浅谈大模型时代下的检索增强：发展趋势、挑战与展望

冯掌印[1], **朱坤**[1], **马伟涛**[1], **黄磊**[1], **秦兵**[1,2], **刘挺**[1,2], **冯骁骋**[1,2*]
[1]哈尔滨工业大学，社会计算与信息检索研究中心，哈尔滨，150006
[2] 鹏城实验室，深圳，518055
{zyfeng,kzhu,wtma,lhuang,qinb,tliu,xcfeng*}@ir.hit.edu.cn

摘要

大型语言模型(LLM) 在各种自然语言任务上表现出了卓越的性能，但它们很容易受到过时数据和特定领域限制的影响。为了应对这些挑战，研究人员整合不同来源的外部信息来增强大语言模型，具体方法如检索增强等。在本文中，我们综合讨论了检索增强技术的发展趋势，包括检索时机规划、检索技术、以及检索结果的利用。此外，我们介绍了当前可用于检索增强任务的数据集和评价方法，并指出了应用和潜在研究方向。我们希望这项综述能够为社区提供对该研究领域的快速了解和全面概述，以启发未来的研究工作。

关键词： 检索增强

Enhancing Large Language Models with Retrieval-Augmented Techniques: Trends, Challenges, and Prospects

Zhangyin Feng[1], **Kun Zhu**[1], **Weitao Ma**[1], **Lei Huang**[1],
Bing Qin[1,2], **Ting Liu**[1,2], **Xiaocheng Feng**[1,2*]
[1]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, 150006
[2] Peng Cheng Laboratory, Shenzhen, 518055
{zyfeng,kzhu,wtma,lhuang,qinb,tliu,xcfeng*}@ir.hit.edu.cn

Abstract

Large language models (LLMs) have demonstrated outstanding performance in various natural language tasks, but they are susceptible to issues caused by outdated data and specific domain limitations. To address these challenges, researchers have integrated external information from different sources to enhance the capabilities of large language models, including retrieval-augmented generation methods. In this paper, we present a review to discuss the development trends of retrieval-augmented techniques, including retrieval timing strategies, retrieval paradigms, and utilization of retrieval results. Additionally, we introduce the datasets and evaluation methods currently available for retrieval-augmented tasks, and highlight applications and future potential research directions. We hope that this survey will provide the community with quick access and a comprehensive overview of this research area, to inspire future research efforts.

Keywords: retrieval-augmented

*通讯作者

第二十三届中国计算语言学大会论文集，第151页-第168页，太原，中国，2024年7月25日至28日。
卷2：前沿综述
(c) 2024 中国中文信息学会计算语言学专业委员会

1 引言

大规模预训练语言模型已表现出将现实世界知识编码到参数中的强大潜力，以及解决各种自然语言处理任务的非凡能力 (Brown et al., 2020; Hoffmann et al., 2022; Zeng et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023; Zhao et al., 2023b)。然而，在面对需要大量现实世界知识作为参考的知识密集型任务时 (Petroni et al., 2021)，大模型仍然面临着严峻的挑战。近期研究表明，大语言模型很难学习长尾知识 (Kandpal et al., 2023; Mallen et al., 2023)，并且无法及时更新参数以捕捉不断变化的世界 (De Cao et al., 2021; Kasai et al., 2024)（例如，ChatGPT[0]的参数仅包含2021年9月之前的信息，完全不了解最新的世界知识。）。此外，大模型还会遇到幻觉问题 (Zhang et al., 2023; Rawte et al., 2023; Huang et al., 2023)。为了缓解这些问题，大量研究尝试利用检索外部知识的手段增强语言模型的知识能力 (Mallen et al., 2023; Shi et al., 2023b; Trivedi et al., 2023)，这类方法一般采用现成的检索模型从外部语料库中获取相关文档，以帮助大型语言模型更好地进行内容生成。

检索增强技术能够在推理阶段以非参数化形式利用外部知识，其框架通常由检索器和生成器组成。当前已有工作探索了以端到端方式训练整个检索器-语言模型系统的不同方法：使用检索增强序列对数似然 (Lewis et al., 2020b; Borgeaud et al., 2022)、解码器中融合注意力蒸馏 (Borgeaud et al., 2022; Izacard et al., 2023)或知识图 (Ju et al., 2022)。当越来越多的独特需求出现时，这种微调的成本可能会很高 (Maronikolakis and Schütze, 2021)。更重要的是，许多大型语言模型只能通过黑盒API 访问 (Ouyang et al., 2022; Achiam et al., 2023)。这些API 允许用户提交查询并接收响应，但通常不支持微调。

在本文中，我们以检索策略和内容处理策略为中心，整理现有的检索增强技术。我们先规范化检索增强技术的范式（§2），然后从检索前、检索时和检索后三个阶段总结现有的检索增强方法会面临的三个重要问题：（§3）什么时候需要通过检索来增强大型语言模型？（§4）如何准确高效地检索得到相关信息？（§5）如何利用检索得到的信息和知识优化大模型生成的内容？此外，我们介绍了当前适用于检索增强任务的数据集（§6）和评价方法（§7），并提供了一些对未来研究的前瞻性思考以促进该领域的进一步发展（§8）。

2 检索增强定义

当前文本生成框架遵循以下范式：给定输入上文信息 x，θ 参数化的预训练语言模型通过概率建模 $p_\theta(y|x) = \prod_i p_\theta(y_i|y_{<i}, x)$ 生成相关内容 y。由于大规模预训练语言模型的 θ 参数难以跟随现实的变化实时更新，检索增强技术则尝试在生成范式中添加辅助的检索知识 $p_\theta(y|x, d) = \prod_i p_\theta(y_i|y_{<i}, x, d)$ 以扩展语言模型的能力，其中 d 是从外部知识或语料库 $d_1, \ldots, d_n \in D$ 中获取的辅助信息。检索增强的工作流程可以简单举例阐述：假设用户向预训练大模型 p_θ 询问关于最近实事动态 x 的评论，检索增强系统需要将 x 转写成合适的查询 q 并以此从最新的外部知识库 D 中来检索相关信息 $[d_1, \ldots, d_i]$，进行必要的处理后作为 d 交给语言模型生成对应的回复。这样一种检索生成范式需要面对的问题主要有三点：

1. 检索的时机：面对当前上文信息 x 时，语言模型 p_θ 的能力是否足够生成所需内容，是否需要进行检索补充信息？即需要能够提前估计生成内容 $y \sim p_\theta(y|x)$ 的质量。

2. 检索的策略：如何将当前上文 x 转变为有效的查询 q，并快速准确地从外部知识库 D 中获取相关文档 $[d_1, \ldots, d_i]$？

3. 检索结果的使用：如何将相关文档 $[d_1, \ldots, d_i]$ 转换为适合语言模型 p_θ 理解与使用的辅助信息 d？

3 检索时机规划

检索时机规划是指在检索开始之前，模型发起检索请求是被动还是主动。我们称预先设定好的检索流程为被动规划，基于大模型反馈的检索请求发起为主动检索。由于检索算法和检索数据来源的限制，检索得到的内容并不是百分百相关且准确。Shi et al. (2023a) 和 Wu et

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

[0] https://chat.openai.com

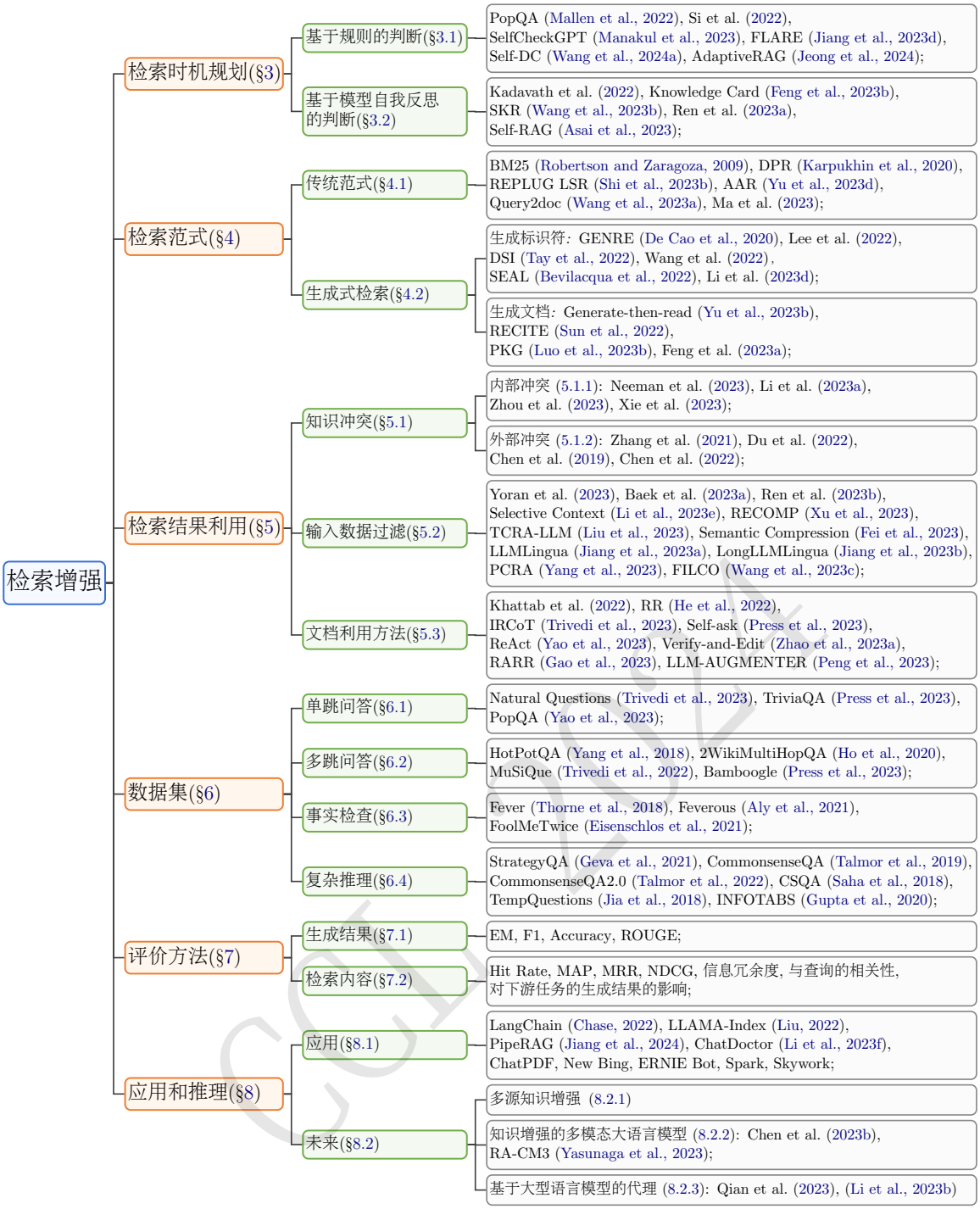

Figure 1: 以检索策略和内容处理策略为中心的检索增强技术分类

al. (2024) 的研究表明：大模型解决问题的能力很容易被输入上下文中不相关但具有误导性的内容干扰。Chen et al. (2023a) 认为，虽然大模型具有一定程度的噪声鲁棒性，但它们在负面信息过滤、信息整合和处理虚假信息方面仍有很大的困难。因此当模型固有的参数化知识足以回答相关问题时，过度检索并不能为最终结果带来增益。由此，本章将深入研究主动检索方法。

对于主动检索来说，一个非常重要的问题是了解大模型的知识边界 (Yin et al., 2023a)并确定何时检索补充知识。根据判断知识边界的方法可将目前的主动检索判断为两类：基于规则的判断和基于模型自我反思的判断。

(c) 2024 中国中文信息学会计算语言学专业委员会

3.1 基于规则的判断

一个简单直观的想法是设置一个指标和阈值。当指标高于或低于阈值时，我们触发检索器来获取相关文档。Kandpal et al. (2022)研究大型语言模型记忆的知识与预训练数据集的信息之间的关系。他们观察到在某些问答数据集上的准确率和相关文档数量之间存在很强的相关性和因果关系，因此得到结论：语言模型回答基于事实的问题的能力与预训练期间看到的与该问题相关的文档数量有关。为了深入分析大模型的参数化知识与数据流行度之间的关系，Mallen et al. (2022)构建一个开放域问答数据集PopQA，其中包含来自维基百科的实体流行度。然后，他们设计了一种自适应检索方法，仅对流行度低于流行度阈值的问题使用检索。除了流行度之外，Jiang et al. (2020)表明大模型往往经过良好校准，低概率或置信度通常表明缺乏相关知识。Si et al. (2022)和 Manakul et al. (2023)利用词语概率来指示其输出的不确定性。遵循这个想法，Jiang et al. (2023d)提出了一种基于置信度的主动检索方法，名为FLARE。如果生成的句子中每个单词的置信度高于阈值，则它们接受该句子而不检索附加信息。否则，它们主动触发检索并利用检索到的相关信息重新生成当前句子。Self-DC (Wang et al., 2024a)则将模型响应的置信度得分划分为三组：未知、不确定和已知。引导大模型按需自适应性地调用不同的方法，被视为未知的查询通过顺序检索增强流水线进行处理，而那些具有不确定性的查询则被分解为子问题以生成答案。AdaptiveRAG (Jeong et al., 2024)则训练一个较小的语言模型作为分类器，判断查询的复杂性，动态决定是否检索，

3.2 基于模型自我反思的判断

基于模型自我反思的判断方法即让大模型根据问题以生成的方式判断是否触发检索。考虑到大语言模型具有非常强大的能力，一些研究人员直接使用大语言模型来确定是否需要检索。Yin et al. (2023b)通过评估大模型识别无法回答或不可知问题的能力来研究他们的自我认知边界。Kadavathet al. (2022)提示大模型预测他们的回答是否可靠的概率。这些不可靠的回答表明大模型需要额外的信息来回答相应的问题。Feng et al. (2023b)询问大模型“您需要更多信息吗？（是或否）”来通过情境学习来确定给定问题是否需要外部知识。SKR (Wang et al., 2023b)构建了一个二元分类数据集，给定问题输出是否可回答。该数据集用于训练小模型或引导大模型情境学习。Ren et al. (2023a)采用先验和后验判断指令来调查大模型是否能够在正常设置和检索设置下感知自己的事实知识边界。先验判断询问大模型是否可以提供问题的答案。事后判断要求大模型评估问题答案的正确性。他们观察到大模型对他们的事实知识边界的认识不准确，并且在正常情况下有过度自信的倾向。Self-RAG (Asai et al., 2023)训练生成器来直接生成检索标记来确定是否执行检索。

4 检索技术

在检索增强系统中，外部知识的质量和语言模型的能力共同决定了生成内容的上限。在当前的语言模型规模和预训练开销下，微调语言模型以追踪最新的知识成本过高难以实现。而利用高质量实时更新的外部知识则是目前最为有效的一种手段。获取得到外部知识的质量主要受到两方面影响，一方面是检索数据源的质量，另一方面是检索方法的性能。传统的检索范式是使用检索器从外部语料库（例如维基百科、知识图谱、网络文本等）获取相关文档。最近有一种观点认为，既然大型语言模型可将现实世界知识编码到其参数中，那么是否可以通过提示模型输出相关的内部知识以获得更准确内容。本章将从传统检索范式和生成式检索两方面介绍现有检索范式。

4.1 传统检索

给定输入上下文 x，检索器旨在从语料库 $D = d_1, \ldots, d_m$ 中检索与 x 相关的一小组文档。检索器有不同类型，包括基于术语的稀疏检索器、基于向量表示的稠密检索器和商业搜索引擎。稀疏检索器通常使用 TF-IDF 或 BM25 实现 (Robertson and Zaragoza, 2009)，它通过倒排索引有效地匹配关键字。然而，术语匹配方法对高度选择性的关键词和短语很敏感。稠密检索器 (Karpukhin et al., 2020)将文本编码到连续的稠密语义空间中，其中由完全不同的词语组成的同义词或复述仍然可以映射到彼此接近的向量。商业搜索引擎，例如谷歌和百度，是能够检索最新世界知识的复杂系统。这三种方法各有不同的优势和应用场景，接下来我们重点关注稠密检索器。

(c) 2024 中国中文信息学会计算语言学专业委员会

给定文本段落的集合，稠密检索器的目标是在低维连续空间中索引所有段落，以便它可以在运行时有效地检索与读者输入问题相关的前 k 个段落。稠密检索器使用稠密编码器 $E(\cdot)$，它将任何文本段落映射到 d 维实值向量。具体来说，编码器通过对 d 中词语的最后一个隐藏表示进行平均池化，将每个文档 $d \in D$ 映射到向量表示 $E(d)$。在查询时，将相同的编码器应用于输入上下文 q 以获得查询向量表示 $E(q)$。查询向量表示和文档向量表示之间的相似度通过它们的余弦相似度计算:

$$s(d, q) = cos(E(d), E(q))$$

通过以上步骤检索获得与输入 q 具有最高相似度分数的前 k 个文档。

由于微调大型语言模型的资源、成本以及黑盒限制，先前使语言模型适应检索器的工作已不适用于当前，最近的工作尝试使检索器适应语言模型。REPLUG LSR (Shi et al., 2023b)利用来自黑盒语言模型的监督信号进一步改进了 REPLUG 中的初始检索模型，即GPT-3 Curie (Brown et al., 2020)。AAR (Yu et al., 2023d)建议利用小规模语言模型为检索器训练提供语言模型偏好的信号。训练后的检索器可以通过插入检索到的文档来直接用于辅助大型目标语言模型。

与之前专注于调整检索器的研究不同，另一个研究方向侧重于弥合输入文本和查询真正需要的知识之间的语义差距。Query2doc (Wang et al., 2023a)通过采用几次提示范例来提示大模型生成伪文档。随后，通过合并伪文档来扩展原始查询。检索器模块使用这个新查询来检索相关文档的列表。Ma et al. (2023)引入了用于检索增强的重写-检索-读取框架，可以进一步调整该框架以适应大模型。他们还在检索器之前添加了查询重写步骤。与 Query2doc 不同的是，它们采用可训练的语言模型来执行重写步骤。重写语言模型通过强化学习进行训练，以 LLM 表现作为奖励。

4.2 生成式检索

生成式检索是一种新的检索范式，主要包括生成文档标识符字符串和生成完整文档两种。

前者使用标识符来减少无用信息量，使模型更容易记忆和学习 (Li et al., 2023d)。De Cao et al. (2020)提出 GENRE，它通过生成实体文本本身来检索实体。GENRE 还可以应用于页面级检索，其中每个文档都包含唯一的标题作为标识符。Lee et al. (2022)将生成检索引入多跳设置，检索的项目是短句。Tay et al. (2022)提出了 DSI 方法，该方法采用数字 ID 作为文档的标识符。Wang et al. (2022)通过生成更多查询作为额外的训练数据来改进 DSI。然而，基于 Id 的数值方法通常在小型数据集上进行评估，部分原因是它们面临大规模扩展问题。Bevilacqua et al. (2022)提出 SEAL，它采用子字符串作为标识符。检索过程是在 FM-Index结构上有效完成的。Li et al. (2023d)提出了多视图标识符，从不同的角度代表一段段落，以增强生成检索并实现最先进的性能。

后者的目标不是生成标识符，而是使用大型语言模型直接生成完整的文档。Generate-then-read (Yu et al., 2023b)表明，生成的上下文文档比检索到的最热门文档更频繁地包含正确答案，并且显著优于直接从大型语言模型生成答案，尽管没有包含任何新的外部信息。RECITE (Sun et al., 2022)采用了类似的方法，它通过首先“背诵”相关信息然后生成输出来处理知识稠密型 NLP 任务。PKG (Luo et al., 2023b)为大模型配备了背景知识生成模块来获取相关知识。

考虑到构建参数化知识模块的知识时效性和微调模型成本，可通过离线高效微调开源小语言模型来存储任何知识。Feng et al. (2023a)建议通过集成专业语言模型，为通用大型语言模型提供模块化和协作来源的知识。专业语言模型是在来自不同来源和领域的语料库上进行训练的。他们还提出了三个级别的知识过滤器，以动态选择和细化生成的文档，并控制主题相关性、文档简洁性和知识真实性。

总结：传统检索范式使用检索器从外部语料库中获取相关文档。然而，检索到的文档可能包含与问题无关的嘈杂信息。另一种选择是使用大型语言模型直接生成相关文档，但无法获得实时信息。所以我们应该根据实际场景做出合适的选择。

5 检索结果利用

在检索得到相关知识和信息之后，如何更有效地利用这些信息，辅助语言模型生成更高质量的内容则至关重要。在语言模型进行生成之前，检索知识的质量需要严格把关，可利用噪声过滤、思维链推理扩展、重新检索等手段进行输入过滤。同时，我们也需要探索如何让语言模

型接受并理解检索知识，并利用该知识进行生成。在语言模型生成之后，检索知识同样可以作为关键信息修正语言模型等生成结果。

5.1 知识冲突

在检索增强的大模型中，有两种知识来源有助于模型推理，但分工不明确且不透明。第一个是通过预训练和微调灌输的隐式参数知识（即它们学习的权重）。第二个是上下文知识，通常来源于检索器的文本段落。知识冲突是指所包含的信息不一致、矛盾。知识冲突有两种类型：内部冲突和外部冲突。内部冲突是指大型语言模型中的知识与检索到的文档中的知识之间的不一致。外部冲突是指检索到的多个文档之间不一致。

5.1.1 内部冲突

随着世界不断发展，记忆的事实可能会变得过时 (Liška et al., 2022; Kasai et al., 2024)。利用包含相关知识的外部背景来增强大模型是一个有前途的方向。然而，此类方法面临的挑战是大模型可能会坚持记住的事实并忽略所提供的上下文 (Longpre et al., 2021)。为了应对这一挑战，最近的著作 (Neeman et al., 2023; Li et al., 2023a)对反事实背景下的大模型进行了微调，其中原始事实被反事实的内容所取代。他们发现这种微调过程可以有效提高大模型对上下文的利用率，而不是仅仅依赖他们的参数知识。Zhou et al. (2023)提出了一种使用提示来提高大模型上下文忠实度的方法，无需额外的微调，这为大模型提供了一种更通用、更具成本效益的方法。他们提出了各种提示策略来提高大模型的忠诚度，包括设计有效的提示和选择适当的上下文演示。Xie et al. (2023)对大模型在遇到反记忆时的行为进行了全面和受控的调查。他们发现，大模型的参数记忆既有支持性的证据，也有矛盾的证据，他们表现出强烈的确认偏差，并倾向于坚持他们的参数记忆。

5.1.2 外部冲突

这种情况在某些段落已用新信息更新而其他段落仍然过时的环境中很常见 (Zhang and Choi, 2021)。当段落被敌对地编辑为包含虚假信息时 (Du et al., 2022)，或者当段落由对答案有不同意见的多人撰写时 (Chen et al., 2019)，也会发生此类冲突。Chen et al. (2022)模拟一种设置，其中证据段落的子集受到干扰，以提出不同的答案，以反映检索返回混合信息包的现实场景。他们发现，当不同的段落提出多个相互冲突的答案时，模型更喜欢与其参数知识相匹配的答案。除了分析简单的内部和外部冲突之外，Xie et al. (2023)还对更复杂的知识冲突场景进行了实验。在提供相关和不相关证据的情况下，大模型可以在一定程度上过滤掉不相关的证据。然而，随着不相关证据数量的增加，这种能力就会减弱。

综上，知识冲突是一个非常重要的问题。然而，目前的研究主要集中在知识冲突的分析上。下一步应该从不同方面解决知识冲突，比如输入数据过滤、改进文档利用方法等。

5.2 输入数据过滤

对于输入过滤，最直觉的方法是首先检索所有问题的相关文档，然后判断相关文档是否可以回答问题。如果相关，大型语言模型会利用检索到的文档来生成答案。如果不相关，大型语言模型会直接生成答案。Yoran et al. (2023)将检索到的文档和问题的相关性判断视为自然语言推理(NLI) 问题 (Dagan et al., 2005; Bowman et al., 2015)，并使用训练有素的 BART-Large NLI (Lewis et al., 2020a)模型来识别不相关的检索文档。检索到的文档作为前提，而问题和生成的答案连接起来作为假设。Baek et al. (2023b)建议使用指令数据对大模型进行微调，以识别输入问题和检索到的文档之间的相关性，并整合各种指令的结果，以进一步提高准确性。Ren et al. (2023b)发现，纳入相关文献后，大模型自我评估的准确性有所提高，动态引入检索文献对大模型来说是有效的。

除了二元判断检索得到的文档是否有帮助外，另一种思考输入过滤方法的角度是对检索得到的内容进行文本摘要，保留其中与目标答案最相关的部分。文本摘要可分为抽取式摘要和生成式摘要。抽取式摘要从检索结果中删除不重要的词语、句子或文档，保留关键句和关键词组成摘要。生成式摘要则根据原文，生成新的词语、短语来组成摘要。Selective Context (Li et al., 2023e)将词语合并为单元，然后基于自信息指标（即负对数似然）应用单元级提示剪枝。RECOMP (Xu et al., 2023)基于表示相似度构造抽取式摘要数据训练 BERT，使用 ChatGPT 生成式摘要数据微调 T5-Large。TCRA-LLM (Liu et al., 2023)则通过去掉句

(c) 2024 中国中文信息学会计算语言学专业委员会

子中语义影响最小的词，基于语义压缩算法构造抽取式数据；利用 ChatGPT 构造不同长度的生成式摘要构造生成式数据。SemanticCompression (Fei et al., 2023)提出了一种语义压缩方法。它首先将文本分解成句子。接下来，它按主题将句子分组，然后总结每组内的句子。LLMLingua (Jiang et al., 2023a)引入了一种从粗到细的方案进行过滤。最初，它执行段落级别删减，然后根据生成模型提供的困惑度计算每个词语的重要性，同时保留重要的词语。为了提高性能，LLMLingua 还提出了一个预算控制器，可以在输入提示的不同部分动态分配修剪预算。LongLLMLingua (Jiang et al., 2023b)基于 LLMLingua 构建，引入了查询感知压缩，并根据计算的重要性得分对检索到的文档进行重新排序。PCRA (Yang et al., 2023)则利用强化学习范式，基于生成结果的反馈约束优化过检索内容过滤器，使其适用于黑盒大模型和不同检索器。FILCO (Wang et al., 2023c)基于后验概率分布预测先验分布，即标注对生成标准答案影响最大的句子作为目的摘要。Zhu et al. (2024)则从信息瓶颈的角度综合后验概率分布和内容冗余程度两方面评估检索上下文质量。

5.3 文档利用方法

有了相关文档后，我们该如何利用它们来提高大语言模型的能力？一种常见的方法是将检索到的文档添加到原始输入中，再将其输入到大型语言模型中以进行最终预测 (Khattab et al., 2022; Yu et al., 2023b; Luo et al., 2023b; Feng et al., 2023b)。大型语言模型能够通过生成逐步的自然语言推理步骤（称为思维链（CoT））来回答复杂的问题 (Wei et al., 2022)。由此引申出将思维链和检索过程相结合来推理复杂问题，通过反复循环检索和生成来逐步细化结果。

He et al. (2022) 提出了一种称为检索重新思考(RR) 的后处理方法，用于在大模型中利用外部知识。他们首先使用思想链（CoT）提示方法来生成一组多样化的推理路径。然后，他们使用这些路径中的每个推理步骤来检索相关的外部知识，这使得RR 能够提供更忠实的解释和更准确的预测。IRCoT (Trivedi et al., 2023)提出了一种交错方法，使用检索来指导思想链推理步骤，并使用 CoT 推理来指导检索。Self-ask (Press et al., 2023) 建立在思想链提示的基础上，但不是输出一个连续的、无界限的思想链。Self-ask 清楚地划分了每个子问题的开始和结束，并使用搜索引擎来回答子问题。Luo et al. (2023a) 构建 CoT 数据集并训练模型在生成阶段首先输出对检索上下文的相关性分析，再回答问题。

ReAct (Yao et al., 2023) 促使大模型以交错的方式生成言语推理轨迹和动作，这使得模型能够执行动态推理来创建、维护和调整高级行动计划，同时还与外部环境将附加信息纳入推理。Verify-and-Edit (Zhao et al., 2023a)旨在通过根据外部知识对推理链进行后期编辑来提高预测的真实性。Self-RAG (Asai et al., 2023)提出“自我反思检索-增强生成”框架。首先训练一个可按需自适应性地生成特殊检索标记的语言模型，获得检索内容后令生成模型反思检索到的段落及其自身的生成，基于最优结果继续生成内容。此外，最近的几项工作 (Shao et al., 2023; Feng et al., 2024; Yu et al., 2023c; Cheng et al., 2024; Wang et al., 2024b)首先生成初始输出，然后基于该输出利用检索模型从大型文档集合中获取相关信息，最后合并将检索到的信息放入输入上下文中以进行输出细化，迭代多次。

RARR (Gao et al., 2023)提出了一种与模型无关的方法来改进任何现有 LM 的归因，而不是限制 LM 生成归因文本。生成文本后，RARR 获取相关证据，然后修改文本以使其与证据一致，同时保留风格或结构等质量，使修改后的文本能够无缝地代替原始文本。RARR 可以被视为增强检索模型，其中检索发生在生成之后而不是生成之前。Peng et al. (2023) 提出 LLM-AUGMENTER，通过外部知识和自动反馈来改进大模型。给定用户查询，LLMAUGMENTER 首先从外部知识中检索证据，并通过将检索到的原始证据与相关上下文联系起来并执行推理以形成证据链来进一步巩固证据。然后，LLM-AUGMENTER 通过检查候选人的回答是否产生幻觉来验证其回答。

综上，可将使用相关文档的时机分为三个不同阶段：使用相关文档作为输入提示的一部分，使用相关文档来确保推理过程的正确性，利用相关文档修改大语言模型的原始答案。三种阶段的方法可综合使用。

6 数据集

因为解决知识密集型任务需要访问大量信息，所以知识密集型任务非常适合评估检索增强型大型语言模型。我们详细研究了常用的数据集，并按任务类型将它们分为以下几类。

6.1 单跳问答

单跳问题的结构相对简单，可以使用段落中包含的信息来回答。常用的数据集有 Natural Questions (NQ)、TriviaQA 和PopQA。Natural Questions (Kwiatkowski et al., 2019a)是由从 Google 搜索引擎聚合的问题组成，答案由专家人工注释。TriviaQA (Joshi et al., 2017)由琐事爱好者撰写的问题组成，证据收集自维基百科和网络。PopQA (Mallen et al., 2022)是一个大规模的以实体为中心的问答数据集，其构造时采样了更多长尾数据，并且明显包含更多的低受欢迎度实体。

6.2 多跳问答

多跳问答是无法通过单一来源或段落直接解决的问题，模型需要执行多个推理步骤才能回答问题。常用的数据集有 HotPotQA、2WikiMultiHopQA、MuSiQue和Bamboogle。HotPotQA (Yang et al., 2018)收集自明确需要对多个支持上下文文档进行推理的问题。2WikiMultihopQA (Ho et al., 2020)也是通过组合构建的，但它们使用一组有限的人工编写的组合规则。MuSiQue (Trivedi et al., 2022)是通过仔细选择和编写单跳问题以自下而上的过程构建的。MuSiQue 具有六种组成结构，比 HotPotQA 和 2WikiMultihopQA 更具挑战性且不易作弊。Bamboogle (Press et al., 2023)是一个由作者编写的包含2 跳问题的小型数据集，其中所有问题都足够困难，以确保常用的互联网搜索引擎无法回答，但所需的两个支持证据都可以在维基百科中找到。

6.3 事实检查

事实验证，也称为事实检查，需要从纯文本中检索相关证据并使用这些证据来验证给定的主张。常用的数据集有 Fever、Feverous 和 FoolMeTwice (FM2)。Fever (Thorne et al., 2018)是一个用于事实验证的大型数据集，需要检索句子级证据来支持某个主张是否得到支持或反驳。除了非结构化文本证据之外，Feverous (Aly et al., 2021)还将维基百科表格视为一种证据形式。Feverous 中的证据检索考虑了维基百科文章的整体，因此证据可以位于文章中除参考部分之外的任何部分。Feverous 的证据分布平衡，要么仅包含文本、表格，要么两者都作为证据，两种情况的实例数量几乎相同。FoolMeTwice (Eisenschlos et al., 2021)是通过一个有趣的多人游戏收集的数据集。该游戏鼓励对抗性示例，其中可以使用捷径解决的实例数量远少于其他数据集。

6.4 复杂推理

复杂推理包括不同类型的推理，如常识推理、表格推理等。常识推理是人类认识的基础，植根于日常生活和社会实践中积累的基础知识和生活经验，概述了世界如何运转的实用知识 (Sap et al., 2020)。常识推理任务评估模型在物理世界中的推理能力。StrategyQA、CommonsenseQA 和 CommonsenseQA2.0 是广泛使用的常识推理数据集。StrategyQA (Geva et al., 2021)是一个专注于开放领域问题的问答基准，其中所需的推理步骤隐含在问题中，应使用策略来推断。CommonsenseQA (Talmor et al., 2019)和CommonsenseQA2.0 (Talmor et al., 2022)的提出是为了探索大型语言模型的常识理解能力，其中包括关于日常常识知识的是/否问题（或论断）。CSQA (Saha et al., 2018)是一个要求长输出的问答数据集，旨在为寻求复杂信息的问题生成全面的答案。TempQuestions (Jia et al., 2018)旨在研究时间推理。该数据集包含1,271 个时间问题，分为四类：显式时间、隐式时间、时间答案和序数约束。INFOTAB (Gupta et al., 2020)由23, 738 个人工编写的文本假设组成，这些假设以从维基百科信息框提取的表格内容为前提。

6.5 其他

除了以上传统任务外，最近也有一些工作针对检索增强任务的特点专门构建数据集。检索增强范式在处理时效性强的任务上具有天然优势，但由于生成模型的训练语料库中可能包含早期数据集的背景知识，致使模型不需要外部知识即可回答问题，为了准确评估检索增强系统的时效性能力，RealTime QA (Kasai et al., 2024)创建了一个动态问答平台，每周定期发布关于最新事件或信息的问题。针对检索上下文中可能包含的噪声、事实冲突，Chen et al. (2024)提出不同大语言模型在检索增强生成中应具有4种基本能力：噪声鲁棒性、负面拒绝、信息集成和反事实鲁棒性，并为此构建一个中英文的新数据集。

(c) 2024 中国中文信息学会计算语言学专业委员会

任务类型	数据集	评价指标
单跳问答	Natural Questions (Kwiatkowski et al., 2019b)	EM / F_1
	TriviaQA (Joshi et al., 2017)	EM / F_1
	PopQA (Mallen et al., 2023)	Accuracy
多跳问答	2WikiMultiHopQA (Ho et al., 2020)	EM / F_1
	HotPotQA (Yang et al., 2018)	EM / F_1
	MuSiQue (Trivedi et al., 2022)	EM / F_1
	Bamboogle (Press et al., 2023)	EM / F_1
事实检查	Fever (Thorne et al., 2018)	Accuracy
	Feverous (Aly et al., 2021)	Accuracy
	FoolMeTwice (Eisenschlos et al., 2021)	Accuracy
复杂推理	StrategyQA (Geva et al., 2021)	Accuracy
	CommonsenseQA (Talmor et al., 2019)	Accuracy
	CommonsenseQA2.0 (Talmor et al., 2022)	Accuracy
	CSQA (Saha et al., 2018)	EM / ROUGE
	TempQuestions (Jia et al., 2018)	EM / F_1
	INFOTABS (Gupta et al., 2020)	EM / F_1

Table 1: 不同任务下数据集及相应评估方法

7 评价方法

检索增强技术有效性的评估主要从两个角度进行。最直观的方式是直接使用具体下游任务的评价方法，如问答任务上的EM、F1，事实检查任务上的Accuracy，生成任务上的ROUGE等。然而仅依靠端到端的评价结果无法全面评估检索增强系统中各组件的性能，我们需要从更细致的角度来衡量检索增强的有效性(Hoshi et al., 2023)。

7.1 生成结果

针对有标准答案的生成结果，表1中展示了第6章中提及的数据集对应的评价方法。而对于没有具体标准答案的生成结果，可针对具体任务采用不同角度的人工或自动评价，如流畅性、连贯性、事实一致性等模型生成能力指标。具体方法与对应任务相关，本文不作深入探讨。

7.2 检索内容

如何评估检索得到内容的质量是当前的难点，关键点在于如何定义检索上下文的有效性。传统搜索引擎的评价方式有命中率（Hit Rate）、平均准确率（Mean Average Precision，MAP）、平均倒数排名（Mean reciprocal rank，MRR）和归一化折损累积增益（Normalized Discounted Cumulative Gain，NDCG）等指标。然而针对检索增强系统，检索上下文的有效性可取决于其信息冗余度、与查询的相关性、对下游任务的生成结果的影响等。综上，我们总结第5.2节中的过滤方法，将过滤检索内容的监督信号总结为以下三种：

- 对于整体语义的贡献：例如，删去不影响句子语义的词语，删去不影响篇章语义的句子。
- 基于向量表示的相似性：例如，保留与查询问题相似度更高的检索信息，对检索信息间相似性高的内容进行酌情删减。
- 对于生成结果的影响程度：例如，保留对生成结果困惑度影响程度高的文章或句子。

8 应用和未来

8.1 应用

事实证明，使用从各种知识存储中检索的相关信息来增强语言模型可以有效提高各种知识密集型任务的性能。在开放域问答和事实验证中，模型可以通过在大型语料库或网络上搜

索相关文档来更准确地回答问题。除了经典的自然语言处理任务之外，随着检索增强大型语言模型的发展，还出现了许多新的应用程序。LangChain (Chase, 2022)是一个功能强大的框架，提供了一组工具、组件和接口来简化创建由大型语言模型和聊天模型支持的应用程序的过程。LangChain 可以轻松管理与大型语言模型的交互、将多个组件链接在一起并集成其他资源。类似的工作有 LLAMA-Index (Liu, 2022) 和 PipeRAG (Jiang et al., 2024)。ChatPDF[1]是一款帮助用户理解 PDF 文档并与之聊天的 AI 工具。它可以识别关键信息、提供简洁的摘要并回答您的问题。ChatDoctor (Li et al., 2023f)是一种专为医疗应用而设计的高级语言模型。患者可以通过聊天界面与 ChatDoctor 模型进行交互，询问有关他们的健康、症状或医疗状况的问题。然后，该模型将分析输入并提供适合患者独特情况的响应。New Bing 通过将 ChatGPT 与微软的搜索引擎相结合来使用检索增强。新的 Bing Chat 根据提示生成搜索查询，检索相关文档，并将它们用作结果的上下文。新必应还提供其生成的句子的信息源链接。综上所述，大语言模型具有更强大的知识理解能力，以及通过检索相关文档进行推理的能力，将会有更多的应用场景。百度、科大讯飞和昆仑万维也提供类似的服务，例如 ERNIE Bot[2]、Spark[3]和 Skywork[4]。

8.2 未来

检索增强的发展仍处于初级阶段，因此还有很大的改进空间。在本节中，我们对未来的研究进行简要概述。

8.2.1 多源知识增强

现有的知识增强方法主要在整合知识的格式和种类方面表现出局限性。当前的检索增强方法主要集中于维基百科或网络的非结构化文本检索 (Shi et al., 2023b; Feng et al., 2023b; Vu et al., 2023)。虽然有诸如 KAPING (Baek et al., 2023a)（知识图谱）、StructGPT(Jiang et al., 2023c)（数据库）等方法探索了基于大模型从结构化文本进行检索以完成增强任务的方法。但在现实场景中，一个复杂的问题可能需要从不同来源收集零散的证据才能得出最终答案。不同来源和不同格式的证据对大型语言模型的影响值得探讨。此外，找到一种合适的方法来整合不同来源的证据也极其重要。

8.2.2 知识增强的多模态大语言模型

多模态学习作为视觉到语言推理的基本技术，由于其巨大的应用潜力而引起了越来越多的研究关注 (Li et al., 2023c; Yu et al., 2023a; Chen et al., 2023b; Yu et al., 2021)。如何赋予大语言模型多模态推理能力正成为研究热点。Cheng et al. (2023b)探索了当前知识编辑方法在细化多模态模型中的应用，并揭示了效果仍在进一步改善。RA-CM3 (Yasunaga et al., 2023)提出了一种检索增强的多模态模型，该模型使基本多模态模型能够引用检索器从外部存储器获取的相关文本和图像。未来的研究可以进一步研究知识和多模态大语言模型的整合，以应对现实世界中的复杂挑战。

8.2.3 基于大型语言模型的代理

自主代理长期以来一直是学术界和工业界的一个突出研究焦点 (Padgham and Winikoff, 2005)。通过获取大量的网络知识，法学硕士在实现人类水平的智能方面表现出了巨大的潜力，并为智能体的进一步发展带来了一线希望 (OpenAI, 2023; Sumers et al., 2023)。这些基于LLM的智能体可以表现出推理和规划能力，并已应用于各种现实场景 (Qian et al., 2023; Li et al., 2023b)。由于现实世界的多样性，基于法学硕士的代理人需要额外的信息来做出决策。探索实际复杂场景中知识与大语言模型方法的融合对于基于大模型的代理的开发非常重要。

9 结论

在本文中，我们对检索增强技术进行了调查，并对其主要方向提供了具体介绍。此外，我们总结了常用的数据集和前沿应用，并指出了一些有前景的研究方向。我们希望这项调查能让读者清楚地了解当前的进展并激发更多的工作。

[1] https://www.chatpdf.com/
[2] https://yiyan.baidu.com/
[3] https://xinghuo.xfyun.cn/
[4] https://search.tiangong.cn/

参考文献

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023a. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 78–106.

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C. Park, and Sung Ju Hwang. 2023b. Knowledge-augmented language model verification.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems*, 35:31668–31683.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Harrison Chase. 2022. Langchain. https://github.com/langchain-ai/langchain.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hung-Ting Chen, Michael JQ Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *arXiv preprint arXiv:2210.13701*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023a. Benchmarking large language models in retrieval-augmented generation. In *AAAI Conference on Artificial Intelligence*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023b. Measuring and improving chain-of-thought reasoning in vision-language models. *arXiv preprint arXiv:2309.04461*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2024. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

(c) 2024 中国中文信息学会计算语言学专业委员会

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Yibing Du, Antoine Bosselut, and Christopher D Manning. 2022. Synthetic disinformation attacks on automated fact verification systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10581–10589.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online, June. Association for Computational Linguistics.

Weizhi Fei, Xueyan Niu, Pingyi Zhou, Lu Hou, Bo Bai, Lei Deng, and Wei Han. 2023. Extending context window of large language models via semantic compression. *arXiv preprint arXiv:2312.09571*.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023a. Cook: Empowering general-purpose language models with modular and collaborative knowledge. *arXiv preprint arXiv:2305.09955*.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023b. Knowledge card: Filling llms' knowledge gaps with plug-in specialized language models. In *The Twelfth International Conference on Learning Representations*.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2024. Retrieval-generation synergy augmented large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11661–11665. IEEE.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online, July. Association for Computational Linguistics.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Yasuto Hoshi, Daisuke Miyashita, Youyang Ng, Kento Tatsuno, Yasuhiro Morioka, Osamu Torii, and Jun Deguchi. 2023. Ralle: A framework for developing and evaluating retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 52–69.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

CCL 2024

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1057–1062.

Zhengbao Jiang, J. Araki, Haibo Ding, and Graham Neubig. 2020. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023a. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023b. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Structgpt: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023d. Active retrieval augmented generation. *ArXiv*, abs/2305.06983.

Wenqi Jiang, Shuai Zhang, Boran Han, et al. 2024. Piperag: Fast retrieval-augmented generation via algorithm-system co-design. *arXiv:2403.05676*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July. Association for Computational Linguistics.

Mingxuan Ju, Wenhao Yu, Tong Zhao, Chuxu Zhang, and Yanfang Ye. 2022. Grape: Knowledge graph enhanced passage reader for open-domain question answering. *arXiv preprint arXiv:2210.02933*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221.

Nikhil Kandpal, H. Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2024. Realtime qa: What's the answer right now? *Advances in Neural Information Processing Systems*, 36.

(c) 2024 中国中文信息学会计算语言学专业委员会

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019a. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019b. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Hyunji Lee, Sohee Yang, Hanseok Oh, and Minjoon Seo. 2022. Generative multi-hop retrieval. In *Conference on Empirical Methods in Natural Language Processing*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. Camel: Communicative agents for "mind" exploration of large language model society.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023d. Multiview identifiers enhanced generative retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6636–6648, Toronto, Canada, July. Association for Computational Linguistics.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023e. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023f. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Junyi Liu, Liangzhi Li, Tong Xiang, Bowen Wang, and Yiming Qian. 2023. Tcra-llm: Token compression retrieval augmented large language model for inference cost reduction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9796–9810.

Jerry Liu. 2022. LlamaIndex, 11.

Adam Liška, Tomáš Kočiský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

(c) 2024 中国中文信息学会计算语言学专业委员会

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023a. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*.

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023b. Augmented large language models with parametric knowledge guiding. *arXiv preprint arXiv:2305.04757*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.

Alex Troy Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Annual Meeting of the Association for Computational Linguistics*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada, July. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark John Francis Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv*, abs/2303.08896.

Antonis Maronikolakis and Hinrich Schütze. 2021. Multidomain pretrained language models for green nlp. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 1–8.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Lin Padgham and Michael Winikoff. 2005. *Developing intelligent agent systems: A practical guide*. John Wiley & Sons.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online, June. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models.

Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, J. Liu, Hao Tian, Huaqin Wu, Ji rong Wen, and Haifeng Wang. 2023a. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *ArXiv*, abs/2307.11019.

(c) 2024 中国中文信息学会计算语言学专业委员会

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023b. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Amrita Saha, Vardaan Pahuja, Mitesh Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online, July. Association for Computational Linguistics.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan L. Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *ArXiv*, abs/2210.09150.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2023. Cognitive architectures for language agents.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *arXiv preprint arXiv:2210.01296*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. Commonsenseqa 2.0: Exposing the limits of ai through gamification. *arXiv preprint arXiv:2201.05320*.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition.

(c) 2024 中国中文信息学会计算语言学专业委员会

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, et al. 2022. A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems*, 35:25600–25614.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023b. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023c. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

Hongru Wang, Boyang Xue, Baohang Zhou, Tianhua Zhang, Cunxiang Wang, Guanhua Chen, Huimin Wang, and Kam-fai Wong. 2024a. Self-dc: When to retrieve and when to generate? self divide-and-conquer for compositional unknown questions. *arXiv preprint arXiv:2402.13514*.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024b. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *ArXiv*, abs/2404.03302.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge clashes.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November. Association for Computational Linguistics.

Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. 2023. Prca: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5364–5375.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Retrieval-augmented multimodal language modeling.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023a. Do Large Language Models Know What They Don't Know? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada, July. Association for Computational Linguistics.

(c) 2024 中国中文信息学会计算语言学专业委员会

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023b. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada, July. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. 2021. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. *Advances in Neural Information Processing Systems*, 34:26462–26474.

Weijiang Yu, Haofan Wang, Guohao Li, Nong Xiao, and Bernard Ghanem. 2023a. Knowledge-aware global reasoning for situation recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, S Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023b. Generate rather than retrieve: Large language models are strong context generators. In *International Conference on Learning Representations*.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023c. Improving language models via plug-and-play retrieval feedback.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023d. Augmentation-adapted retriever improves generalization of language models as generic plug-in. *arXiv preprint arXiv:2305.17331*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Michael Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023a. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada, July. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023b. A survey of large language models.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556.

Kun Zhu, Xiaocheng Feng, Xiyuan Du, Yuxuan Gu, Weijiang Yu, Haotian Wang, Qianglong Chen, Zheng Chu, Jingchang Chen, and Bing Qin. 2024. An information bottleneck perspective for effective noise filtering on retrieval-augmented generation. In *Proceedings of the 62th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

(c) 2024 中国中文信息学会计算语言学专业委员会