3 ^[113] , Re-Imagen ^[114]	图像生成	知识增强	生成阶段, 基于检索的知识丰富上下文信息
RAC ^[115]	图像分类	长尾分布	生成阶段, 融合原始图像和检索内容特征
文献 [116], Make-an-Audio ^[117]	语音翻译	数据增强	基于检索构建多样化样本
RAG-Driver ^[118] , 文献 [119]	自动驾驶	可解释性	生成阶段, 基于 RAG 提取相似场景案例

信息, 并通过信息增强对多个信息源进行整合、动态生成和更新, 应用于智能助手问答系统, 从而根据用户的查询提供更全面、个性化的回复; 在个性化信息检索系统中, 利用 RAG 技术可以有效集成外部数据库、知识库以及用户的兴趣、喜好记录, 通过将用户查询与外部数据库进行匹配, 为用户提供更准确、相关、有针对性的推荐和建议。

3.1.2 代码生成

将程序视为语言序列, 以递归神经网络和 Transformer^[120] 为代表的神经序列架构可以自然地用于实现代码生成。近年来, 具有 120 亿参数的 OpenAI Codex^[121] 模型展示出在数十亿行公共代码上预训练的大型代码生成模型的潜力, 开创了基于 LLM 的代码生成应用。通过使用生成预训练策略, Codex 可以极高概率解决 Python 语言的入门级编程问题, 用户研究也表明由 Codex 提供支持的付费服务 GitHub Copilot 的 88% 的用户能够使用它进行更高效的编码^[122]。自此, 更多的预训练代码生成模型纷纷出现, 如 DeepMind AlphaCode^[123]、Salesforce CodeGen^[124]、Meta-InCoder^[125]、Google PaLM-Coder-540B^[126] 等。然而, 由于程序编写涉及的需求多样性和专业复杂性, 大部分的代码生成模型在多语言代码模型的功能正确性、生成代码的完整性与准确性方面仍然与专业程序员有着极大的差距。基于 LLM 的代码生成涉及大量的代码库和技术文档等信息, 通过引入 RAG 技术, 结合检索、增强和生成, 模型可以从海量数据中检索到相关信息, 并结合生成模型形成高质量的代码, 能够在一定程度上克服传统生成模型中存在的信息不完整和重复性等问题, 并帮助模型更好地理解程序员的意图, 提高代码生成的准确性和效率^[127]。例如, 最近发布的具有 130 亿个参数的 CodeGeeX 多语言代码生

成模型^[127], 通过引入 RAG 算法, 构建流行公有仓库和私有仓库的代码向量数据库, 缓解代码生成模型幻觉性问题, 如避免生成错误的私有函数调用、让模型拥有最新的代码仓库知识、对私有代码仓库建立知识库等。此外, CodeGeeX 还支持语言对之间的代码解释和代码翻译任务, 并支持几种不同的模式(代码完成、函数级生成、代码翻译、代码解释和自定义提示), 以实时帮助用户执行编程任务, 填补多语言代码生成的空白^[127]。基于 RAG 构建的模型扩展显示了其在提高代码多语性、编码效率、编码准确性方面的显著潜力, 能够帮助研究人员和开发人员广泛利用 LLM 进行代码生成。

总的来说, 通过外部数据库和知识库的扩展以及实时信息和上下文的应用, RAG 技术应用可以帮助 LLM 系统提供更准确和个性化的回答、更好的交互体验, 增强用户对系统的信任和满意度。RAG 技术可以很容易地扩展和适应不同的应用场景和领域需求, 但对于 LLM 泛化性与专业性的平衡, 即如何在泛化能力和特定领域的深度知识之间找到平衡点仍然有待探索。此外, 外部数据库的可靠性、计算资源的受限性以及数据隐私保护问题涉及的操作合规性也是 RAG 应用中亟需考虑的问题。

3.2 多模态场景下的应用

知识不仅以文本形式存储, 还可以以其他形式或模态进行存储, 例如图像、视频和音频等。这些知识通常在传统的文本语料库中无法访问、无法使用或无法描述。为解决这些问题, 针对多模态数据的检索增强生成技术应运而生。每种模态都有不同的检索和生成方法、任务和挑战。本节针对不同模态类型进行讨论和分析, 涵盖图像、视频、音频和其他模态的数据。

3.2.1 图像和视频领域

随着一系列视觉语言大模型的提出^[67], 相关的各类检索增强生成的方法也陆续提出, 旨在更好地将来自图像和视频的外部知识纳入到模型中. 在一般文本生成任务中, 图像和视频模态的信息也可以作为额外的知识来源来提高本文的生成质量.

在视觉问答领域, RA-VQA^[109] 联合训练文档检索器和答案生成模块, 首先将目标图像转换为文本数据, 然后使用密集通道检索获取与图像相关的文本文档, 最后将每个文档与初始问题连接以生成最终答案. KAT^[110] 为解决外部知识视觉问答任务 (Outside knowledge visual question answering, OK-VQA), 对 GPT3 进行检索并获取相关知识, 支撑最终视觉问答的答案生成过程.

在图像描述领域, Plug-and-Play^[111] 利用 Grad-CAM^[128] 方法检索和输入的问题相关的图像, 然后将检索到的图像结合原始的图像获得增强的上下文, 最终再进行图像描述. 类似地, MuRAG^[112] 也通过对文本和图像数据的检索增强文本上下文来提高模型的生成能力. 在生物医学领域, RAMM^[129] 检索相似的生物医学图像和标题, 并设计不同的网络结构对检索到的信息进行编码. 为生成多种风格的描述, SACO^[130] 使用风格感知的视觉编码器对相关的图像内容进行检索, 然后再进行描述的生成.

RA-CM3^[113] 可以同时生成图像和文本. 研究表明, 基于检索增强的图像生成在知识密集型生成任务上表现更好, 并展示了多模态上下文学习等新能力. Re-Imagen^[114] 利用多模态知识库检索图像-文本对, 以促进图像生成, 超越了在生成文本前检索图像的范畴. RAC^[115] 通过从一个由预编码图像和文本组成的非参数存储器中检索, 改进图像分类中长尾分布问题. K-LITE^[131] 是一种知识增强方法, 利用外部知识源 (如 WordNet^[132] 和 Wiktionary^[133]) 来丰富自然语言监督. 它在语言-图像预训练和基于提示的评估阶段都使用这些知识, 并在图像分类和目标检测等经典计算机视觉问题上表现出较高的通用性和有效性. REACT^[134] 是一个即插即用的框架, 利用大规模图文语料库作为外部知识, 可以有效地对下游任务的模型进行定制. 在超过 20 个不同的数据集上, 包括图像分类、图像-文本检索、目标检测和语义分割等多种任务中证明了该方法的通用性和有效性.

在视频领域, KaVD^[135] 是一种基于检索的视频描述生成方法. 对于新闻视频, 它检索与主题相关的新闻文档, 然后使用基于知识的视频描述网络生成描述. VGNMN^[136] 是一种从视频片段中提取视觉

线索以增强视频对话的方法. 通过构建特定的网络结构来进行视频检索, 并使用先前对话中的实体和动作实例化这些网络.

3.2.2 音频领域

音频 RAG 在特定音频语言任务中发挥着重要作用, 如音乐字幕、音乐文本生成和语音识别. 对于文本音频任务, 最重要的挑战之一是缺乏音频-文本对的训练数据. 而对音频和文本知识的检索可以有效缓解这一数据稀缺问题, 提高模型的生成性能. 因此, 将音频 RAG 应用于音频处理领域具有广阔的发展前景^[22].

为应对语音翻译领域数据稀缺的挑战, Zhao 等^[116] 提出 Spoken-Vocab 技术, 通过将机器翻译转换成合成的语音翻译数据来扩充现有的语音数据集. 具体而言, Spoken-Vocab 技术通过检索并拼接与机器翻译句子中的单词对应的音频片段来生成合成音频. 实验证实, 利用拼接的音频片段能够有效提升翻译质量. Kim 等^[137] 利用基于预训练语言模型的方法来解决数据稀缺问题. 该方法从输入的音频中提取特征, 并使用深度神经网络将其映射为连续向量, 然后将这些向量作为前缀向量对预训练语言模型进行前缀微调. 利用从检索到的音频中获取的额外信息, 模型在文本生成音频的任务中的性能得到显著增强. 类似地, Mestre 等^[138] 将音频特征整合到预训练语言模型中, 从而在数据稀缺的情况下提高模型的生成性能. Huang 等^[117] 应用音频-文本检索方法来获取伪文本提示, 进而改善在数据稀缺情况下的音频生成.

3.2.3 其他模态数据

幻觉问题是生成式模型一直未有效解决的难题之一, 即模型有可能输出虚假信息. 因此, 一种可能的解决方案是利用检索到的结构化知识 (例如知识图谱、表格和数据库) 来对生成的内容进行辅助增强.

在知识图谱问答领域, Shu 等^[139] 采用多粒度检索来提取相关的知识库上下文, 并使用约束解码来控制输出空间. 在表格问答领域, Pan 等^[140] 提出一种使用基于 Transformer 的模型来检索最相关的表格并定位正确单元格的方法. 此外, 通过对知识的选择和利用, 还可以提高推理任务模型的性能及可解释性. Yang 等^[141] 提出一种数学推理器, 首先检索高度相关的代数知识, 然后将其作为提示以改善生成任务的语义表示. 随着 CoT^[79] 技术的发展, He 等^[142] 和 Li 等^[143] 根据链式推理从知识图谱和知识库 (如 Wikidata) 中进行检索.

3.3 垂直场景下的应用

当前, LLM 已经在医疗保健、法律金融、工业、自动驾驶等许多垂直场景得到应用, 但其固有的幻觉、知识更新速度慢、缺乏领域知识等缺陷制约各类应用的效果. 为解决这些问题, RAG 技术广泛应用于各个垂直场景当中, 以提高 LLM 的性能. 对于医疗、法律等高风险场景, RAG 技术的使用可以有效提高大模型输出的内容生成的可追溯性和可解释性, 以减少幻觉带来的风险. 此外, RAG 技术还可以结合特定场景的私有数据, 完成特定领域的知识问答, 提高 LLM 的人机交互能力.

3.3.1 医疗健康场景

医疗健康场景包含大量知识密集型和专家依赖型任务, 这为通用医疗大模型的构建提出巨大挑战^[144]. RAG 通过检索外部信息来增强内容生成, 在外部知识的帮助下减少知识密集型任务中的事实错误. 例如, Jiang 等^[105] 结合 RAG 技术构建基于假设知识图谱增强的医疗对话系统, 该系统利用 LLM 对用户的输入进行补充, 为检索相应的知识提供尽可能全面的信息, 并在检索阶段采用一种基于假设输出的片段粒度感知方法, 以增强用户查询与外部知识边缘推理路径的一致性. 在 MMCU-Medical 医疗数据集上的问答任务中, 相比于非 RAG 的基线方法, 基于 RAG 的方法^[105] 在 Exact match (EM)、Partial correct rate (PCR) 指标上分别提升 14.1% 和 14.4%. 为增强 LLM 在小众医学领域的性能, Kang 等^[145] 提出基于自然语言提示的 RAG 方法, 该方法消除对向量嵌入的依赖, 采用更直接和灵活的基于自然语言提示的检索过程, 并在两项面向韩国医学领域的测试中取得良好的性能. 为实现面向多发性骨髓瘤的精准医疗, Quidwai 等^[146] 提出一种基于 RAG 的聊天机器人框架, 利用自然语言处理技术和 LLM 来管理和分析该疾病的基因组数据和文献等资料, 并根据患者的基因组数据提供个性化的治疗建议. 此外, 制药行业当前存在复杂且大量的法规政策, 企业通常需要花费大量的时间来浏览并校准生产行为, 以确保符合要求. 因此, Kim 等^[147] 提出问答 RAG 模型 (QA-RAG), 可精确地为用户提供有关制药的法规政策指南, 以提高用户决策能力.

3.3.2 法律金融场景

LLM 在法律和金融场景的应用可以优化对专业化信息的处理和分析流程^[148-149]. 针对细分和动态变化的信息, RAG 技术的引入能够实现 LLM 能力的有效拓展. 其中, 为增强 LLM 查询与输出的准

确性, Wiratunga 等^[106] 提出基于案例推理的 RAG 模型 (CBR-RAG), 通过组织非参数记忆, 使案例更有效地与查询匹配, 从而增强 RAG 模型中的检索过程. 该模型的性能在面向法律数据集的实验中得到验证, 在 95% 的置信水平上显著优于没有使用 RAG 的模型. 针对住房纠纷等法律问题, Rafat^[148] 基于 RAG 技术开发聊天机器人原型系统, 并开展实验评估该机器人的理解能力和响应的准确性、完整性. Ryu 等^[150] 提出 Eval-RAG 方法, 基于检索器收集的相关文档来评估 LLM 生成的法学相关的文本质量, 并在面向韩国法律问答任务的实验中验证了该方法的有效性.

3.3.3 自动驾驶场景

当前, 已有大量工作利用 LLM 来增强智能车辆的感知和决策能力^[151], 然而 LLM 固有的幻觉风险可能会损害自动驾驶系统的安全性和可靠性, 因此 Dai 等^[26] 提出 VistaRAG 框架, 该框架采用动态检索机制, 从外部数据库获取高度相关的驾驶经验、实时道路网络状态和其他相关信息, 以帮助 LLM 进行推理和决策, 从而提高复杂交通场景下基于 LLM 的自动驾驶系统的安全性和可信度. 为提高自动驾驶的可解释性, Yuan 等^[118] 开发 RAG-Driver 模型, 以产生高质量的驾驶动作解释和控制信号预测, 在 BDD-X^[152] 和 Spoken-SAX^[118] 两个数据集上的实验表明, RAG-Driver 模型^[118] 能够在用于评估行为解释质量的三个性能指标上, 即 BLEU (简记为 B4)^[153]、METEOR (简记为 M)^[154]、CIDEr (简记为 C)^[155], 优于专家基线和大模型基线方法 (DriveGPT4^[156]). 相比于 DriveGPT4, RAG-Driver 的行为解释准确率质量在 B4 指标、C 指标以及 M 指标上分别提升 14.3%、21.9% 和 3.0%. 此外, 为对行人和驾驶员等道路使用者的行为进行可解释的预测, Hussien 等^[119] 将 RAG 技术和基于知识图谱的贝叶斯推理方法结合起来, 实现对行人过马路行为和车辆变道机动行为的高质量预测.

3.3.4 其他场景

除上述场景外, 还有其他一些垂直应用场景使用 RAG 技术来处理复杂而大量的领域知识, 并提供个性化服务. Bornea 等^[157] 提出 Telco-RAG 方法, 该方法旨在处理电信领域复杂的标准文档, 特别是第三代合作伙伴计划 (3GPP) 文档的特定需求. Gaddala^[158] 综述基于 RAG 构建的智能体在供应链管理场景中的实际作用, 包括需求预测、库存管理、供应链操作和实时决策中的作用. 此外, 为使 LLM 能够像农业专家一样, 根据不同地区的气

候和传统提供个性化的农业知识服务, Gupta 等^[159]基于 RAG 技术提出一个 LLM 工作流程, 包括识别和收集涵盖广泛农业主题的相关文件, 然后对这些文档进行清理和结构化嵌入, 以便 LLM 生成高质量的问答对, 并根据问答对质量进行评估和过滤.

3.4 RAG 应用总结

表 2 展示了不同 RAG 方法在 NLP、CV 及垂直领域的应用情况, 从表 2 中可以看出, 在 NLP 任务中, RAG 方法主要提升 LLM 在个性化问答、专业知识理解以及实时信息获取方面的能力. 而在 CV 任务中, RAG 技术则主要通过案例检索、背景上下文补充等提升视觉模型的知识理解水平, 同时, 由于检索能够有效克服长尾分布问题, 因此 RAG 方法能够有效提升对低频类别目标的感知能力. 在音频领域, 当前 RAG 方法主要基于检索过程对语音数据进行合成和增强, 提升数据的多样性. 在诸如自动驾驶的实际应用中, RAG 技术能够通过通过对相关案例的检索给当前智能体的行为决策提供建议和指导, 从而提升智能体行为决策的可解释性. 由上可见, RAG 在不同领域应用时所发挥的作用存在着一定的差异. 在 RAG 技术的应用推广中, 开源 RAG 平台发挥了重要作用. 表 3 给出了部分热门开源平台的汇总及特点介绍, 包括功能多样且可拓展性强的 LangChain, 专注高数据搜索的 LlamaIndex, 轻量化和灵活的 Embedchain, 知识图谱增强信息理解的 GraphRAG, 具有自动化 RAG 构建能力的 RagFlow 等, 这些平台推动了 RAG 技术在各领域的创新与发展.

4 SAGE: 搜索增强的生成与扩展技术

通过在外部数据库和知识库中检索相关信息,

RAG 技术使 LLM 各方面的性能都得到增强. 然而, 当前 RAG 技术仍然存在一些挑战, 例如知识库依赖、信息损耗和检索计算延迟等. 这些挑战限制 RAG 技术在更大规模和更加复杂的任务上的应用. 首先, RAG 技术中使用的检索方法大多依赖于关键字和基于相似度的搜索, 当输入查询的质量不高时, 将会严重影响 LLM 系统的输出质量; 其次, 随着外部数据源的不断扩展, 当在外部数据源中无法检索到查询的相关内容时, LLM 系统可能会输出不准确的结果. 当前 RAG 技术已经应用于许多领域, 用于生成文字、代码、图像、音频和视频等多模态的信息, 但基于信息和知识外挂的思想仍有进一步拓展至生成任务之外的其他任务上的需求.

基于以上分析, 本文给出从 RAG 技术到 SAGE (Search-augmented generation and extension) 技术的转变^[160], 其基本架构如图 9 所示. 检索强调在现有数据库中的搜索, 然而随着 LLM 系统的应用规模和复杂性不断上升, 外部数据源的需要不断扩展, 传统的检索技术将难以应对这种趋势. 因此, 在信息获取方面, SAGE 采用基于搜索的技术来增强系统的主动信息挖掘能力. 此外, 考虑到搜索过程可能带来的非确定性时延, SAGE 引入内/缓存系统的分层管理来处理外部数据源和 LLM 有限输入内容的交换. 参考传统的操作系统架构, 分别采用内存管理单元和缓存管理单元对不同规模和重要性的内容进行动态管理, 实现对历史对话、上下文信息的高效精准配置^[161]. 就实现的功能而言, SAGE 将 RAG 从内容生成拓展至更广泛的领域, 以支持一般性任务, 形成检索增强感知 (Retrieval-augmented perception, RAP)、检索增强思维 (Retrieval-augmented thoughts, RAT)、检索增强决策 (Retrieval-augmented decision-making, RAD)、检索增强行动 (Retrieval-augmented actions,

表 3 RAG 开源平台
Table 3 Open-source platforms of RAG

名称	发布日期	特点	链接
LangChain	2022 年 10 月	功能多样, 可拓展性强	https://github.com/langchain-ai/langchain
LlamaIndex	2023 年 05 月	数据搜索检索效率高	https://github.com/jerryliu/llama_index
HayStack	2019 年 11 月	侧重文本检索和问答应用开发	https://github.com/deepset-ai/haystack
Embedchain	2023 年 07 月	轻量化, 灵活, 可拓展性强	https://github.com/mem0ai/embedchainjs
NeumAI	2023 年 12 月	高吞吐分布式架构	https://github.com/NeumTry/NeumAI
GraphRAG	2023 年 07 月	知识图谱增强的全面信息理解	https://github.com/microsoft/graphrag
Quivr	2023 年 05 月	基于 LangChain 的知识库应用平台	https://github.com/QuivrHQ/quivr
Dify	2023 年 05 月	生成式 AI 开发框架	https://github.com/langgenius/dify
RagFlow	2024 年 07 月	自动化 RAG 构建, 流程精简	https://github.com/infiniflow/ragflow
Open-WebUI	2024 年 02 月	支持友好界面以及完全离线运行	https://github.com/open-webui/open-webui

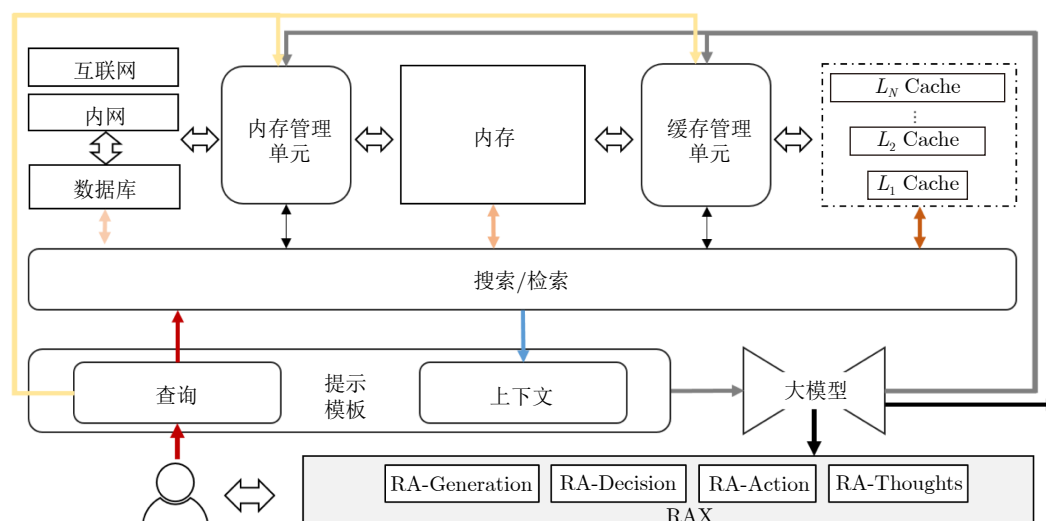


图 9 SAGE 的框架

Fig.9 The framework of SAGE

RAT) 等 RAX (Retrieval-augmented x) 应用拓展, 从而将多样任务集成到统一的框架内, 以提供更加高效的类人思考与行为模式, 实现对智能体尤其是具身智能体知识系统的灵活支撑^[23, 162]. 下文将从信息获取、知识管理以及应用拓展三个方面对 SAGE 进行介绍.

1) 信息获取. SAGE 技术在特定向量数据库的基础上引入基于搜索的增强机制, 以实现开放信息的主动获取. 与传统的 RAG 技术依赖现有数据库不同, SAGE 增加动态地从外部数据源中挖掘相关信息的功能, 并引入智能路由单元实现搜索和检索过程的自适应切换. 动态路由可分为前期路由、中期路由以及后期路由三种模式. 其中, 前期路由在进行外部知识的获取之前对采取特定数据库的检索策略还是采取开放信息搜索策略进行选择. 为此, 需要对内部知识库建立多层次信息摘要, 并利用摘要和用户查询信息的匹配度判断特定知识库是否足以支撑对用户查询的响应, 从而在特定数据库无法支撑用户查询的情况下转而采取面向互联网开放信息的搜索过程. 后期路由则首先执行对特定数据库的检索过程, 并根据检索结果判断是否需要来自互联网的搜索信息. 中期路由用于多步问答场景 (例如 CoT), 在每个子查询任务中, 持续进行检索和搜索机制的选择与执行, 以满足多次内容生成中的不确定和多样性需求.

2) 知识管理. 在应对外部数据源搜索带来的不确定性时延问题时, SAGE 引入层级化的内存和缓存管理机制. SAGE 的知识管理包括预测性知识管理以及历史性知识管理两部分, 并借助多级内存和多级缓存机制来实现. 预测性知识管理根据用户的

实施查询需求, 在执行针对特定查询的检索和搜索过程之外, 通过兴趣和偏好预测模型, 对查询相关的更广泛的信息进行预先加载; 历史性知识管理考虑用户长期的信息访问频率, 赋予高频访问的内容更高的优先权, 从而优化不同规模数据的存取效率. 通过这种分层管理, SAGE 能够缩短对外部数据源访问造成的延迟, 提高对外部知识与系统内部存储的交互效率.

3) 应用拓展. SAGE 不仅仅局限于生成内容, 还将检索增强技术扩展至更广泛的智能体应用领域, 从而在对生物人类的查询进行响应之外, 还可以对数字人以及实体机器人的行为决策提供支撑. 整体而言, SAGE 将感知、思考、决策以及行为等多样任务集成到统一的框架中, 提供一种类人思维与行为模式. 其中, RAP 将智能体所接收到的单模态或多模态信息在经过初步处理后, 与外部资源进行检索和对比, 从而实现对开放场景的准确感知. 例如, 借助 RAP 技术, 智能体可以根据看到的广告照片, 对其中的商品类别进行精准识别, 即使是在训练过程中, 智能体的感知网络并没有接触过这类商品数据. RAT 则将 CoT 等类人思考机制与 RAG 进行结合, 利用外部信息支撑智能体进行全面思考. RAD 针对特定的决策任务, 通过查询历史案例库信息获取过往的专家级经验数据, 从而为当前决策过程提供引导. RAA 过程则通过引入工具或功能模块的操控接口信息, 快速动态地赋予智能体以更多能力.

相比于 RAG 技术方法, SAGE 一方面拓展知识的边界, 将更为广泛的互联网知识通过搜索过程引入 LLM 的生成流程中. 同时, 基于内存和缓存的

知识管理提供更为高效的信息获取效率。另一方面,通过在感知、决策、思维和行动等更为普遍的任务中引入针对外部信息的增强机制, SAGE 为智能体的能力拓展提供新的思路。

5 总结与展望

本文总结了大模型在知识获取和更新方面存在的问题,围绕检索增强生成方法介绍其基本架构以及关键技术改进方法,并通过 LLM 和多模态大模型两个方向介绍其应用现状。检索增强生成技术的灵活性、可解释性、隐私安全性和成本优势,使其成为打造行业或私人专属大模型的有效支撑手段。为方便读者阅读,表 4 总结了本文出现的术语。未来,伴随着数字人和机器人技术的发展,检索增强生成方法有望进一步强化多元化场景下的专用智能提升,打造生物人、数字人及机器人协同的平行人体系,赋能智业时代的生产力提升。本节将对 RAG 技术的优势和存在的挑战以及 RAG 技术潜在研究方向展开分析。

5.1 优势分析

RAG 的主要优势归纳为五个方面,提升下游任务性能、知识接入灵活、可解释性高、迁移代价低和隐私可控性强。

1) 提升下游任务性能。使用 RAG 技术可以普遍提升下游自然语言处理任务的性能^[18]。Jiang 等^[86]提出一种基于 RAG 主动多次检索的 FLARE 方法,并在三项不同的任务及对应数据集——常识推理(数据集: StrategyQA^[163])、长篇问答(数据集: ASQA^[164]、ASQA-hint^[86])和开放域摘要(数据集: WikiAsp^[165])上评估所提方法相比无检索方法的性能提升。结果如表 5 所示,其中,EM、D-F1、R-L、DR、E-F1 分别为绝对匹配、消歧 F1、ROUGE-L、ROUGE 和基于实例的 F1。在所有任务中,相比于无检索基线方法,FLARE 均表现出优越或具有竞争力的性能,展示了 RAG 技术的有效性和普适性。本文进一步分析 RAG 技术在各种任务上对性能的提升,发现使用 RAG 在多跳问答任务上有最显著的提升。这主要是由于该任务具有清晰的定义和通

表 4 中英文术语对照表
Table 4 Glossary of Chinese-English terms

中文名称	英文名称
检索增强生成技术	Retrieval-augmented generation (RAG) ^[19-23]
大语言模型	Large language model (LLM) ^[5-8]
自然语言处理	Natural language processing (NLP)
计算机视觉	Computation vision (CV)
数据分块	Data chunking ^[38-41]
独热编码	One-hot encoding ^[49]
词袋模型	Bag of words (BOW) ^[50]
词频-逆向文件频率	Term frequency-inverse document frequency (TF-IDF) ^[51]
N 元模型	N -Gram ^[52]
海量文本语义向量基准测试	Massive text embedding benchmark (MTEB) ^[55]
退后提示	Step back prompting ^[81]
多路召回	Multi query retrieval ^[82]
假想文档嵌入	Hypothetical document embeddings (HyDE) ^[83-84]
外部知识视觉问答任务	Outside knowledge visual question answering (OKVQA) ^[110]
思维链	Chain of thought (CoT) ^[79]
搜索增强的生成与扩展技术	Search-augmented generation and extension (SAGE)

表 5 基于 RAG 的 FLARE 方法^[86]与无检索基线方法的实验结果对比
Table 5 Comparison of experimental results between the RAG-based FLARE method^[86] and the non-retrieval baseline method

指标	StrategyQA		ASQA			ASQA-hint				WikiAsp		
	EM	EM	D-F1	R-L	DR	EM	D-F1	R-L	DR	UniEval ^[166]	E-F1	R-L
无检索	72.9	33.8	24.2	33.3	28.4	40.1	32.5	36.4	34.4	47.1	14.1	26.4
FLARE	77.3	41.3	28.2	34.3	31.1	46.2	36.7	37.7	37.2	53.4	18.9	27.6

过推理过程生成最终答案的具体目标,这使得语言模型更容易生成相关的输出. ASQA 和 WikiAsp 更为开放,增加了生成和评估的难度. ASQA-hint 的提升幅度大于 ASQA,因为 ASQA-hint 为引导语言模型在生成答案时保持正确提供了一个关于问题哪部分信息模糊的简短提示.

2) 知识接入灵活. RAG 技术通过与外部数据源的实时交互,实现对知识的动态接入. 这种设计使得 RAG 能够根据具体任务的需求,灵活选择并融合适宜的知识库资源. 与传统的静态模型相比, RAG 能够进行灵活的能力切换,从而有效提升任务处理效率.

3) 可解释性高. RAG 方法通过检索机制对外部信息进行链接,这一步骤有助于降低生成不准确信息的风险. RAG 输出的答案具备较高的可追溯性和可解释性,用户可以通过审查答案的信息来源,验证其准确性,从而确保所获取信息的可靠性.

4) 迁移代价低. 在大型语言模型的应用中,领域迁移往往需要大量的存储和计算资源,以实现模型的重新训练或微调. RAG 模型通过直接更新知识库中的数据,实现知识的快速更新和领域迁移. 这种更新机制避免了对整个模型进行重新训练的需求,从而显著降低了领域迁移的成本和复杂性.

5) 隐私可控性强. RAG 系统将存储在数据库的知识和存储在模型参数中的知识进行显式分离,这为企业和个人应用提供了隐私保障的基础. 为获取具备领域特色的大模型系统,用户无需向大模型方案商提供私域数据,从而增强了对数据和隐私安全的控制.

5.2 挑战分析

RAG 目前的挑战归纳为三类,算法优化、适用性需求与应用范围拓展^[21].

1) 算法优化. 如第 2 节所述,检索和生成质量是 RAG 技术的关键,目前主要算法的优化在于知识库构建、检索和生成过程. 具体地,知识库构建时可采取提高数据粒度、优化索引结构、加入元数据等方式;检索时可优化查询,即使用户原始问题更清晰,更适合检索任务,常见方法包括查询转换和查询增强等. 检索时可优化索引,如采用对齐优化和混合检索等方法;生成优化可采用重排序、压缩上下文等优化检索到的相关上下文与查询的整合过程,以及微调 LLM.

2) 适用性需求. 噪声鲁棒性、负面拒绝、信息整合和反事实鲁棒性^[62, 167]等能力对于模型在各种挑战和复杂场景下的表现至关重要,影响最终的生成质量. 具体地,噪声鲁棒性评估模型处理与问题相

关但缺乏实质性信息的噪声文档的能力;负面拒绝评估模型在检索到的文档不包含回答问题所需知识时拒绝作答的辨别能力;信息整合评估模型从多个文档中综合信息以回答复杂问题的熟练程度;反事实鲁棒性测试模型识别并忽略文档中已知的不准确信息的能力. 其他适用性需求的指标还包括时延、多样性等. 时延衡量系统查找信息并响应的速度,对于用户体验至关重要. 多样性评估系统是否能检索到各种相关文档并生成多样化的响应.

3) 应用范围拓展. RAG 的应用范围拓展主要包括超长上下文、多模态和工具箱三个方面. LLM 上下文输入的拓展为 RAG 的发展提供了新机会,使其能够解决更多复杂问题和需要阅读大量材料才能回答的综合性问题^[168]. 在超长上下文背景下开发新的 RAG 方法是未来的研究趋势之一. 检索多模态信息以增强生成将成为另一个重要研究方向. 常见的模态包括图像^[169]、音视频和代码^[170]等,需要构建相应的多模态知识库、检索和生成策略. 这有助于更好地将生成回复植根于现实世界中,构建一个更加符合直觉、能更好地与世界互动的模型. 最后, RAG 生态系统的发展受到其技术栈进步的巨大影响. 随着 ChatGPT 的出现, LangChain 和 LlamaIndex 等关键工具迅速流行,提供了广泛的 RAG 相关 API,并在 LLM 领域中成为必不可少的工具. 随着 RAG 的发展, RAG 工具箱正汇聚成一个基础技术栈,为高级企业应用奠定基础. 然而,一个完全集成、综合的平台概念初具雏形,需要进一步的发展. 在个性化对话生成中,带有情感和风格的实例可能更理想;包含特定术语的平行数据在机器翻译中更有帮助,等等. 另外,使用通用的检索度量标准可能导致检索结果缺乏多样性. 收集多样化的检索结果集可以提高有用信息的覆盖范围. 因此,考虑多个不同的检索度量标准可能会在未来带来更高质量的生成.

其他的挑战还包括可控文本生成、知识异质化^[171]等. 可控的文本生成可以在个性化对话生成中生成带有情感和风格的内容;在专业领域机器翻译时使用专业术语. 这一功能的实现相比使用通用的词汇相似度进行检索,需要制定并使用多个不同维度的检索度量标准. 知识异质化指从异质知识源学习知识,以提高知识覆盖率并有更多空间检索合适的知识. 目前 RAG 的工作大多仅利用单一来源的同质知识检索空间,即维基百科段落. 然而,其性能可能受到同质知识源的限制. 例如,在许多开放域问答数据集中,只有有限部分的问题可以从维基百科段落中回答,而其余的问题只能依赖输入查询,因为无法检索到相关文档. 参考人类学习知识的过程,

一个直观的想法是将检索语料库从维基百科扩展到整个万维网. 具体研究方向包括将异质知识进行一致化表征, 在异质知识上进行虚拟多跳检索和基于结构化知识的检索文档推理等.

References

- Tian Yong-Lin, Wang Yu-Tong, Wang Jian-Gong, Wang Xiao, Wang Fei-Yue. Key problems and progress of vision transformers: The state of the art and prospects. *Acta Automatica Sinica*, 2022, **48**(4): 957–979
(田永林, 王雨桐, 王建功, 王晓, 王飞跃. 视觉 Transformer 研究的关键问题: 现状及展望. *自动化学报*, 2022, **48**(4): 957–979)
- Casper S, Davies X, Shi C, Gilbert T K, Scheurer J, Rando J, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.
- Croitoru F A, Hondru V, Ionescu R T, Shah M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(9): 10850–10869
- Muennighoff N, Rush A M, Barak B, le Scao T, Piktus A, Tazi N, et al. Scaling data-constrained language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM, 2023. Article No. 2191
- Wang Y T, Pan Y H, Yan M, Su Z, Luan T H. A survey on ChatGPT: AI-generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society*, 2023, **4**: 280–302
- Wang Xiao, Zhang Xiang-Yu, Zhou Rui, Tian Yong-Lin, Wang Jian-Gong, Chen Long, et al. An intelligent architecture for cognitive autonomous driving based on parallel testing. *Acta Automatica Sinica*, 2024, **50**(2): 356–371
(王晓, 张翔宇, 周锐, 田永林, 王建功, 陈龙, 等. 基于平行测试的认知自动驾驶智能架构研究. *自动化学报*, 2024, **50**(2): 356–371)
- Fan L L, Guo C, Tian Y L, Zhang H, Zhang J, Wang F Y. Sora for foundation robots with parallel intelligence: Three world models, three robotic systems. *Frontiers of Information Technology and Electronic Engineering*, 2024, **25**(7): 917–923
- Moor M, Banerjee O, Abad Z S H, Krumholz H M, Leskovec J, Topol E J, et al. Foundation models for generalist medical artificial intelligence. *Nature*, 2023, **616**(7956): 259–265
- Lu Jing-Wei, Guo Chao, Dai Xing-Yuan, Miao Qing-Hai, Wang Xing-Xia, Yang Jing, et al. The ChatGPT after: Opportunities and challenges of very large scale pre-trained models. *Acta Automatica Sinica*, 2023, **49**(4): 705–717
(卢经纬, 郭超, 戴星原, 缪青海, 王兴霞, 杨静, 等. 问答 ChatGPT 之后: 超大预训练模型的机遇和挑战. *自动化学报*, 2023, **49**(4): 705–717)
- Currie G M. Academic integrity and artificial intelligence: Is ChatGPT hype, hero or heresy? *Seminars in Nuclear Medicine*, 2023, **53**(5): 719–730
- Hirano Y, Hanaoka S, Nakao T, Miki S, Kikuchi T, Nakamura Y, et al. GPT-4 Turbo with vision fails to outperform text-only GPT-4 Turbo in the Japan diagnostic radiology board examination. *Japanese Journal of Radiology*, 2024, **42**(8): 918–926
- Ding Y J, Fan W Q, Ning L B, Wang S J, Li H Y, Yin D W, et al. A survey on RAG meets LLMs: Towards retrieval-augmented large language models. arXiv preprint arXiv: 2405.06211, 2024.
- Xiong G Z, Jin Q, Lu Z Y, Zhang A D. Benchmarking retrieval-augmented generation for medicine. In: Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: ACL, 2024. 6233–6251
- Zhao A, Huang D, Xu Q, Lin M, Liu Y J, Huang G. ExpeL: LLM agents are experiential learners. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 19632–19642
- Zhai Y X, Tong S B, Li X, Cai M, Qu Q, Lee Y J, et al. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In: Proceedings of the Conference on Parsimony and Learning. Hong Kong, China: PMLR, 2024. 202–227
- Gupta S, Jegelka S, Lopez-Paz D, Ahuja K. Context is environment. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024.
- Ji Z W, Yu T Z, Xu Y, Lee N, Ishii E, Fung P. Towards mitigating LLM hallucination via self reflection. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore, Singapore: ACL, 2023. 1827–1843
- Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM, 2020. Article No. 793
- Huang Y Z, Huang J. A survey on retrieval-augmented text generation for large language models. arXiv preprint arXiv: 2404.10981, 2024.
- Zhu Y T, Yuan H Y, Wang S T, Liu J N, Liu W H, Deng C L, et al. Large language models for information retrieval: A survey. arXiv preprint arXiv: 2308.07107, 2023.
- Gao Y F, Xiong Y, Gao X Y, Jia K X, Pan J L, Bi Y X, et al. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv: 2312.10997, 2023.
- Li H Y, Su Y X, Cai D, Wang Y, Liu L M. A survey on retrieval-augmented text generation. arXiv preprint arXiv: 2202.01110, 2022.
- Hu Y C, Lu Y X. RAG and RAU: A survey on retrieval-augmented language model in natural language processing. arXiv preprint arXiv: 2404.19543, 2024.
- Tian Yong-Lin, Wang Xing-Xia, Wang Yu-Tong, Wang Jian-Gong, Guo Chao, Fan Li-Li, et al. RAG-PHI: RAG-driven parallel human and parallel intelligence. *Chinese Journal of Intelligent Science and Technology*, 2024, **6**(1): 41–51
(田永林, 王兴霞, 王雨桐, 王建功, 郭超, 范丽丽, 等. RAG-PHI: 检索增强生成驱动的平行人与平行智能. *智能科学与技术学报*, 2024, **6**(1): 41–51)
- Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. arXiv preprint arXiv: 2307.10169, 2023.
- Dai X Y, Guo C, Tang Y, Li H C, Wang Y T, Huang J, et al. VistaRAG: Toward safe and trustworthy autonomous driving through retrieval-augmented generation. *IEEE Transactions on Intelligent Vehicles*, 2024, **9**(4): 4579–4582
- Dave T, Athaluri S A, Singh S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence*, 2023, **6**: Article No. 1169595
- Louis A, van Dijk G, Spanakis G. Interpretable long-form legal question answering with retrieval-augmented large language models. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 22266–22275
- Wu S J, Irsoy O, Lu S, Dabrovolski V, Dredze M, Gehrmann S, et al. BloombergGPT: A large language model for finance. arXiv preprint arXiv: 2303.17564, 2023.
- Wang Fei-Yue, Wang Yan-Fen, Chen Yi-Zhu, Tian Yong-Lin, Qi Hong-Wei, Wang Xiao, et al. Federated ecology: From federated data to federated intelligence. *Chinese Journal of Intelligent Science and Technology*, 2020, **2**(4): 305–311
(王飞跃, 王艳芬, 陈慧竹, 田永林, 齐红威, 王晓, 等. 联邦生态: 从联邦数据到联邦智能. *智能科学与技术学报*, 2020, **2**(4): 305–311)
- Gemini Team Google. Gemini: A family of highly capable multimodal models. arXiv preprint arXiv: 2312.11805, 2023.
- Lewis M, Liu Y H, Goyal N, Ghazvininejad M, Mohamed A,

- Levy O, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Washington, USA: ACL, 2020. 7871–7880
- 33 Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: ACL, 2019. 4171–4186
- 34 Wang Y X, Sun Q X. M3E: Moka massive mixed embedding model [Online], available: <https://github.com/wangyuxinwhy/uniem>. December 31, 2023
- 35 Neelakantan A, Xu T, Puri R, Radford A, Han J M, Tworek J, et al. Text and code embeddings by contrastive pre-training. arXiv preprint arXiv: 2201.10005, 2022.
- 36 Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense passage retrieval for open-domain question answering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Virtual Event: ACL, 2020. 6769–6781
- 37 Yang Z L, Qi P, Zhang S Z, Bengio Y, Cohen W, Salakhutdinov R, et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2018. 2369–2380
- 38 Yu H, Gan A R, Zhang K, Tong S W, Liu Q, Liu Z F. Evaluation of retrieval-augmented generation: A survey. In: Proceedings of the CCF Conference on Big Data. Singapore, Singapore: Springer Nature, 2024. 102–120
- 39 Chen H Y, Yu H. Intent-based web page summarization with structure-aware chunking and generative language models. In: Proceedings of the ACM Web Conference 2023. Austin, USA: ACM, 2023. 310–313
- 40 Xiao S T, Liu Z, Zhang P T, Muennighof N. C-pack: Packaged resources to advance general Chinese embedding. arXiv preprint arXiv: 2309.07597, 2023.
- 41 Chen T, Wang H W, Chen S H, Yu W H, Ma K X, Zhao X R, et al. Dense X retrieval: What retrieval granularity should we use? In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami, USA: ACL, 2024. 15159–15177
- 42 Chung H W, Hou L, Longpre S, Zoph B, Tai Y, Fedus W, et al. Scaling instruction-finetuned language models. *The Journal of Machine Learning Research*, 2024, **25**(1): Article No. 70
- 43 Manning C D, Schütze H. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press, 1999.
- 44 Zhang T, Damerou F, Johnson D. Text chunking based on a generalization of winnow. *Journal of Machine Learning Research*, 2002, **2**(3): 615–637
- 45 Barzilay R, Elhadad M. Using lexical chains for text summarization. In: Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization. Madrid, Spain: ACL, 1997. 10–17
- 46 Moens M F, Uyttendaele C, Dumortier J. Information extraction from legal texts: The potential of discourse analysis. *International Journal of Human-Computer Studies*, 1999, **51**(6): 1155–1171
- 47 Lin C Y. ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the Text Summarization Branches Out. Barcelona, Spain: ACL, 2004. 74–81
- 48 Wan X J. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu, USA: ACL, 2008. 553–561
- 49 Rodríguez P, Bautista M A, González J, Escalera S. Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing*, 2018, **75**: 21–31
- 50 Zhang Y, Jin R, Zhou Z H. Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 2010, **1**(1–4): 43–52
- 51 Chowdhury G. *Introduction to Modern Information Retrieval* (3rd edition). London: Facet Publishing, 2010.
- 52 Kondrak G. N-gram similarity and distance. In: Proceedings of the 12th International Symposium on String Processing and Information Retrieval. Buenos Aires, Argentina: Springer, 2005. 115–126
- 53 Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the 1st International Conference on Learning Representations. Scottsdale, USA: ICLR, 2013.
- 54 Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: ACL, 2014. 1532–1543
- 55 Muennighoff N, Tazi N, Magne L, Reimers N. MTEB: Massive text embedding benchmark. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia: ACL, 2023. 2014–2037
- 56 Meng R, Liu Y, Joty S, Xiong C M, Zhou Y B, Yavuz S. SFR-embedding-mistral: Enhance text retrieval with transfer learning [Online], available: <https://www.salesforce.com/blog/sfr-embedding/#author-section>, October 28, 2024
- 57 Muennighoff N, Su H J, Wang L, Yang N, Wei F R, Yu T, et al. Generative representational instruction tuning. arXiv preprint arXiv: 2402.09906, 2024.
- 58 Wang L, Yang N, Huang X L, Yang L J, Majumder R, Wei F R. Improving text embeddings with large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand: ACL, 2024. 11897–11916
- 59 Yang A Y, Xiao B, Wang B N, Zhang B R, Bian C, Yin C, et al. Baichuan 2: Open large-scale language models. arXiv preprint arXiv: 2309.10305, 2023.
- 60 Chen J L, Xiao S T, Zhang P T, Luo K, Lian D F, Liu Z. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv: 2402.03216, 2024.
- 61 Xiao S T, Liu Z, Zhang P T, Muennighoff N, Lian D F, Nie J Y. C-pack: Packed resources for general Chinese embedding. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington, USA: ACM, 2024. 641–649
- 62 Chen J W, Lin H Y, Han X P, Sun L. Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI, 2024. 17754–17762
- 63 Luo K, Liu Z, Xiao S T, Liu K. BGE landmark embedding: A chunking-free embedding method for retrieval augmented long-context large language models. arXiv preprint arXiv: 2402.11573, 2024.
- 64 Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv: 2302.13971, 2023.
- 65 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770–778
- 66 Tan M X, Le Q V. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 6105–6114
- 67 Wan X J. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In: Proceedings of the Conference on Empirical methods in Natural language Pro-

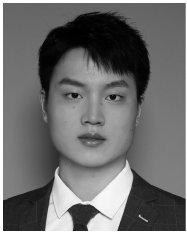
- cessing. Virtual Event: ACL, 2008.
- 68 Tsalera E, Papadakis A, Samarakou M. Comparison of pre-trained CNNs for audio classification using transfer learning. *Journal of Sensor and Actuator Networks*, 2021, **10**(4): Article No. 72
- 69 Yuan Y, Xun G X, Suo Q L, Jia K B, Zhang A D. Wave2Vec: Learning deep representations for biosignals. In: Proceedings of the IEEE International Conference on Data Mining (ICDM). New Orleans, USA: IEEE, 2017. 1159–1164
- 70 Lin G H, Zhang Y M, Xu G, Zhang Q X. Smoke detection on video sequences using 3D convolutional neural networks. *Fire Technology*, 2019, **55**(5): 1827–1847
- 71 Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: ICML, 2021. 813–824
- 72 Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends? In Information Retrieval*, 2009, **3**(4): 333–389
- 73 McCandless M, Hatcher E, Gospodnetic O. *Lucene in Action* (2nd edition). Greenwich: Manning Publications, 2010.
- 74 Gormley C, Tong Z. *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*. Sebastopol: O'Reilly Media, Inc., 2015.
- 75 Chang X Y. The analysis of open source search engines. *Highlights in Science, Engineering and Technology*, 2023, **32**: 32–42
- 76 Cuconasu F, Trappolini G, Siciliano F, Filice S, Campagnano C, Maarek Y, et al. The power of noise: Redefining retrieval for RAG systems. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington, USA: ACM, 2024. 719–729
- 77 Singh J, Prasad M, Prasad O K, Meng Joo E, Saxena A K, Lin C T. A novel fuzzy logic model for pseudo-relevance feedback-based query expansion. *International Journal of Fuzzy Systems*, 2016, **18**(6): 980–989
- 78 Kim L, Yahia E, Segonds F, Véron P, Mallet A. i-Dataquest: A heterogeneous information retrieval tool using data graph for the manufacturing industry. *Computers in Industry*, 2021, **132**: Article No. 103527
- 79 Wei J, Wang X Z, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM, 2022. Article No. 1800
- 80 Ma X B, Gong Y Y, He P C, Zhao H, Duan N. Query rewriting in retrieval-augmented large language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore, Singapore: ACL, 2023. 5303–5315
- 81 Zheng H S, Mishra S, Chen X Y, Cheng H T. Take a step back: Evoking reasoning via abstraction in large language models. arXiv preprint arXiv: 2310.06117, 2023.
- 82 Wang Z Y, Wu Y, Narasimhan K, Russakovsky O. Multi-query video retrieval. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 233–249
- 83 Gao L Y, Ma X G, Lin J, Callan J. Precise zero-shot dense retrieval without relevance labels. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: ACL, 2023. 1762–1777
- 84 Wang L, Yang N, Wei F R. Query2doc: Query expansion with large language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore, Singapore: ACL, 2023. 9414–9423
- 85 Asai A, Wu Z Q, Wang Y Z, Sil A, Hajishirzi H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024.
- 86 Jiang Z B, Xu F, Gao L Y, Sun Z Q, Liu Q, Dwivedi-Yu J, et al. Active retrieval augmented generation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore, Singapore: ACL, 2023. 7969–7992
- 87 Wang H R, Huang W Y, Deng Y, Wang R, Wang Z Z, Wang Y F, et al. UniMS-RAG: A unified multi-source retrieval-augmented generation for personalized dialogue systems. arXiv preprint arXiv: 2401.13256, 2024.
- 88 Yuan Z Q, Zhang W K, Tian C Y, Mao Y Q, Zhou R X, Wang H Q, et al. MCRN: A multi-source cross-modal retrieval network for remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 2022, **115**: Article No. 103071
- 89 Bouchakwa M, Ayadi Y, Amous I. Multi-level diversification approach of semantic-based image retrieval results. *Progress in Artificial Intelligence*, 2020, **9**(1): 1–30
- 90 Li W H, Yang S, Wang Y, Song D, Li X Y. Multi-level similarity learning for image-text retrieval. *Information Processing and Management*, 2021, **58**(1): Article No. 102432
- 91 Jeong S, Baek J, Cho S, Hwang S J, Park J. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Mexico City, Mexico: ACL, 2024. 7036–7050
- 92 Malkov Y A, Yashunin D A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **42**(4): 824–836
- 93 Li W T, Li J K, Ma W Z, Liu Y. Citation-enhanced generation for LLM-based Chatbots. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand: ACL, 2024. 1451–1466
- 94 Glass M, Rossiello G, Chowdhury M F M, Naik A, Cai P S, Gliozzo A. Re2G: Retrieve, rerank, generate. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA: ACL, 2022. 2701–2715
- 95 Wang S T, Xu X, Wang M, Chen W P, Zhu Y T, Dou Z C. RichRAG: Crafting rich responses for multi-faceted queries in retrieval-augmented generation. In: Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: ACL, 2025. 11317–11333
- 96 Xu Z P, Liu Z H, Liu Y B, Xiong C Y, Yan Y K, Wang S, et al. ActiveRAG: Revealing the treasures of knowledge via active learning. arXiv preprint arXiv: 2402.13547, 2024.
- 97 Izacard G, Grave E. Leveraging passage retrieval with generative models for open domain question answering. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Virtual Event: ACL, 2021. 874–880
- 98 Zhang J, Wang X, Zhang H Y, Sun H L, Liu X D. Retrieval-based neural source code summarization. In: Proceedings of the 42nd ACM/IEEE International Conference on Software Engineering. Seoul, South Korea: ACM, 2020. 1385–1397
- 99 Khandelwal U, Levy O, Jurafsky D, Zettlemoyer L, Lewis M. Generalization through memorization: Nearest neighbor language models. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR, 2020.
- 100 Poesia G, Polozov A, Le V, Tiwari A, Soares G, Meek C, et al. Synchronesh: Reliable code generation from pre-trained language models. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022.
- 101 Joshi H, Sanchez J C, Gulwani S, Le V, Verbruggen G, Radiček I. Repair is nearly generation: Multilingual program repair with LLMs. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI, 2023. 5131–5140

- 102 Zheng L M, Chiang W L, Sheng Y, Li T L, Zhuang S Y, Wu Z H, et al. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024.
- 103 Zhu Y H, Ren C Y, Xie S Y, Liu S K, Ji H Y, Wang Z X, et al. REALM: RAG-driven enhancement of multimodal electronic health records analysis via large language models. arXiv preprint arXiv: 2402.07016, 2024.
- 104 Shi Y X, Zi X, Shi Z J, Zhang H M, Wu Q, Xu M. ERAGent: Enhancing retrieval-augmented language models with improved accuracy, efficiency, and personalization. arXiv preprint arXiv: 2405.06683, 2024.
- 105 Jiang X K, Zhang R Z, Xu Y X, Qiu R H, Fang Y, Wang Z Y, et al. Think and retrieval: A hypothesis knowledge graph enhanced medical large language models. arXiv preprint arXiv: 2312.15883, 2023.
- 106 Wiratunga N, Abeyratne R, Jayawardena L, Martin K, Massie S, Nkisi-Orji I, et al. CBR-RAG: Case-based reasoning for retrieval augmented generation in LLMs for legal question answering. In: Proceedings of the 32nd International Conference on Case-Based Reasoning Research and Development. Merida, Mexico: Springer, 2024. 445–460
- 107 Salemi A, Zamani H. Towards a search engine for machines: Unified ranking for multiple retrieval-augmented large language models. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington, USA: ACM, 2024. 741–751
- 108 Komeili M, Shuster K, Weston J. Internet-augmented dialogue generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: ACL, 2022. 8460–8478
- 109 Lin W Z, Byrne B. Retrieval augmented visual question answering with outside knowledge. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: ACL, 2022. 11238–11254
- 110 Gui L K, Wang B R, Huang Q Y, Hauptmann A, Bisk Y, Gao J F. KAT: A knowledge augmented transformer for vision-and-language. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA: ACL, 2022. 956–968
- 111 Tiong A M H, Li J N, Li B Y, Savarese S, Hoi S C H. Plug-and-play VQA: Zero-shot VQA by conjoining large pretrained models with zero training. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: ACL, 2022. 951–967
- 112 Chen W H, Hu H X, Chen X, Verga P, Cohen W. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: ACL, 2022. 5558–5570
- 113 Yasunaga M, Aghajanyan A, Shi W J, James R, Leskovec J, Liang P, et al. Retrieval-augmented multimodal language modeling. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: ACM, 2022. Article No. 1659
- 114 Chen W H, Hu H X, Saharia C, Cohen W W. Re-Imagen: Retrieval-augmented text-to-image generator. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2023.
- 115 Long A, Yin W, Ajanthan T, Nguyen V, Purkait P, Garg R, et al. Retrieval augmented classification for long-tail visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 6949–6959
- 116 Zhao J, Haffar G, Shareghi E. Generating synthetic speech from spokenvocab for speech translation. In: Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023. Dubrovnik, Croatia: ACL, 2023. 1975–1981
- 117 Huang R J, Huang J W, Yang D C, Ren Y, Liu L P, Li M Z, et al. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: ICML, 2023. 13916–13932
- 118 Yuan J H, Sun S Y, Omeiza D, Zhao B, Newman P, Kunze L, et al. RAG-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. arXiv preprint arXiv: 2402.10828, 2024.
- 119 Hussien M M, Melo A N, Ballardini A L, Maldonado C S, Izquierdo R, Sotelo M Á. RAG-based explainable prediction of road users behaviors for automated driving using knowledge graphs and large language models. *Expert Systems With Applications*, 2025, **265**: Article No. 125914
- 120 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: ACM, 2017. 6000–6010
- 121 Chen M, Tworek J, Jun H, Yuan Q M, de Oliveira Pinto H P, Kaplan J, et al. Evaluating large language models trained on code. arXiv preprint arXiv: 2107.03374, 2021.
- 122 Ziegler A, Kalliamvakou E, Li X A, Rice A, Rifkin D, Simister S, et al. Productivity assessment of neural code completion. In: Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming. San Diego, USA: ACM, 2022. 21–29
- 123 Li Y J, Choi D, Chung J, Kushman N, Schrittwieser J, Leblond R, et al. Competition-level code generation with AlphaCode. *Science*, 2022, **378**(6624): 1092–1097
- 124 Nijkamp E, Pang B, Hayashi H, Tu L F, Wang H, Zhou Y B, et al. A conversational paradigm for program synthesis. arXiv preprint arXiv: 2203.13474, 2022.
- 125 Fried D, Aghajanyan A, Lin J, Wang S D, Wallace E, Shi F, et al. InCoder: A generative model for code infilling and synthesis. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2023.
- 126 Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling language modeling with pathways. *The Journal of Machine Learning Research*, 2023, **24**(1): Article No. 240
- 127 Zheng Q K, Xia X, Zou X, Dong Y X, Wang S, Xue Y F, et al. CodeGeeX: A pre-trained model for code generation with multilingual benchmarking on HumanEval-X. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Long Beach, USA: ACM, 2023. 5673–5684
- 128 Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 618–626
- 129 Yuan Z, Xi Q, Tan C Q, Zhao Z Y, Yuan H Y, Huang F, et al. RAMM: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM, 2023. 547–556
- 130 Zhou Y C, Long G D. Style-aware contrastive learning for multi-style image captioning. In: Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023. Dubrovnik, Croatia: ACL, 2023. 2257–2267
- 131 Shen S, Li C Y, Hu X W, Yang J W, Xie Y J, Zhang P C, et al. K-LITE: Learning transferable visual models with external knowledge. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM, 2022. Article No. 1132
- 132 Fellbaum C. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, 1998.
- 133 Zesch T, Müller C, Gurevych I. Using wiktionary for comput-

- ing semantic relatedness. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence. Chicago, USA: AAAI, 2008. 861–866
- 134 Liu H T, Son K, Yang J W, Liu C, Gao J F, Lee Y J, et al. Learning customized visual models with retrieval-augmented knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 15148–15158
- 135 Whitehead S, Ji H, Bansal M, Chang S F, Voss C. Incorporating background knowledge into video description generation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2018. 3992–4001
- 136 Le H, Chen N, Hoi S. Vgmmn: Video-grounded neural module networks for video-grounded dialogue systems. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA: ACL, 2022. 3377–3393
- 137 Kim M, Sung-Bin K, Oh T H. Prefix tuning for automated audio captioning. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023. 1–5
- 138 Mestre R, Middleton S E, Ryan M, Gheasi M, Norman T, Zhu J T. Augmenting pre-trained language models with audio feature embedding for argumentation mining in political debates. In: Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023. Dubrovnik, Croatia: ACL, 2023. 274–288
- 139 Shu Y H, Yu Z W, Li Y H, Karlsson B, Ma T T, Qu Y Z, et al. TIARA: Multi-grained retrieval for robust question answering over large knowledge base. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: ACL, 2022. 8108–8121
- 140 Pan F F, Canim M, Glass M, Gliozzo A, Fox P. CLTR: An end-to-end, transformer-based system for cell-level table retrieval and table question answering. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Bangkok, Thailand: ACL, 2021. 202–209
- 141 Yang Z C, Qin J H, Chen J Q, Lin L, Liang X D. LogicSolver: Towards interpretable math word problem solving with logical prompt-enhanced learning. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: ACL, 2022. 1–13
- 142 He H F, Zhang H M, Roth D. Rethinking with retrieval: Faithful large language model inference. arXiv preprint arXiv: 2301.00303, 2022.
- 143 Li X X, Zhao R C, Chia Y K, Ding B S, Bing L D, Joty S, et al. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. arXiv preprint arXiv: 2305.13269, 2023.
- 144 Zhou H J, Gu B Y, Zou X Y, Li Y R, Chen S S, Zhou P L, et al. A survey of large language models in medicine: Progress, application, and challenge. arXiv preprint arXiv: 2311.05112, 2023.
- 145 Kang B, Kim J, Yun T R, Kim C E. Prompt-RAG: Pioneering vector embedding-free retrieval-augmented generation in niche domains, exemplified by Korean medicine. arXiv preprint arXiv: 2401.11246, 2024.
- 146 Quidwai M A, Lagana A. A RAG chatbot for precision medicine of multiple myeloma. *medRxiv*, DOI: [10.1101/2024.03.14.24304293](https://doi.org/10.1101/2024.03.14.24304293)
- 147 Kim J, Min M. From RAG to QA-RAG: Integrating generative AI for pharmaceutical regulatory compliance process. arXiv preprint arXiv: 2402.01717, 2024.
- 148 Rafat M I. AI-powered Legal Virtual Assistant: Utilizing RAG-optimized LLM for Housing Dispute Resolution in Finland [Master thesis], Haaga-Helia University of Applied Sciences, Finland, 2024.
- 149 Li Y H, Wang S F, Ding H, Chen H. Large language models in finance: A survey. In: Proceedings of the 4th ACM International Conference on AI in Finance. Brooklyn, USA: ACM, 2023. 374–382
- 150 Ryu C, Lee S, Pang S, Choi C, Choi H, Min M, et al. Retrieval-based evaluation for LLMs: A case study in Korean legal QA. In: Proceedings of the Natural Legal Language Processing Workshop 2023. Singapore, Singapore: ACL, 2023. 132–137
- 151 Cui C, Ma Y S, Cao X, Ye W Q, Zhou Y, Liang K Z, et al. A survey on multimodal large language models for autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Waikoloa, USA: IEEE, 2024. 958–979
- 152 Kim J, Rohrbach A, Darrell T, Canny J, Akata Z. Textual explanations for self-driving vehicles. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 577–593
- 153 Papineni K, Roukos S, Ward T, Zhu W J. Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: ACL, 2002. 311–318
- 154 Elliott D, Keller F. Image description using visual dependency representations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Seattle, USA: ACL, 2013. 1292–1302
- 155 Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 4566–4575
- 156 Xu Z H, Zhang Y J, Xie E Z, Zhao Z, Guo Y, Wong K Y K, et al. DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024, **9**(10): 8186–8193
- 157 Bornea A L, Ayed F, de Domenico A, Piovesan N, Maatouk A. Telco-RAG: Navigating the challenges of retrieval augmented language models for telecommunications. In: Proceedings of the IEEE Global Communications Conference. Cape Town, South Africa: IEEE, 2024. 2359–2364
- 158 Gaddala V S. Unleashing the power of generative AI and RAG agents in supply chain management: A futuristic perspective. *IRE Journals*, 2023, **6**(12): 1411–1417
- 159 Gupta A, Shirgaonkar A, de Luis Balaguer A, Silva B, Holstein D, Li D W, et al. RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. arXiv preprint arXiv: 2401.08406, 2024.
- 160 Wang F Y. Foundation worlds for parallel intelligence: From foundation/infrastructure models to foundation/infrastructure intelligence. *Alfred North Whitehead Laureate Lectures*. Beijing: 2021.
- 161 Packer C, Fang V, Patil S G, Lin K, Wooders S, Gonzalez J E. MemGPT: Towards LLMs as operating systems. arXiv preprint arXiv: 2310.08560, 2023.
- 162 Pouplin T, Sun H, Holt S, van der Schaar M. Retrieval-augmented thought process as sequential decision making. arXiv preprint arXiv: 2402.07812, 2024.
- 163 Geva M, Khashabi D, Segal E, Khot T, Roth D, Berant J. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 2021, **9**: 346–361
- 164 Stelmakh I, Luan Y, Dhingra B, Chang M W. ASQA: Factoid questions meet long-form answers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: ACL, 2022. 8273–8288
- 165 Hayashi H, Budania P, Wang P, Ackerson C, Neervannan R, Neubig G. WikiAsp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 2021, **9**: 211–225
- 166 Zhong M, Liu Y, Yin D, Mao Y N, Jiao Y Z, Liu P F, et al. Towards a unified multi-dimensional evaluator for text genera-

tion. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: ACL, 2022. 2023–2038

- 167 Yu H, Gan A, Zhang K, Tong S W, Liu Q, Liu Z F. Evaluation of retrieval-augmented generation: A survey. arXiv preprint arXiv: 2405.07437, 2024.
- 168 Chen H T, Xu F Y, Arora S, Choi E. Understanding retrieval augmentation for long-form question answering. arXiv preprint arXiv: 2310.12150, 2023.
- 169 Chen W H, Hu H X, Saharia C, Cohen W W. Re-imagen: Retrieval-augmented text-to-image generator. arXiv preprint arXiv: 2209.14491, 2022.
- 170 Nashid N, Sintaha M, Mesbah A. Retrieval-based prompt selection for code-related few-shot learning. In: Proceedings of the 45th International Conference on Software Engineering (ICSE). Melbourne, Australia: IEEE, 2023. 2450–2462
- 171 Yu W H. Retrieval-augmented generation across heterogeneous knowledge. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. Washington, USA: ACL, 2022. 52–58



田永林 中国科学院自动化研究所多模态人工智能系统全国重点实验室助理研究员。2022 年获得中国科学技术大学控制科学与工程专业博士学位。主要研究方向为平行智能, 自动驾驶, 智能交通系统。

E-mail: yonglin.tian@ia.ac.cn

(**TIAN Yong-Lin** Assistant researcher at the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree in control science and engineering from University of Science and Technology of China in 2022. His research interest covers parallel intelligence, autonomous driving, and intelligent transportation systems.)



王雨桐 中国科学院自动化研究所多模态人工智能系统全国重点实验室副研究员。2021 年获得中国科学院大学控制理论与控制工程专业博士学位。主要研究方向为计算机视觉, 智能感知。E-mail: yutong.wang@ia.ac.cn

(**WANG Yu-Tong** Associate re-

searcher at the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. She received her Ph.D. degree in control theory and control engineering from University of Chinese Academy of Sciences in 2021. Her research interest covers computer vision and intelligent perception.)



王兴霞 中国科学院自动化研究所多模态人工智能系统全国重点实验室博士研究生。2021 年获得南开大学工学硕士学位。主要研究方向为平行智能, 平行油田, 多智能体系统。

E-mail: wangxingxia2022@ia.ac.cn

(**WANG Xing-Xia** Ph.D. candi-

date at the State Key Laboratory for Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. She received her master degree in engineering from Nankai University in 2021. Her research interest covers parallel control, parallel oilfields, and multi-agent systems.)



杨 静 中国科学院自动化研究所多模态人工智能系统全国重点实验室博士研究生。2020 年获得北京化工大学自动化专业学士学位。主要研究方向为众包, 平行制造, 社会制造, 预训练语言模型和社会物理信息系统。

E-mail: yangjing2020@ia.ac.cn

(**YANG Jing** Ph.D. candidate at the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. She received her bachelor degree in automation from Beijing University of Chemical Technology in 2020. Her research interest covers crowdsourcing, parallel manufacturing, social manufacturing, pre-trained language models, and cyber-physical-social systems.)



沈甜雨 北京化工大学信息科学与技术学院副教授。2021 年获得中国科学院自动化研究所博士学位。主要研究方向为智能感知与智能无人系统。

E-mail: tianyu.shen@buct.edu.cn

(**SHEN Tian-Yu** Associate profes-

sor at the College of Information Science and Technology, Beijing University of Chemical Technology. She received her Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2021. Her research interest covers intelligent perception and intelligent unmanned systems.)



王建功 中国航空系统工程研究所工程师。2023 年获得中国科学院自动化研究所博士学位。主要研究方向为大模型, 计算机视觉, 航空工程。

E-mail: wangjg055@avic.com

(**WANG Jian-Gong** Engineer at

Aviation System Engineering Institute of China. He received his Ph.D. degree from the

Institute of Automation, Chinese Academy of Sciences in 2023. His research interest covers large models, computer vision, and aeronautical engineering.)



范丽丽 北京理工大学信息与电子学院博士后. 2022 年获得吉林大学博士学位. 主要研究方向为计算机视觉, 跨模态感知与理解, 类脑认知与决策. E-mail: lilifan@bit.edu.cn

(FAN Li-Li Postdoctor at the School of Information and Electronics, Beijing Institute of Technology. She received her Ph.D. degree from Jilin University in 2022. Her research interest covers computer vision, cross-modal perception and understanding, and neuromorphic cognition and decision-making.)



郭超 中国科学院自动化研究所助理研究员. 主要研究方向为人工智能艺术创作, 人机协作, 智能机器人系统, 机器学习, 强化学习. E-mail: chao.guo@ia.ac.cn

(GUO Chao Assistant professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers AI for art creation, human-machine collaboration, intelligent robotic systems, machine learning, and reinforcement learning.)



王寿文 澳门科技大学创新工程学院智能科学与系统专业博士研究生. 主要研究方向为智能系统和复杂系统的建模、分析与控制. E-mail: 2109853pmi3004@student.must.edu.mo

(WANG Shou-Wen Ph.D. candidate at the Faculty of Innovation Engineering, Macau University of Science and Technology. His research interest covers modeling, analysis and control of intelligent systems and complex systems.)



赵勇 国防科技大学系统工程学院博士研究生. 2021 年获得国防科技大学控制科学与工程专业硕士学位. 主要研究方向为群智感知和人机交互. E-mail: zhaoyong15@nudt.edu.cn

(ZHAO Yong Ph.D. candidate at the College of Systems Engineering, National University of Defense Technology. He received his master degree in control science and engineering from National University of Defense Technology in 2021. His research interest covers crowdsensing and human-computer interaction.)



武万森 国防科技大学系统工程学院博士研究生. 2018 年获得国防科技大学学士学位. 主要研究方向为视觉语言多模态, 机器人. E-mail: wuwansen14@nudt.edu.cn

(WU Wan-Sen Ph.D. candidate at the College of Systems Engineering, National University of Defense Technology. He received his bachelor degree from National University of Defense Technology in 2018. His research interest covers vision-and-language multi-modality and robot.)



王飞跃 中国科学院自动化研究所复杂系统管理与控制国家重点实验室研究员. 主要研究方向为智能系统和复杂系统的建模、分析与控制. 本文通信作者. E-mail: feiyue.wang@ia.ac.cn

(WANG Fei-Yue Professor at the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interest covers modeling, analysis, and control of intelligent systems and complex systems. Corresponding author of this paper.)