利用 LLM 辅助查询理解 进行实时检索增强生成

Guanting Dong
Renmin University of China
Beijing, China
dongguanting@ruc.edu.cn

Xiaoxi Li
Renmin University of China
Beijing, China
xiaoxi_li@ruc.edu.cn

Yuyao Zhang
Renmin University of China
Beijing, China
2020201710@ruc.edu.cn

Mengjie Deng
Renmin University of China
Beijing, China
dengmengjie_777@163.com

ABSTRACT

现实世界中的实时检索增强生成（RAG）系统在处理用户查询时面临重大挑战，这些查询通常噪声大、模糊且包含多种意图。虽然 RAG 增强了大型语言模型（LLMs）的外部知识，但当前系统往往难以应对这种复杂输入，因为它们通常是基于更干净的数据进行训练或评估的。本文介绍了 Omni-RAG，这是一个旨在提高 RAG 系统在实时开放领域设置中的鲁棒性和有效性的新框架。Omni-RAG 采用 LLM 辅助查询理解通过三个关键模块预处理用户输入：（1）深度查询理解和分解，利用带有定制提示词的 LLMs 对查询进行去噪（例如，纠正拼写错误）并将多意图查询分解为结构化的子查询；（2）意图感知知识检索，从语料库中（即。，FineWeb 使用 OpenSearch）针对每个子查询执行检索并聚合结果；以及（3）重排序和生成，在此步骤中一个重排序器（即。，BGE）在最终响应由 LLM（即。，Falcon-10B）根据思维链提示生成之前细化文档选择。Omni-RAG 旨在弥合当前 RAG 能力与现实世界应用需求之间的差距，例如 SIGIR 2025 LiveRAG 挑战所强调的那样，通过稳健地处理复杂和噪声查询。

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS CONCEPTS

• Information systems → Retrieval models and ranking.

KEYWORDS

检索增强生成，查询理解，去噪，文档排名

ACM Reference Format:
Guanting Dong, Xiaoxi Li, Yuyao Zhang, and Mengjie Deng. 2025. 利用 LLM 辅助查询理解 进行实时检索增强生成. In Proceedings of ACM Conference (Conference'17). ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 介绍

大型语言模型（LLMs）[10, 38, 50] 的快速发展已在其广泛应用于自然语言处理任务方面带来了变革性的进展 [25, 52] 。然而，在处理知识密集型任务时，LLMs 仍然仅依赖于其内部知识，这常常在事实一致性及幻觉问题上表现不足 [15] 。为了解决这些问题，研究人员提出了检索增强生成（RAG）[13] ，该方法结合了外部知识来源以辅助 LLMs 的内容生成，显著提高了输出的准确性和可靠性。

然而，在实际的实时 RAG 应用程序中，用户查询很少是原子化的或单一意图的。相反，它们通常涉及多个意图、复杂的结构和各种类型的噪声 [6, 21]。虽然现有的 RAG 方法在标准基准上表现良好，但它们通常选择简单且无噪声的数据集进行微调或对齐。因此，当面对开放领域设置中嘈杂的、模糊的或多意图查询时，这些系统难以准确解释意图并生成可靠的响应。

为了推进这一方向的研究，SIGIR 2025 LiveRAG 挑战赛引入了首个专门设计用于评估在线 RAG 系统实时问题解决能力的比赛。该挑战为所有团队提供了一个固定的知识语料库（FineWeb）[31] 和一个预训练的语言模型（Falcon3-10B-Instruct）[32]，同时使用可配置的合成 DataMorgana [11] 模拟器动态生成多样化的用户查询，以模拟实时的人类查询交互。参与的 RAG 系统必须在两小时内完成任务，需要高效处理复杂、嘈杂和多意图查询。因此，核心挑战在于稳健且高效地理解用户查询中的潜在意图和语义噪声，这对构建实用的实时 RAG 系统构成了重大障碍。

为了弥合这一差距，我们设计了 Omni-RAG 框架，该框架利用大语言模型的理解能力来预处理用户查询——通过去噪和意图分解——增强了 RAG 系统在实际在线环境中的鲁棒性。该框架包括三个关键模块：

- **深度查询理解和分解：** 基于大语言模型在语言理解方面的优势，我们应用定制化的提示来引导模型对用户查询进行预处理。这包括重写噪声输入并将具有多个意图的复杂查询分解为更清晰、结构化的子查询。
- **意图感知的知识检索：** 为了检索全面且相关的支持信息，我们使用一个 OpenSearch 系统在 FineWeb 语料库上对每个重写和分解后的子查询执行检索。检索到的文档随后被汇总成一个统一的语料库，以捕捉生成所需的更广泛的语义背景。
- **重排序和生成：** 在生成之前，我们应用一个 BGE 重新排序器来对所有子查询的候选文档进行重新排序，选择最相关的前 10 个。然后将这些与重写的主要查询集成到一个思维链提示中，该提示被输入到 Falcon-10B 模型以生成最终响应。

实验结果表明，Omni-RAG 框架实现了**第 2 名在** SIGIR LiveRAG **挑战赛第 1 阶段的整体表现**。此外，我们按照官方评估指标在 dry-test 集上进行了伪标记实验，进一步证明了该框架在生成效率和事实一致性方面的有效性。

2 相关工作

检索增强生成。检索增强生成（RAG）[13] 已经成为一种强大的范式，它将外部信息或知识融入以提高生成文本的质量、事实性和相关性。最近的努力 [5, 7, 23, 24] 利用了 RAG 来应对幻觉挑战，并在一系列任务中提高了 LLM 的性能。为了进一步提升检索质量，已经引入了几种后检索策略 [18, 36, 46] 以填补检索器和生成器之间的差距，包括重新排序、改进和压缩。重新排序器 [17, 19, 44] 重新排列了从检索器获取的结果，使它们更好地与 LLM 的信息需求保持一致。此外，一些研究 [2, 43] 引入了技术来减轻检索到的知识文档中的噪声问题。为了解决长上下文限制的问题，各种方法 [14, 45] 集中于压缩检索到的引用以符合输入长度限制，并移除不相关的内容以增强鲁棒性。

查询理解。查询理解 [3, 37] 包括一系列旨在提高检索增强生成系统在预检索阶段效率和准确性的技术，包括查询改写、消歧、分解和扩展。最近的进展 [1, 6] 强调了 LLMs 在查询理解中的关键作用，以提升检索

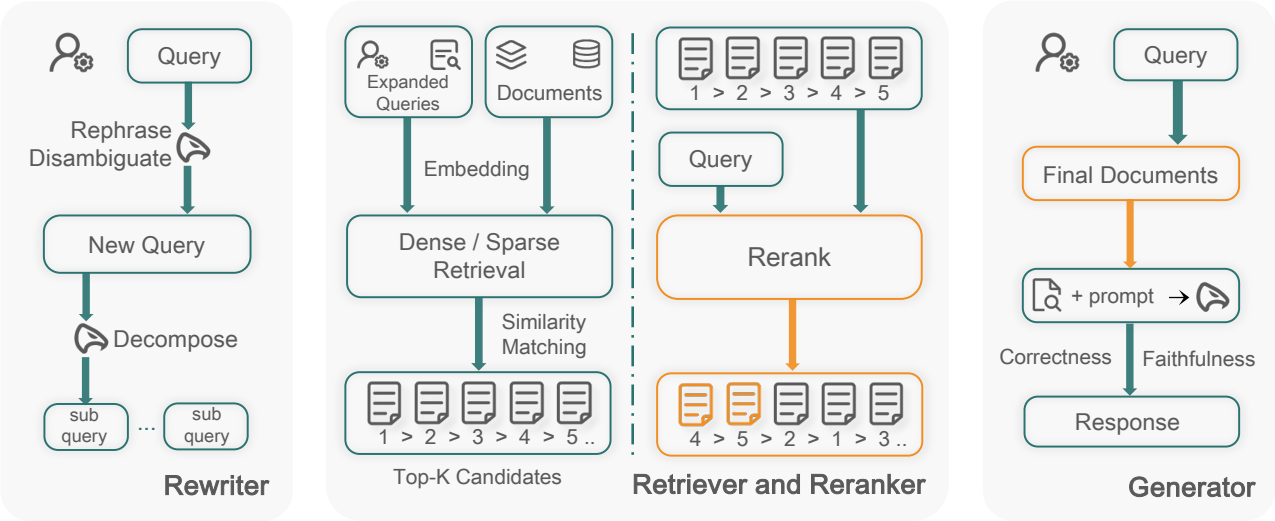

Figure 1: **我们的** Omni-RAG **的整体管道**。

质量。查询改写 [26, 29] 涉及将原始查询重新表述为更符合有效检索所需信息的版本，从而解决人类意图与模型解释之间的常见不匹配问题。查询消歧 [20, 27, 28, 33] 侧重于通过将模棱两可或多轮次的查询转换成更具具体性和语境感知性的搜索输入来澄清用户意图。查询分解 [34, 49, 51] 将复杂查询分解为更简单的子查询，以提高检索效果并实现全面的答案生成。基于推理的查询分解方法 [8, 39, 40, 47] 侧重于生成解决复杂任务的推理轨迹或计划。查询扩展通过丰富原始查询来提升检索性能，增加的信息来源于内部或外部知识源。内部扩展方法 [12, 16, 41, 48] 侧重于利用大语言模型中的参数化知识来增强原始查询，而外部扩展方法 [30, 35] 则结合来自知识库等外部来源的补充信息。

3 方法

概述。为了确保生成的 RAG 回应稳健且高质量，我们引入了由大语言模型驱动的 Omini-RAG 流水线，如图 1 所示。给定一个查询，大语言模型首先进行深度理解，包括重写和分解。然后应用检索和重新排序以获取每个子意图的候选文档，接着通过 Falcon-10B 生成回应。接下来的小节详细介绍了流水线的每一阶段。

3.1 问题定义

与标准文本生成相比，RAG 经常采用一种检索然后阅读范式 [9, 22]，在这种范式中引入了一个额外的检索器来收集外部知识并增强生成过程。给定输入用户查询用 q 表示。RAG 系统的主要目标是生成一个全面且相关的响应 R。形式上，该系统旨在找到一个最优响应 R^*，使得：

$$R^* = \arg\max_R P(R \mid q, \mathcal{K}) \tag{1}$$

其中 $\mathcal{K}$ 代表可以从其检索信息的知识库或语料库。下面描述的流程概述了逼近此最优响应的步骤。

3.2 查询理解与分解

现实世界中的用户查询经常包含噪声，比如拼写错误或模糊的措辞。直接使用这样的查询进行搜索可能会导致不准确或无关的检索结果。因此，我们首先采用一个 LLM 来进行查询理解和重写。设原始查询为 q。重写过程可以表示为：

$$q' = \text{Rewrite}(q, \theta_{\text{rewrite}}) \tag{2}$$

其中 q' 是重写的查询，而 θ_{rewrite} 代表了用于重写任务的 LLM 的微调参数。

重写后，虽然查询 q' 的语义变得正确且直观，但其意图可能仍然复杂或多方面。为了进一步实现更精确的检索目标并提高相关文档的召回率，我们执行查询分解。具体来说，重写的查询 q' 被输入到一个大语言模型中，该模型输出一组 M 子查询 $\{q'_1, q'_2, \ldots, q'_M\}$。上述过程可以表示为：

$$\{q'_s\}_{s=1}^{M} = \text{Decompose}(q', \theta_{\text{decompose}}) \tag{3}$$

其中每个子查询 q'_s 都被设计用来针对原始查询的一个特定方面。这些子查询通常以结构化格式生成，例如 JSON，以便于提取和处理。这为搜索过程中检索更广泛且准确的知识奠定了坚实的基础。

3.3 意图感知知识检索

为了获得更加全面和广泛的信息，一个直观的方法是从复杂查询的分解中为每个子意图检索信息。对于每个分解出的子查询 q'_s（其中 $s \in \{1, \ldots, M\}$），我们利用搜索函数从知识库 $\mathcal{K}$ 中检索出前 K 个相关文档。令 D_s 为子查询 q'_s 检索到的文档集：

$$D_s = \text{Search}(q'_s, \mathcal{K}, K) = \{d_{s,1}, d_{s,2}, \ldots, d_{s,K}\} \tag{4}$$

其中 $d_{s,k}$ 是子查询 q'_s 检索到的第 k 个文档。最初检索到的文档集合，$D_{\text{retrieved}}$，是通过对所有子查询检索到的文档进行并集运算形成的，$D_{\text{retrieved}} = \bigcup_{s=1}^{M} D_s$。这确保了对原始查询不同方面的广泛信息覆盖。

3.4 重新排序和生成：

重新排序： 为了平衡子查询引入的冗余检索结果，重新排序和过滤信息是必不可少的。在获得初始检索文档集 $D_{\text{retrieved}}$ 后，我们使用一个重新排序模型来细化这些文档的选择和顺序。为此，我们使用了一个复杂的重新排序模型，例如 BGE-reranker-large。该重新排序器计算原始查询 q（或重写后的查询 q'）与每个文档 $d \in D_{\text{retrieved}}$ 的相关性分数。令 $\text{score}(q, d)$ 为重新排序器分配的相关性分数。然后根据这些分数按降序对 $D_{\text{retrieved}}$ 中的文档进行排序。我们从这个排序列表中选择前 N 个文档来形成最终的上下文文档集，D_{reranked}：

$$D_{\text{reranked}} = \{d_1^*, d_2^*, \ldots, d_N^*\} \subseteq D_{\text{retrieved}} \tag{5}$$

使得 $\text{score}(q, d_i^*) \geq \text{score}(q, d_{i+1}^*)$ 对于所有的 $i \in \{1, \ldots, N-1\}$ 成立，并且 N 是用于生成的一个预定义文档数量。在重新排序后，我们获得了与查询相关的高质量文档，为准确生成提供了重要支持。

生成： 最后，使用原始查询 q 和重新排序后的前 N 个文档 D_{reranked}，我们利用一个 LLM 生成最终的响应 R。该 LLM 同时依赖于查询和所选文档提供的上下文信息：

$$R = \text{Generate}(q, D_{\text{reranked}}, \theta_{\text{generate}}) \tag{6}$$

其中 θ_{generate} 表示用于生成的 LLM 的参数。此步骤旨在将检索到的文档中的信息综合成一个连贯、准确且符合上下文的回答来回应用户的查询。

3.5 伪标签与评估

由于参考数据不包含真实答案，我们提出了一种基于大语言模型的伪标签生成和一致性评估策略，以支持 RAG 系统在开发过程中的性能评估和迭代优化。需要注意的是，伪标签的使用严格遵循竞赛指南：它们仅用于系统的评估以及干燥测试中无答案样本的分析，绝不会在实际的答案生成过程中使用。

具体来说，我们首先采用 Qwen2.5-7B-Instruct（少于 10B 参数）[4] 作为参考模型，通过向其提供输入查询及其检索到的文档来生成伪答案。为了实现多维度评估，我们根据官方评估标准定义了两个核心指标——“相关性” 和 “忠实度”，并为每个指标设计了专门的提示（如下）。

以相关性评估提示为例，它首先定义了模型的角色，然后是关键评估点和评分指南。此外，还包括四个手工制作的示例来表示不同的评级水平，这增强了模型的情境理解并提高了评分一致性。忠实度评估的提示结构与相关性评估相同。最后，我们使用 Falcon-10B 独立执行相关性和忠实度评估，为每个候选答案生成伪分数。

Table 1: **第 1 轮实时 RAG 挑战的团队排名**

排名	团队名称	正确性	忠实的
1	RMIT-ADMS	1.1993	0.4774
2	RUC_DeepSearch（**我们提出的**）	0.9693	0.3878
3	Ped100X	0.9289	0.0434
4	PRMAS-DRCA	0.9228	0.4106
5	Hybrid Search w. Graph	0.8751	0.3158
6	BagBag	0.6941	-0.9114
7	UniClustRAG	0.6851	0.4601
8	METURAG	0.6735	0.3253
9	DeepRAG	0.5661	0.0978
10	UiS-IAI	0.5523	0.4337
11	SNU-LDILab	0.5174	0.1030
12	Gravitational Lens	0.3766	-0.9881

Relevance Evaluation Prompt Template

你是一名专家评估员，负责根据提供的参考答案（黄金答案）来评估一个预测答案的质量。你的任务是基于以下评分标准，根据预测的语义等价性和相关性分配分数：

你的评估考虑了：
- 等价性：预测是否传达了与标准答案相同的含义？
- 相关性：预测是否直接回答了问题而没有添加无关信息？

评分标准：
- 2：正确且相关（无无关信息）。
- 1：正确但包含无关信息。
- 0：未提供答案（弃权）。
- -1：答案错误。

指令：
根据其与黄金答案的对齐程度以及直接回答问题的程度评估预测。仅返回数字分数（2、1、0、-1）。

上下文示例：{examples}
问题：{question}
黄金答案：{answer}
预测：{prediction}

Table 2: **不同 top-k 设置下的性能对比使用** OpenSearch。5 (sc4) **表示使用前** 5 **个文档和** 4 **条采样推理路径生成以保持自洽性**。

方法	顶部-k	相关性					忠实性			
		Avg	-1	0	1	2	Avg	-1	0	1
全 RAG	1	140	4	2	44	50	62	4	30	66
	2	136	6	2	42	50	66	2	30	68
	3	146	6	0	36	58	70	0	30	70
	4	154	2	0	40	58	76	2	20	78
	5	156	4	0	32	64	80	0	20	80
全 RAG	5 (sc4)	170	2	0	24	74	72	0	28	72
	5 (sc8)	148	4	0	40	56	80	0	20	80

3.6 实验

实验设置。. 我们严格遵循 LiveRAG 竞赛的要求，使用 OpenSearch 从 Falcon 语料库中检索，BGE 作为重排序器，Falcon-10B 作为生成器。

值得注意的是，在表 2 中，我们实验了自洽（sc）策略 [42]，并且这些指标基于 Qwen2.5-72B-Instruct 生成的内部伪相关性和忠实度分数。

主要结果。. 主要结果呈现在主表中，其中我们的团队（*RUC_DeepSearch*）在第一环节的 12 支参赛队伍中总排名第二。

值得注意的是，与排名第三的团队 Team Ped100X 相比，我们的系统在正确性上提高了超过 4%，在忠实度上大约提高了 34%。同样地，与总体排名第 5 的 Team BagBag 相比，我们的系统在正确性上领先约 9%，在忠实度上高出约 7%。这些结果清楚地证明了我们 RobustRAG 框架的可靠性和有效性。

干试分析。. 干测试是我们研究中的一个代表性评估设置。为了进一步分析我们的 RAG 性能，我们从干测试集中选取了 50 个样本，并使用基于前五篇文档的 Qwen2.5-7B-instruct 的回答作为参考来评估我们的模型。关键发现如下。

性能随文档数量变化： 随着检索到的文档数量增加，Omni-RAG 模型在生成质量上表现出强大的可扩

展性。观察到保真度和相关性都有所提高，表明性能得益于更丰富的文档上下文。

自洽路径设置中的权衡：为了提高推理稳定性，我们引入了自洽机制。然而，5（sc8）配置并没有达到预期的性能提升，这表明更多的推理路径并不一定会带来更好的结果。有趣的是，5（sc4）设置提高了相关性，但导致忠实度适度下降，强调了需要平衡路径数量与生成质量的关系。

4 结论

本文介绍了 Omni-RAG，一个通过 LLM 辅助查询理解来增强 RAG 系统的强大且可扩展的框架。通过集成深度查询分解、意图感知检索和重排序引导生成，Omni-RAG 有效解决了实时开放领域环境中复杂和噪声查询带来的挑战。我们的方法展示了在实际应用中的强大力量，在 SIGIR LiveRAG Challenge 的第一轮中取得了第二名的整体表现，并为更可靠和智能的检索增强系统提供了一个实用步骤。

REFERENCES

[1] Abhijit Anand, Venktesh V, Vinay Setty, and Avishek Anand. 2023. Context Aware Query Rewriting for Text Rankers using LLM. CoRR abs/2308.16753 (2023). https://doi.org/10.48550/ARXIV.2308.16753 arXiv:2308.16753

[2] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. CoRR abs/2310.11511 (2023). https://doi.org/10.48550/ARXIV.2310.11511 arXiv:2310.11511

[3] Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: A survey. Inf. Process. Manag. 56, 5 (2019), 1698–1735. https://doi.org/10.1016/J.IPM.2019.05.009

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. CoRR abs/2309.16609 (2023). https://doi.org/10.48550/ARXIV.2309.16609 arXiv:2309.16609

[5] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162), Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 2206–2240. https://proceedings.mlr.press/v162/borgeaud22a.html

[6] Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation. CoRR abs/2404.00610 (2024). https://doi.org/10.48550/ARXIV.2404.00610 arXiv:2404.00610

[7] Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. 2025. Tool-Star: Empowering LLM-Brained Multi-Tool Reasoner via Reinforcement Learning. arXiv:2505.16410 [cs.CL] https://arxiv.org/abs/2505.16410

[8] Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. 2025. Self-play with Execution Feedback: Improving Instruction-following Capabilities of Large Language Models. In The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net. https://openreview.net/forum?id=cRR0oDFEBC

[9] Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Zhicheng Dou, and Ji-Rong Wen. 2024. Understand What LLM Needs: Dual Preference Alignment for Retrieval-Augmented Generation. CoRR abs/2406.18676 (2024). https://doi.org/10.48550/ARXIV.2406.18676 arXiv:2406.18676

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert,

Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 Herd of Models. CoRR abs/2407.21783 (2024). https://doi.org/10.48550/ARXIV.2407.21783 arXiv:2407.21783

[11] Simone Filice, Guy Horowitz, David Carmel, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2025. Generating Diverse Q&A Benchmarks for RAG Evaluation with DataMorgana. arXiv preprint arXiv:2501.12789 (2025).

[12] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 1762–1777. https://doi.org/10.18653/V1/2023.ACL-LONG.99

[13] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. CoRR abs/2312.10997 (2023). https://doi.org/10.48550/ARXIV.2312.10997 arXiv:2312.10997

[14] Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2023. Fid-light: Efficient and effective retrieval-augmented text generation. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1437–1447.

[15] 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. CoRR abs/2311.05232 (2023). https://doi.org/10.48550/ARXIV.2311.05232 arXiv:2311.05232

[16] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023. Association for Computational Linguistics, 7969–7992. https://aclanthology.org/2023.emnlp-main.495

[17] Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan Ö. Arik. 2025. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. In The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net. https://openreview.net/forum?id=oU3tpaR8fm

[18] Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Qi Ye, and Zhicheng Dou. 2025. Hierarchical Document Refinement for Long-context Retrieval-augmented Generation. arXiv:2505.10413 [cs.CL] https://arxiv.org/abs/2505.10413

[19] Zixuan Ke, Weize Kong, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Bridging the Preference Gap between Retrievers and LLMs. arXiv:2401.06954 [cs.CL]

[20] Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of Clarifications: Answering Ambiguous Questions with Retrieval-Augmented Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 996–1009. https://doi.org/10.18653/V1/2023.EMNLP-MAIN.63

[21] Myeonghwa Lee, Seonho An, and Min-Soo Kim. 2024. Plan-RAG: A Plan-then-Retrieval Augmented Generation for Generative Large Language Models as Decision Makers. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024, Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (Eds.). Association for Computational Linguistics, 6537–6555. https://doi.org/10.18653/V1/2024.NAACL-LONG.364

[22] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[23] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic Search-Enhanced Large Reasoning Models. CoRR abs/2501.05366 (2025). https://doi.org/10.48550/ARXIV.2501.05366 arXiv:2501.05366

[24] Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capability. CoRR abs/2504.21776 (2025). https://doi.org/10.48550/ARXIV.2504.21776 arXiv:2504.21776

[25] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From Matching to Generation: A Survey on Generative Information Retrieval. ACM Trans. Inf. Syst. 43, 3, Article 83 (May 2025), 62 pages. https://doi.org/10.1145/3722552

[26] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query Rewriting for Retrieval-Augmented Large Language Models. CoRR abs/2305.14283 (2023). https://doi.org/10.48550/ARXIV.2305.14283 arXiv:2305.14283

[27] Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang.

2024. RaFe: Ranking Feedback Improves Query Rewriting for RAG. CoRR abs/2405.14431 (2024). https://doi.org/10.48550/ARXIV.2405.14431 arXiv:2405.14431

[28] Raja Sekhar Reddy Mekala, Yasaman Razeghi, and Sameer Singh. 2024. EchoPrompt: Instructing the Model to Rephrase Queries for Improved In-context Learning. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024, Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (Eds.). Association for Computational Linguistics, 399–432. https://doi.org/10.18653/V1/2024.NAACL-SHORT.35

[29] Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop Reading Comprehension through Question Decomposition and Rescoring. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 6097–6109. https://doi.org/10.18653/V1/P19-1613

[30] Jeonghyun Park and Hwanhee Lee. 2024. Conversational Query Reformulation with the Guidance of Retrieved Documents. CoRR abs/2407.12363 (2024). https://doi.org/10.48550/ARXIV.2407.12363 arXiv:2407.12363

[31] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. Advances in Neural Information Processing Systems 37 (2024), 30811–30849.

[32] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023. http://papers.nips.cc/paper_files/paper/2023/hash/fa3ed726cc5073b9c31e3e49a807789c-Abstract-Datasets_and_Benchmarks.html

[33] Wenjun Peng, Guiyang Li, Yue Jiang, Zilong Wang, Dan Ou, Xiaoyi Zeng, Derong Xu, Tong Xu, and Enhong Chen. 2024. Large Language Model based Long-tail Query Rewriting in Taobao Search. In Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024, Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw (Eds.). ACM, 20–28. https://doi.org/10.1145/3589335.3648298

[34] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 5687–5711. https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.378

[35] Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Yibin Lei, Tianyi Zhou, Michael Blumenstein, and Daxin Jiang. 2024. Retrieval-Augmented Retrieval: Large Language Models are Strong Zero-Shot Retriever. In Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 15933–15946. https://doi.org/10.18653/V1/2024.FINDINGS-ACL.943

[36] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models. CoRR abs/2301.12652 (2023). https://doi.org/10.48550/ARXIV.2301.12652 arXiv:2301.12652

[37] Mingyang Song and Mao Zheng. 2024. A Survey of Query Optimization in Large Language Models. CoRR abs/2412.17558 (2024). https://doi.org/10.48550/ARXIV.2412.17558 arXiv:2412.17558

[38] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. CoRR abs/2307.09288 (2023). https://doi.org/10.48550/ARXIV.2307.09288 arXiv:2307.09288

[39] Venktesh V, Sourangshu Bhattacharya, and Avishek Anand. 2023. In-Context Ability Transfer for Question Decomposition in Complex QA. CoRR abs/2310.18371 (2023). https://doi.org/10.48550/ARXIV.2310.18371 arXiv:2310.18371

[40] Prakhar Verma, Sukruta Prakash Midigeshi, Gaurav Sinha, Arno Solin, Nagarajan Natarajan, and Amit Sharma. 2024. Plan×RAG: Planning-guided Retrieval Augmented Generation. arXiv:2410.20753 [cs.CL] https://arxiv.org/abs/2410.20753

[41] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 9414–9423. https://doi.org/10.18653/V1/2023.EMNLP-MAIN.585

[42] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 13484–13508. https://doi.org/10.18653/V1/2023.ACL-LONG.754

[43] Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-Knowledge Guided Retrieval Augmentation for Large Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2023, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10303–10315. https://doi.org/10.18653/v1/2023.findings-emnlp.691

[44] Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to Filter Context for Retrieval-Augmented Generation. arXiv:2311.08377 [cs.CL]

[45] Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. RECOMP: Improving Retrieval-Augmented LMs with Context Compression and Selective Augmentation. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net. https://openreview.net/forum?id=mlJLVigNHp

[46] 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, EMNLP 2023, Singapore, December 6-10, 2023. Association for Computational Linguistics, 5364–5375. https://aclanthology.org/2023.emnlp-main.326

[47] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net. https://openreview.net/forum?id=WE_vluYUL-X

[48] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than Retrieve: Large Language Models are Strong Context Generators. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net. https://openreview.net/forum?id=fB0hRu9GZUS

[49] Yuyao Zhang, Zhicheng Dou, Xiaoxi Li, Jiajie Jin, Yongkang Wu, Zhonghua Li, Qi Ye, and Ji-Rong Wen. 2025. Neuro-Symbolic Query Compiler. arXiv:2505.11932 [cs.CL] https://arxiv.org/abs/2505.11932

[50] 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. 2023. A Survey of Large Language Models. CoRR abs/2303.18223 (2023).

[51] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net. https://openreview.net/forum?id=WZH7099tgfM

[52] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large Language Models for Information Retrieval: A Survey. CoRR abs/2308.07107 (2023). https://doi.org/10.48550/ARXIV.2308.07107 arXiv:2308.07107