A Survey of Query Optimization in Large Language Models

Anonymous ACL submission

Abstract

Query Optimization (QO) refers to techniques aimed at improving the operational efficiency and response quality of Large Language Models (LLMs) in processing complex queries, particularly within Retrieval-Augmented Generation (RAG) frameworks. RAG dynamically retrieves current external information to complement model knowledge as a cost-effective solution addressing LLMs' tendencies to generate factually inconsistent outputs. With recent advancements expanding RAG into multi-component systems, QO has become pivotal for optimizing the evidence retrieval phase - critically determining the system's ability to source accurate, multi-faceted supporting information for query resolution. Effective query optimization strategies directly enhance information retrieval performance (e.g., improving recall rates of evidentiary documents) while indirectly strengthening the model's semantic comprehension and final response generation. This paper systematically examines the developmental trajectory of QO techniques through a comprehensive analysis of seminal research. By establishing a structured categorization framework, we aim to synthesize existing QO methodologies in RAG implementations, clarify their technical underpinnings, and emphasize their transformative potential for expanding LLM capabilities across diverse applications.

1 Introduction

Large Language Models (LLMs) have made impressive achievements (Zhao et al., 2023), yet they still encounter notable challenges, particularly in tasks that are domain-specific or heavily reliant on specialized knowledge (Kandpal et al., 2023; Gao et al., 2023b; Zhu et al., 2023b; Huang and Huang, 2024; Verma, 2024; Zhao et al., 2024; Hu and Lu, 2024; Fan et al., 2024; Wu et al., 2024; Peng et al., 2024a; Gupta et al., 2024). One prominent issue is their tendency to produce "hallucinations" when dealing with queries that surpass their

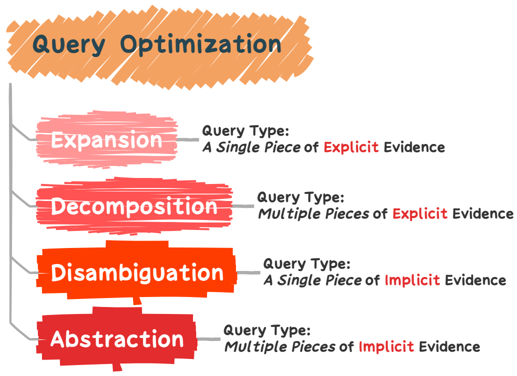

Figure 1: Illustration of four atomic operations in QO. Each atomic operation is classified according to the types of evidence required when solving the query.

training data or necessitate up-to-date information (Zhang et al., 2023b; Tonmoy et al., 2024). To mitigate these challenges, Retrieval-Augmented Generation (RAG) enhances LLMs by retrieving relevant segments, effectively diminishing the production of factually incorrect content. The widespread integration of RAG into LLMs has established it as a crucial technology for the advancement of query solvers and has improved the suitability of LLMs for practical, real-world applications.

Since Lewis et al. (2020) introduced RAG, the field has advanced rapidly, particularly with the emergence of models like ChatGPT. Despite these developments, there is a significant gap in the literature—a thorough analysis of RAG's underlying mechanisms and the progress made in subsequent studies is lacking. Furthermore, the field is characterized by fragmented research focuses and inconsistent terminology for similar methods, which leads to confusion.

RAG typically involves several core concepts, including but not limited to query optimization, information retrieval, and response generation (Zhu

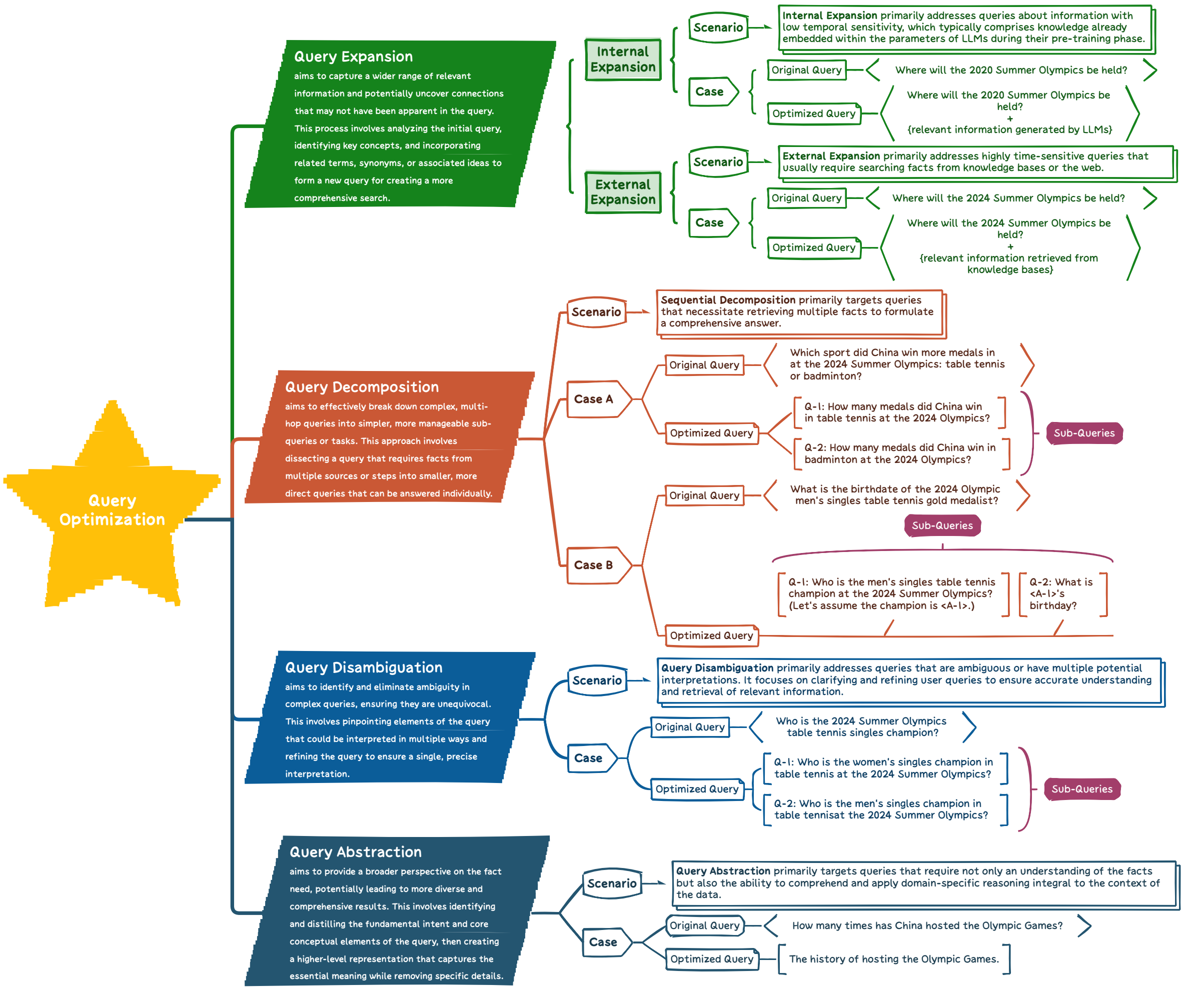

Figure 2: Classification of query optimization techniques in detail.

et al., 2023b; Huang and Huang, 2024; Verma, 2024). Among these, query optimization plays a crucial role in directly determining the relevance of the retrieved information and consequently impacts the quality of the final response. Although query optimization in retrieval-augmented large language models (LLMs) has experienced rapid growth, there has been a lack of systematic synthesis to clarify its broader trajectory. This survey endeavors to fill this gap by mapping out the query optimization process in retrieval-augmented LLMs, charting its evolution, and anticipating future developments. We consider both technical paradigms and research methods, summarizing four main approaches identified in recent LLM-based RAG studies: *Expansion*, *Disambiguation*, *Decomposition*, and *Abstraction*, as shown in Figure 1, and then categorize the corresponding atomic operations for query optimization and map them accordingly. We classify the difficulty of most queries into four types: those that can be solved with a single piece of explicit evidence, those requiring multiple pieces of explicit evidence, those solvable with a single piece of implicit evidence, and those needing multiple pieces of implicit evidence. We then map these queries to different optimization operations respectively for ease of explanation, as shown in Figure 2. Next, we briefly introduce each type of query and the corresponding optimization method, as illustrated in Figure 3.

Overall, this paper aims to meticulously compile and categorize the foundational technical concepts, historical developments, and the range of query optimization methodologies and applications that have emerged since the advent of LLMs. It is designed to equip readers and professionals with a detailed and structured understanding of query optimization in retrieval-augmented LLMs, illumi-

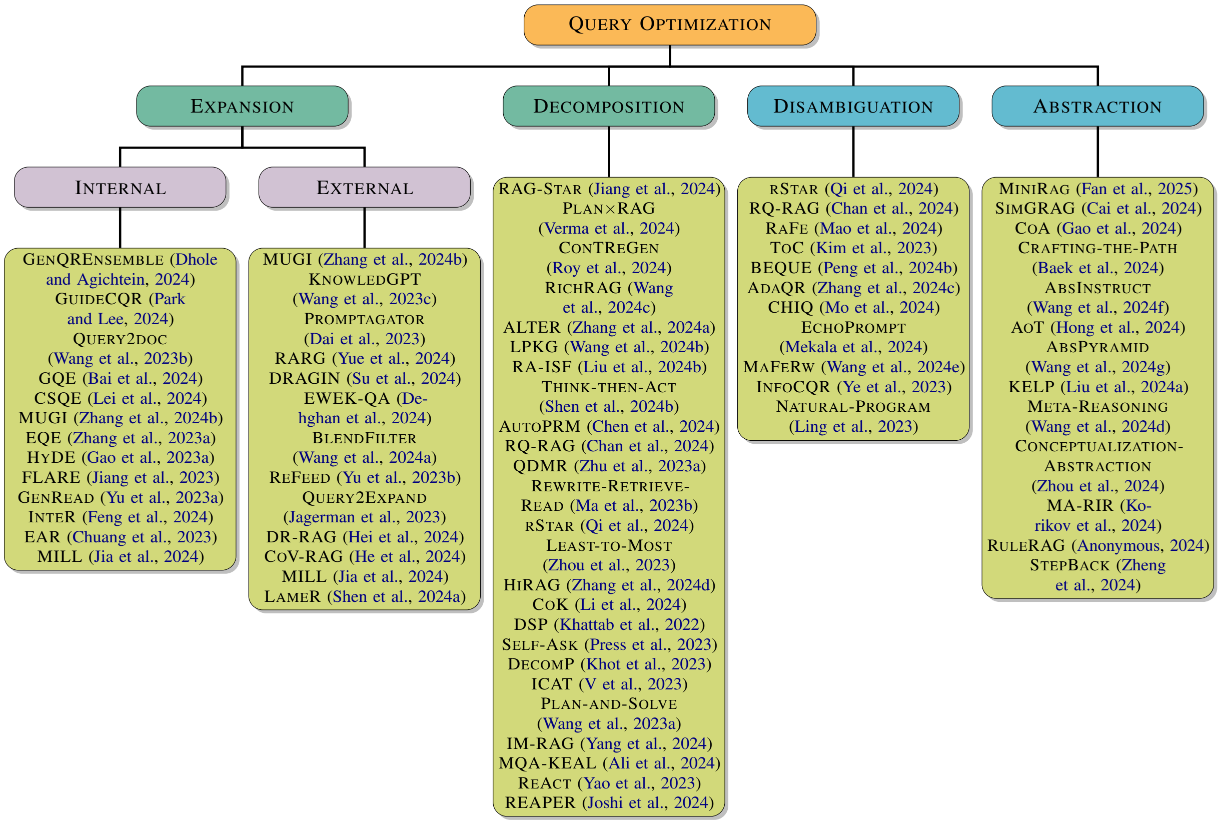

Figure 3: Taxonomy tree of core techniques of query optimization.

nating the evolution of these techniques and speculating on upcoming trends and innovations.

Query optimization techniques summarized in this paper may involve multiple scenarios, including but not limited to retrieval-augmented generation, question answering, etc. Therefore, we uniformly adopt the term "query" to represent terms such as "query", "question", and "problem" in the subsequent content. This survey is organized as follows: Section 2 introduces the stratification of query optimization. The subsequent sections delve into key techniques in query optimization: Section 2.1 explores query expansion, which is further divided into internal expansion (Section 2.1.1) and external expansion (Section 2.1.2). Section 2.2 discusses query decomposition. Section 2.3 and Section 2.4 focus on disambiguation and abstraction. Section 3 addresses the challenges and future directions in this field. Finally, the section of conclusion is presented in Section 4.

2 Stratification of Query Optimization

Query optimization is crucial for enhancing the effectiveness and precision of retrieval-augmented generation using large language models. By refining users' original queries, this process addresses several challenges, including ambiguous semantics, complex requirements, and discrepancies in relevance between the query and target documents. Effective query optimization demands a profound understanding of user intent and query context, especially when dealing with intricate or multifaceted inquiries. When implemented successfully, it significantly improves problem-solving performance, substantially impacting the quality of the model's generated outputs. Ultimately, this enhancement in query processing leads to more accurate and contextually appropriate responses, elevating the overall user experience and increasing the utility of LLMs across various applications.

As previously described, this paper summarizes the query optimization approaches used in recent years, thereby identifying expansion, decomposition, disambiguation, and abstraction as four types of atomic operations in query optimization. Therefore, not only have we classified and matched the queries most suitable for each atomic operation in Figure 1, but we have also distinguished and visualized the effects of each atomic operation in the query processing process, as shown in Figure 4.

In Figure 4, $q_{n,m}$ represents the problem in different states, where $q_{0,0}$ represents the initial problem,

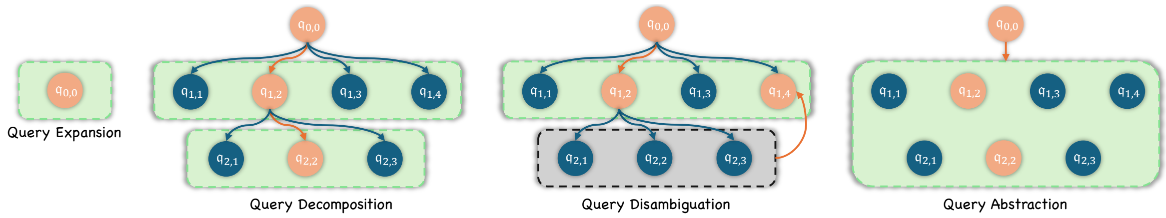

Figure 4: Visualization of the relationship between different queries and query optimization operations.

and $q_{1,1}$ represents the subproblem aimed at solving the first hop, that is, addressing a subproblem within the original problem. Specifically, based on the previous description, query expansion is usually more suitable for solving problems that only require retrieval or obtaining explicit evidence. It thus does not require additional query optimization operations. Query decomposition typically involves complex multi-hop problems that require multiple retrievals. However, each decomposition may not necessarily yield the correct answer, resulting in many redundant exploratory operations. For example, the blue and orange parts represent sub-queries involved in obtaining the correct answer. Query disambiguation refers to reducing the amount of evidence needed to solve the current query by adding additional conditions, which can be understood as a backtracking operation in the figure. Lastly, query abstraction summarizes the problem from a higher level, thereby obtaining background information more conducive to answering the original question. Compared to query decomposition, it can save many retrieval processes.

2.1 Query Expansion

Query expansion techniques (Azad and Deepak, 2019) are critical and effective approaches in enhancing the performance of retrieval-augmented generation, particularly when integrated with LLMs (Weller et al., 2024). Based on the different sources of knowledge, we broadly categorize it into internal expansion and external expansion. The former focuses on maximizing the value of existing information in the original query or the used LLM without relying on external knowledge sources., while the latter introduces supplementary data from outside sources (e.g., Web or Knowledge base) to fill gaps, provide additional context, or broaden the scope of the content.

2.1.1 Internal Expansion

In previous years, researchers have developed various query expansion techniques to enhance information retrieval systems through Large Language Models (LLMs). The seminal GenRead approach (Yu et al., 2023a) employs carefully crafted instructions to prompt LLMs to generate contextual documents that bridge query understanding and answer generation. This paradigm was extended by Query2Doc (Wang et al., 2023b), which uses few-shot prompting to create pseudo-documents containing web-scale knowledge, effectively disambiguating queries and guiding retrieval systems through expanded contextual signals.

Several methods adopt iterative refinement strategies: ReFeed (Yu et al., 2023b) establishes a retrieval-augmented loop by generating initial outputs, retrieving supporting documents, and refining responses through context enrichment. Similarly, InteR (Feng et al., 2024) constructs a synergistic framework where retrieval models expand queries using LLM-generated knowledge, while LLMs enhance prompt formulation through retrieved documents. FLARE (Jiang et al., 2023) introduces anticipatory retrieval based on predicted content trajectories, dynamically triggering new queries when low-confidence tokens emerge.

Alternative approaches leverage hypothetical generation and verification mechanisms: HyDE (Gao et al., 2023a) generates hallucinated documents through zero-shot prompting, then employs contrastive encoding to ground them in real corpus embeddings. MILL (Jia et al., 2024) innovates with mutual verification between LLM-generated sub-queries/documents and retrieved content, ensuring comprehensive coverage through diversity-aware synthesis.

Ensemble strategies further advance the field: GenQREnsemble (Dhole and Agichtein, 2024) enhances retrieval robustness through instruction paraphrasing and keyword set aggregation, while ERRR (Cong et al., 2024) focuses on parametric knowledge distillation and query optimization to maximize relevance precision.

2.1.2 External Expansion

External expansion is a systematic methodology that substantially enriches document content through the strategic integration of relevant information from diverse external knowledge sources. This process effectively enhances the overall contextual depth, informational accuracy, and semantic richness of document corpora by incorporating authoritative facts, current statistical data, and contextual knowledge from curated repositories, specialized datasets, and validated knowledge bases.

LameR (Shen et al., 2024a) employs large language models (LLMs) to augment queries with potential answer candidates obtained through standard retrieval procedures. This approach synthesizes both correct and incorrect in-domain candidates through prompt engineering that combines original queries with retrieved results. GuideCQR (Park and Lee, 2024) addresses conversational query refinement by extracting critical information from initially retrieved documents to guide query reformulation processes. The methodology focuses on distilling essential contextual signals from preliminary search results to optimize subsequent retrieval iterations. CSQE (Lei et al., 2024) utilizes the dual capabilities of LLMs for knowledge extraction and relevance assessment, systematically identifying pivotal sentences within retrieved documents. This corpus-derived knowledge is integrated with LLM-generated expansions through a structured framework that enhances query-document relevance prediction. MUGI (Zhang et al., 2024b) introduces a novel paradigm that leverages LLMs to generate multiple pseudo-references for query expansion. This approach synergistically combines generated references with original queries to optimize performance across both sparse and dense retrieval architectures.

2.2 Question Decomposition

For complex queries, simply searching with the original query often fails to retrieve adequate information. It is crucial for LLMs to first decompose such queries into simpler, answerable sub-queries, and then search for information relevant to these sub-components. By integrating the responses to these sub-queries, LLMs are able to construct a comprehensive response to the original query.

The Demonstrate-Search-Predict (DSP) framework (Khattab et al., 2022) exemplifies this approach through coordinated interaction between LLMs and retrieval models (RMs) within language processing pipelines. This framework orchestrates three core operations: generating bootstrap demonstrations through few-shot learning, executing targeted passage retrieval, and producing evidence-grounded predictions. By decomposing complex tasks into sequential transformations, DSP leverages the complementary strengths of neural reasoning and information retrieval systems for robust problem-solving.

Contemporary prompting strategies reinforce this decomposition paradigm. The LEAST-TO-MOST (Zhou et al., 2023) methodology employs few-shot prompting to recursively divide complex problems into solvable subproblems through chain-of-thought reasoning. Similarly, PLAN-AND-SOLVE (Wang et al., 2023a) prompting operationalizes task decomposition through explicit planning phases, where models first architect solution blueprints before executing stepwise subtask resolution. Both techniques demonstrate enhanced performance through systematic decomposition of cognitive load.

The concept of compositional reasoning is further quantified through SELF-ASK (Press et al., 2023), which identifies the compositionality gap metric. This measure exposes systemic limitations in answer integration by calculating the ratio of failed composite answers relative to correctly solved sub-components. The quantification underscores fundamental challenges in neural reasoning architectures' ability to synthesize partial solutions into coherent final responses.

To address retrieval challenges, approaches like EAR (Chuang et al., 2023) apply a query expansion model to generate a diverse set of queries, using a query reranker to select those that could lead to better retrieval results. Correction of Knowledge (CoK) (Li et al., 2024) first proposes and prepares several preliminary rationales and answers while identifying the relevant knowledge domains. If there is no majority consensus among the answers, CoK corrects the rationales step by step by adapting knowledge from the identified domains, serving as a better foundation for the final response consolidation. ICAT (V et al., 2023) induces reasoning capabilities without any LLM fine-tuning or manual annotation of in-context samples. It transfers the ability to decompose complex queries into simpler ones or generate step-by-step rationales by carefully selecting from available data sources of related tasks.

ReAct (Yao et al., 2023) introduces a paradigm to combine reasoning and acting with LLMs for solving diverse language reasoning and decision-making tasks. ReAct prompts LLMs to generate both verbal reasoning traces and actions on a task in an interleaved manner. This allows the model to perform dynamic reasoning to create, maintain, and adjust high-level plans for acting ("reason to act"), while also interacting with external environments (e.g., Wikipedia) to incorporate additional information into reasoning ("act to reason").

Approaches leveraging query decomposition and iterative refinement have emerged as effective strategies for handling complex queries. AutoPRM (Chen et al., 2024) and RA-ISF (Liu et al., 2024b) both employ multi-stage decomposition frameworks, though with distinct execution mechanisms. AutoPRM decomposes complex problems into manageable sub-queries using a granularity control mechanism, then applies reinforcement learning to optimize sub-query resolution sequentially. RA-ISF integrates text relevance with self-knowledge through iterative sub-query processing, isolating multi-turn queries into independent single-turn tasks before synthesizing their solutions.

Several methods enhance LLM capabilities through structured knowledge integration. RQ-RAG and LPKG (Wang et al., 2024b) exemplify this trend: LPKG improves query planning by grounding knowledge graph patterns into natural language sub-queries, while RQ-RAG employs explicit query rewriting and disambiguation techniques. Similarly, ALTER (Zhang et al., 2024a) enhances table reasoning through multi-perspective question augmentation, generating diverse sub-queries to examine complex problems from complementary angles.

IM-RAG (Yang et al., 2024) introduces a Refiner module to mediate between Retriever and Reasoner components, enabling multi-round knowledge reconciliation. REAPER (Joshi et al., 2024) adopts lightweight planning with smaller LLMs to generate tool-calling blueprints for complex queries. HiRAG (Zhang et al., 2024d) and MQA-KEAL (Ali et al., 2024) both implement multi-hop reasoning through decomposition, with HiRAG employing Chain-of-Thought integration and MQA-KEAL utilizing external structured memory for iterative knowledge retrieval.

Recent advancements focus on sophisticated retrieval-reasoning integration. RichRAG (Wang et al., 2024c) combines latent query facet exploration with multi-faceted document curation, while ConTReGen (Roy et al., 2024) employs tree-structured retrieval for hierarchical information synthesis. Plan×RAG (Verma et al., 2024) formalizes reasoning as directed acyclic graphs, enabling atomic sub-queries with efficient information sharing. Completing this spectrum, RAG-Star (Jiang et al., 2024) implements Monte Carlo Tree Search for deliberative reasoning, autonomously planning intermediate sub-queries through LLM self-knowledge. These approaches collectively demonstrate progressive refinement in aligning retrieval mechanisms with complex reasoning requirements.

2.3 Query Disambiguation

For ambiguous queries with multiple possible answers, relying solely on the original query for information retrieval is inadequate. To deliver complete and nuanced responses, LLMs must learn to clarify the query by identifying the user's intent and then formulate a more targeted search query. After gathering relevant information, LLMs can provide a detailed and comprehensive response. There are mainly two types of approaches for query disambiguation. One is when the query itself is ambiguous, and the other is for multi-turn queries, where it's necessary to rewrite the query by incorporating historical dialogue content to achieve disambiguation (Peng et al., 2024b; Mao et al., 2024).

Ling et al. (2023) early introduces a deductive reasoning format based on the natural language that decomposes the reasoning verification process into a series of step-by-step processes. Each process receives only the necessary context and premises, allowing LLMs to generate precise reasoning steps that are rigorously grounded on prior ones. This approach empowers language models to conduct reasoning self-verification sequentially, significantly enhancing the rigor and trustworthiness of the generated reasoning steps. EchoPrompt (Mekala et al., 2024) introduces a query-rephrasing subtask by employing prompts like "*Let's repeat the query and also think step by step.*". This encourages the model to restate the query in its own words before engaging in reasoning, ensuring better understanding and consistency. Importantly, the prompt used for answer extraction remains consistent across all zero-shot methodologies. ToC (Kim et al., 2023) recursively builds a tree of disambiguations for ambiguous queries by utilizing few-shot prompting and external knowledge. It retrieves relevant facts to generate a comprehensive long-form answer based on this tree, thus providing more accurate and detailed responses. INFOCQR (Ye et al., 2023) introduces a novel "rewrite-then-edit" framework, where LLMs first rewrite the original query and then revise the rewritten query to eliminate ambiguities. The well-designed instructions independently guide the LLMs through the rewriting and editing tasks, resulting in more informative and unambiguous queries.

To further manipulate the disambiguated query, ADAQR (Zhang et al., 2024c) proposes a novel preference optimization approach, which aims to tailor rewriters to better suit retrievers by utilizing conversation answers to model retrievers' preferences. Specifically, the trained rewriter generates several rewrites, which are then used as queries to retrieve passages from a target retriever. Then, ADAQR calculates the conditional probability of the answer given each retrieved passage and the conversation, obtaining the marginal probability of the answer by marginalizing over the set of passages. This marginal probability serves as a reward that quantifies the retrievers' preferences over rewrites and pairs these rewrites based on their rewards to optimize the trained rewriter using direct preference optimization.

MAFERW (Wang et al., 2024e) improves the RAG performance by integrating multi-aspect feedback from both the retrieved documents and the generated responses as rewards to explore the optimal query rewriting strategy. This approach leverages comprehensive feedback to enhance the effectiveness of query rewriting. CHIQ leverages the NLP capabilities of LLMs, such as resolving coreference relations and expanding context, to reduce ambiguity in conversational history. This enhancement improves the relevance of the generated search queries. We investigate various methods for integrating refined conversational history into existing frameworks, including ad-hoc query rewriting, generating pseudo-supervision signals for fine-tuning query rewriting models, and combining both approaches.

2.4 Query Abstraction

For complex multi-hop reasoning tasks, sequential decomposition often fails to produce accurate solutions and may inadvertently introduce additional complexity. Human problem-solvers frequently address this challenge by employing abstraction techniques to derive high-level principles, thereby reducing error propagation in intermediate reasoning steps (Zheng et al., 2024). The STEP-BACK methodology (Zheng et al., 2024) operationalizes this cognitive strategy through structured prompting mechanisms that guide large language models (LLMs) to align their reasoning trajectories with the core intent of the original query, particularly enhancing performance on tasks requiring multi-step logical inference.

This abstraction paradigm has inspired multiple technical implementations. The framework proposed by (Zhou et al., 2024) formalizes conceptual reasoning through abstract query formulations, constraining solutions within verifiable symbolic spaces to promote systematic handling of high-level concepts. Similarly, COA (Gao et al., 2024) transforms conventional chain-of-thought reasoning into abstract variable chains, enabling domain-specific tool integration such as computational modules and web search interfaces. AOT (Hong et al., 2024) advances this approach through a hierarchical skeletal framework that explicitly structures reasoning across multiple abstraction levels, where higher tiers maintain functional objectives while distilling away implementation details—a marked contrast to the less constrained nature of standard chain-of-thought prompting.

Contextual enrichment strategies further enhance reasoning capabilities. Baek et al. (2024) generates meta-level abstraction layers that provide conceptual background for queries, effectively expanding the information landscape available for analysis. For multi-faceted queries, MA-RIR (Korikov et al., 2024) introduces query aspect decomposition, parsing compound queries into distinct topical components to enable targeted reasoning across dimensions.

Recent advancements emphasize structural alignment between queries and knowledge representations. META-REASONING (Wang et al., 2024d) deconstructs query semantics into generalizable symbolic representations, facilitating cross-domain pattern recognition. RULERAG (Anonymous, 2024) implements rule-guided retrieval augmented generation, leveraging logical axioms to retrieve both supportive documents and attributable inference rules. Knowledge graph integration approaches like KELP (Liu et al., 2024a) employ latent semantic path scoring for flexible knowledge extraction, while SIMGRAG (Cai et al., 2024) introduces a two-stage graph alignment process using generated query patterns and graph semantic distance metrics.

For resource-constrained environments, MINI-

RAG (Fan et al., 2025) demonstrates effective abstraction through entity-centric mapping onto heterogeneous knowledge graphs, proving particularly suitable for smaller language models through its emphasis on computationally lightweight entity extraction primitives.

3 Challenges and Future Directions

3.1 Query-Centric Process Reward Model

A promising approach to improving reasoning in LLMs is the use of process reward models (PRMs) (Ma et al., 2023a; Setlur et al., 2024). PRMs provide feedback at each step of a multi-step reasoning process, potentially enhancing credit assignment compared to outcome reward models (ORMs) that only provide feedback at the final step. However, the processes in PRMs generated by chain-of-thought (CoT) prompting methods are usually unpredictable and make it difficult to find the optimal path. Utilizing the optimal path for optimizing complex queries to construct query-centric process reward models may be a simpler and more effective strategy, which means rewards are provided at each sub-query of a multi-step reasoning process.

3.2 Query Optimization Benchmark

Currently, the notable lack of benchmarks for query optimization hinders the consistent assessment and comparison of different query optimization techniques across various scenarios. Typically, the issue is especially prominent in complex contexts, such as optimizing queries for search within multi-turn retrieval-augmented dialogues and in the decomposition of intricate problems. Therefore, developing comprehensive evaluation frameworks and benchmarks may significantly benefit advancements in query optimization techniques, such as existing benchmarks in RAG (Kuo et al., 2024; Xie et al., 2024; Han et al., 2024).

3.3 Improving Query Optimization Efficiency and Quality

Many existing methods fail to pursue the most optimal query optimization paths, relying instead on strategies akin to exhaustive enumeration. This kind of strategy leads to increased computational time and higher search costs, as the system expends resources exploring numerous non-optimal paths. Additionally, it may introduce inconsistent or irrelevant search information, potentially impacting the overall quality and reliability of the results.

Future research should focus on designing efficient algorithms capable of identifying optimal optimization pathways without the need for exhaustive search. Such advancements would reduce time and resource expenditures while enhancing the consistency and accuracy of query optimization outcomes. For example, query decomposition can further be categorized into parallel decomposition and sequential decomposition. Sequential decomposition typically corresponds to multi-hop queries. The reason for this classification is that parallel decomposition usually does not increase additional search time, while sequential decomposition requires iterative searching to solve dependent queries one by one, which typically increases search time as the number of hops increases.

3.4 Enhancing Query Optimization via Post-Performance

A typical paradigm of prompting-based methods involves providing LLMs with several ground-truth optimizing cases (optional) and a task description for the query optimizer. Although LLMs are capable of identifying the potential user intents of a query, they lack awareness of the retrieval quality resulting from the optimized query. This disconnect can result in optimized queries that appear correct but produce unsatisfactory ranking results. While some existing studies have utilized reinforcement learning to adjust the query optimization process based on generation results, a substantial realm of research remains unexplored concerning the integration of ranking results.

4 Conclusion

This in-depth analysis explores the domain of query optimization techniques, with a focus on their application to retrieval-augmented LLMs. Our study encompasses a broad range of optimization methods, providing a comprehensive understanding of the field. By examining the complexities of query optimization, we identify the key challenges and opportunities that arise in this area. As research in this field continues to advance, the development of specialized methodologies tailored to the needs of retrieval-augmented LLMs is crucial for unlocking their full potential across various domains. This survey aims to serve as a valuable resource for retrieval-augmented LLMs, providing a detailed overview of the current landscape and encouraging further investigation into this vital topic.

5 Limitations

The main goal of this paper is to provide a survey of the existing RAG approaches. Since we do not propose new models, there are no potential social risks to the best of our knowledge. Our work may benefit the research community by providing more introspection into the current state-of-the-art retrieval-augmented LLMs.

References

Muhammad Asif Ali, Nawal Daftardar, Mutayyaba Waheed, Jianbin Qin, and Di Wang. 2024. Mqa-keal: Multi-hop question answering under knowledge editing for arabic language. *Preprint*, arXiv:2409.12257.

Anonymous. 2024. RuleRAG: Rule-guided retrieval-augmented generation with language models for question answering. In *Submitted to The Thirteenth International Conference on Learning Representations*. Under review.

Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: A survey. *Inf. Process. Manag.*, 56(5):1698–1735.

Ingeol Baek, Jimin Lee, Joonho Yang, and Hwanhee Lee. 2024. Crafting the path: Robust query rewriting for information retrieval. *CoRR*, abs/2407.12529.

Zechen Bai, Tianjun Xiao, Tong He, Pichao Wang, Zheng Zhang, Thomas Brox, and Mike Zheng Shou. 2024. GQE: generalized query expansion for enhanced text-video retrieval. *CoRR*, abs/2408.07249.

Yuzheng Cai, Zhenyue Guo, Yiwen Pei, Wanrui Bian, and Weiguo Zheng. 2024. Simgrag: Leveraging similar subgraphs for knowledge graphs driven retrieval-augmented generation. *Preprint*, arXiv:2412.15272.

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. *arXiv*, abs/2404.00610.

Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. 2024. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. 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*, pages 1346–1362. Association for Computational Linguistics.

Yung-Sung Chuang, Wei Fang, Shang-Wen Li, Wen-tau Yih, and James R. Glass. 2023. Expand, rerank, and retrieve: Query reranking for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12131–12147. Association for Computational Linguistics.

Youan Cong, Cheng Wang, Pritom Saha Akash, and Kevin Chen-Chuan Chang. 2024. Query optimization for parametric knowledge refinement in retrieval-augmented large language models. *Preprint*, arXiv:2411.07820.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Mohammad Dehghan, Mohammad Ali Alomrani, Sunyam Bagga, David Alfonso-Hermelo, Khalil Bibi, Abbas Ghaddar, Yingxue Zhang, Xiaoguang Li, Jianye Hao, Qun Liu, Jimmy Lin, Boxing Chen, Prasanna Parthasarathi, Mahdi Biparva, and Mehdi Rezagholizadeh. 2024. EWEK-QA : Enhanced web and efficient knowledge graph retrieval for citation-based question answering systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14169–14187. Association for Computational Linguistics.

Kaustubh D. Dhole and Eugene Agichtein. 2024. Genqrensemble: Zero-shot LLM ensemble prompting for generative query reformulation. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part III*, volume 14610 of *Lecture Notes in Computer Science*, pages 326–335. Springer.

Tianyu Fan, Jingyuan Wang, Xubin Ren, and Chao Huang. 2025. Minirag: Towards extremely simple retrieval-augmented generation. *Preprint*, arXiv:2501.06713.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6491–6501. ACM.

Jiazhan Feng, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. 2024. Synergistic interplay between search and large language models for information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9571–9583. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. 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*, pages 1762–1777. Association for Computational Linguistics.

Silin Gao, Jane Dwivedi-Yu, Ping Yu, Xiaoqing Ellen Tan, Ramakanth Pasunuru, Olga Golovneva, Koustuv Sinha, Asli Celikyilmaz, Antoine Bosselut, and Tianlu Wang. 2024. Efficient tool use with chain-of-abstraction reasoning. *Preprint*, arXiv:2401.17464.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (RAG): evolution, current landscape and future directions. *CoRR*, abs/2410.12837.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4354–4374. Association for Computational Linguistics.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. *Preprint*, arXiv:2410.05801.

Zijian Hei, Weiling Liu, Wenjie Ou, Juyi Qiao, Junming Jiao, Guowen Song, Ting Tian, and Yi Lin. 2024. Dr-rag: Applying dynamic document relevance to retrieval-augmented generation for question-answering. *Preprint*, arXiv:2406.07348.

Ruixin Hong, Hongming Zhang, Xiaoman Pan, Dong Yu, and Changshui Zhang. 2024. Abstraction-of-thought makes language models better reasoners. *CoRR*, abs/2406.12442.

Yucheng Hu and Yuxing Lu. 2024. RAG and RAU: A survey on retrieval-augmented language model in natural language processing. *CoRR*, abs/2404.19543.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *CoRR*, abs/2404.10981.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *CoRR*, abs/2305.03653.

Pengyue Jia, Yiding Liu, Xiangyu Zhao, Xiaopeng Li, Changying Hao, Shuaiqiang Wang, and Dawei Yin. 2024. MILL: mutual verification with large language models for zero-shot query expansion. 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*, pages 2498–2518. Association for Computational Linguistics.

Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. 2024. Rag-star: Enhancing deliberative reasoning with retrieval augmented verification and refinement. *Preprint*, arXiv:2412.12881.

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*, pages 7969–7992. Association for Computational Linguistics.

Ashutosh Joshi, Sheikh Muhammad Sarwar, Samarth Varshney, Sreyashi Nag, Shrivats Agrawal, and Juhi Naik. 2024. REAPER: reasoning based retrieval planning for complex RAG systems. *CoRR*, abs/2407.18553.

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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

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. *CoRR*, abs/2212.14024.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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*, pages 996–1009. Association for Computational Linguistics.

Anton Korikov, George Saad, Ethan Baron, Mustafa Khan, Manav Shah, and Scott Sanner. 2024. Multi-aspect reviewed-item retrieval via LLM query decomposition and aspect fusion. *CoRR*, abs/2408.00878.

Tzu-Lin Kuo, Feng-Ting Liao, Mu-Wei Hsieh, Fu-Chieh Chang, Po-Chun Hsu, and Da-Shan Shiu. 2024.

Rad-bench: Evaluating large language models capabilities in retrieval augmented dialogues. *CoRR*, abs/2409.12558.

Yibin Lei, Yu Cao, Tianyi Zhou, Tao Shen, and Andrew Yates. 2024. Corpus-steered query expansion with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 2: Short Papers, St. Julian's, Malta, March 17-22, 2024*, pages 393–401. Association for Computational Linguistics.

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. *CoRR*, abs/2005.11401.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. 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*.

Haochen Liu, Song Wang, Yaochen Zhu, Yushun Dong, and Jundong Li. 2024a. Knowledge graph-enhanced large language models via path selection. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6311–6321. Association for Computational Linguistics.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, and Tianyu Du. 2024b. RA-ISF: learning to answer and understand from retrieval augmentation via iterative self-feedback. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 4730–4749. Association for Computational Linguistics.

Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023a. Let's reward step by step: Step-level reward model as the navigators for reasoning. *CoRR*, abs/2310.10080.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023b. Query rewriting for retrieval-augmented large language models. *CoRR*, abs/2305.14283.

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.

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*, pages 399–432. Association for Computational Linguistics.

Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, and Jian-Yun Nie. 2024. CHIQ: contextual history enhancement for improving query rewriting in conversational search. *CoRR*, abs/2406.05013.

Jeonghyun Park and Hwanhee Lee. 2024. Conversational query reformulation with the guidance of retrieved documents. *CoRR*, abs/2407.12363.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024a. Graph retrieval-augmented generation: A survey. *CoRR*, abs/2408.08921.

Wenjun Peng, Guiyang Li, Yue Jiang, Zilong Wang, Dan Ou, Xiaoyi Zeng, Derong Xu, Tong Xu, and Enhong Chen. 2024b. 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*, pages 20–28. ACM.

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*, pages 5687–5711. Association for Computational Linguistics.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *CoRR*, abs/2408.06195.

Kashob Kumar Roy, Pritom Saha Akash, Kevin Chen-Chuan Chang, and Lucian Popa. 2024. Contregen: Context-driven tree-structured retrieval for open-domain long-form text generation. *Preprint*, arXiv:2410.15511.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. Rewarding progress: Scaling automated process verifiers for LLM reasoning. *CoRR*, abs/2410.08146.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Yibin Lei, Tianyi Zhou, Michael Blumenstein, and Daxin Jiang. 2024a. 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*, pages 15933–15946. Association for Computational Linguistics.

Yige Shen, Hao Jiang, Hua Qu, and Jihong Zhao. 2024b. Think-then-act: A dual-angle evaluated retrieval-augmented generation. *CoRR*, abs/2406.13050.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12991–13013. Association for Computational Linguistics.

S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *CoRR*, abs/2401.01313.

Venktesh V, Sourangshu Bhattacharya, and Avishek Anand. 2023. In-context ability transfer for question decomposition in complex QA. *CoRR*, abs/2310.18371.

Prakhar Verma, Sukruta Prakash Midigeshi, Gaurav Sinha, Arno Solin, Nagarajan Natarajan, and Amit Sharma. 2024. Plan×rag: Planning-guided retrieval augmented generation. *Preprint*, arXiv:2410.20753.

Sourav Verma. 2024. Contextual compression in retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2409.13385.

Haoyu Wang, Tuo Zhao, and Jing Gao. 2024a. Blendfilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering. *CoRR*, abs/2402.11129.

Junjie Wang, Mingyang Chen, Binbin Hu, Dan Yang, Ziqi Liu, Yue Shen, Peng Wei, Zhiqiang Zhang, Jinjie Gu, Jun Zhou, Jeff Z. Pan, Wen Zhang, and Huajun Chen. 2024b. Learning to plan for retrieval-augmented large language models from knowledge graphs. *CoRR*, abs/2406.14282.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. 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*, pages 2609–2634. Association for Computational Linguistics.

Liang Wang, Nan Yang, and Furu Wei. 2023b. 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*, pages 9414–9423. Association for Computational Linguistics.

Shuting Wang, Xin Yu, Mang Wang, Weipeng Chen, Yutao Zhu, and Zhicheng Dou. 2024c. Richrag: Crafting rich responses for multi-faceted queries in retrieval-augmented generation. *CoRR*, abs/2406.12566.

Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. 2023c. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *CoRR*, abs/2308.11761.

Yiming Wang, Zhuosheng Zhang, Pei Zhang, Baosong Yang, and Rui Wang. 2024d. Meta-reasoning: Semantics-symbol deconstruction for large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 622–643. Association for Computational Linguistics.

Yujing Wang, Hainan Zhang, Liang Pang, Binghui Guo, Hongwei Zheng, and Zhiming Zheng. 2024e. Maferw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models. *CoRR*, abs/2408.17072.

Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Y. Wong, and Simon See. 2024f. Absinstruct: Eliciting abstraction ability from llms through explanation tuning with plausibility estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 973–994. Association for Computational Linguistics.

Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2024g. Abspyramid: Benchmarking the abstraction ability of language models with a unified entailment graph. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3991–4010. Association for Computational Linguistics.

Orion Weller, Kyle Lo, David Wadden, Dawn J. Lawrie, Benjamin Van Durme, Arman Cohan, and Luca Soldaini. 2024. When do generative query and document expansions fail? A comprehensive study across methods, retrievers, and datasets. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 1987–2003. Association for Computational Linguistics.

Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, and Chun Jason Xue. 2024. Retrieval-augmented generation for natural language processing: A survey. *CoRR*, abs/2407.13193.

Kaige Xie, Philippe Laban, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. 2024. Do RAG systems cover what matters? evaluating and optimizing responses with sub-question coverage. *CoRR*, abs/2410.15531.

Diji Yang, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Jie Yang, and Yi Zhang. 2024. IM-RAG: multi-round retrieval-augmented generation through learning inner monologues. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 730–740. ACM.

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.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5985–6006. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023a. 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.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023b. Improving language models via plug-and-play retrieval feedback. *Preprint*, arXiv:2305.14002.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. 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*, pages 5628–5643. Association for Computational Linguistics.

Han Zhang, Yuheng Ma, and Hanfang Yang. 2024a. ALTER: augmentation for large-table-based reasoning. *CoRR*, abs/2407.03061.

Le Zhang, Yihong Wu, Qian Yang, and Jian-Yun Nie. 2024b. Exploring the best practices of query expansion with large language models. *Preprint*, arXiv:2401.06311.

Tianhua Zhang, Kun Li, Hongyin Luo, Xixin Wu, James R. Glass, and Helen Meng. 2024c. Adaptive query rewriting: Aligning rewriters through marginal probability of conversational answers. *CoRR*, abs/2406.10991.

Xiaoming Zhang, Ming Wang, Xiaocui Yang, Daling Wang, Shi Feng, and Yifei Zhang. 2024d. Hierarchical retrieval-augmented generation model with rethink for multi-hop question answering. *Preprint*, arXiv:2408.11875.

Yanan Zhang, Weijie Cui, Yangfan Zhang, Xiaoling Bai, Zhe Zhang, Jin Ma, Xiang Chen, and Tianhua Zhou. 2023a. Event-centric query expansion in web search. In *Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics: Industry Track, ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 464–475. Association for Computational Linguistics.

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. 2023b. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna Qiu, and Lili Qiu. 2024. Retrieval augmented generation (RAG) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *CoRR*, abs/2409.14924.

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.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. 2024. Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. In *The Twelfth International Conference on Learning Representations*, volume abs/2310.06117.

Ben Zhou, Hongming Zhang, Sihao Chen, Dian Yu, Hongwei Wang, Baolin Peng, Dan Roth, and Dong Yu. 2024. Conceptual and unbiased reasoning in language models. *CoRR*, abs/2404.00205.

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.

Wang Zhu, Jesse Thomason, and Robin Jia. 2023a. Chain-of-questions training with latent answers for robust multistep question answering. In *Proceedings* *of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 8845–8860. Association for Computational Linguistics.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023b. Large language models for information retrieval: A survey. *CoRR*, abs/2308.07107.