Query Expansion Revisited: From Classical Retrieval to Agentic Search

Amin Bigdeli
University of Waterloo

Radin Hamidi Rad
Mila – Quebec AI Institute

Hai Son Le
Toronto Metropolitan University

Yibin Lei
University of Amsterdam

Andrew Yates
Johns Hopkins University

Negar Arabzadeh
University of California, Berkeley

Charles L. A. Clarke
University of Waterloo

Ebrahim Bagheri
University of Toronto

Abstract

The emergence of large language models and agentic search systems is transforming query expansion, elevating it from a standalone preprocessing step to a central reasoning component in modern retrieval pipelines. The shift from classical term-selection methods to LLM-based generation has opened new possibilities for expansion, while deep search agents, retrieval-augmented generation, and conversational retrieval have introduced fundamentally new settings in which query expansion must operate. This tutorial offers a comprehensive treatment of query expansion methods, tracing their progression from classical statistical techniques through neural and transformer-based approaches to LLM-based generation, while covering evaluation methodology, cross-paradigm deployment, failure analysis, and emerging research opportunities. The tutorial is accompanied by hands-on experimentation for implementing and comparing expansion methods across retrieval settings.

1 Motivation and Overview

Query expansion has been a central technique in information retrieval for decades, originally developed to address the fundamental vocabulary mismatch between user queries and relevant documents [3, 8, 35, 36]. Classical methods such as relevance feedback [35, 36], pseudo-relevance feedback [23], and statistical term co-occurrence [34] established the core principle that enriching a query with additional terms can substantially improve retrieval effectiveness. These techniques, along with related query refinement methods such as query rewriting, decomposition, and clarification, were traditionally treated as standalone components within retrieval pipelines, applied as a preprocessing step before a single retrieval pass [3, 8, 33].

The landscape of query expansion has since evolved considerably. Neural approaches introduced the ability to expand queries using semantic similarity rather than surface-level term overlap, first through word embeddings [13, 22, 47] and later through transformer-based models [11, 31, 42, 51, 52] that condition expansion on the full query context. More recently, large language models have opened fundamentally new possibilities for query expansion. Methods based on LLMs can generate pseudo-documents [16, 18, 40, 48], produce targeted expansion terms grounded in the target corpus [7, 10, 19, 24, 38], and iteratively refine queries based on retrieval feedback [4, 15, 25]. These advances have transformed query expansion from a relatively constrained term-selection problem into a flexible generation and reasoning task.

At the same time, LLMs are reshaping how users interact with search systems and how modern retrieval pipelines are designed. Search is increasingly moving beyond one-shot querying toward conversational search [14, 29], retrieval-augmented generation (RAG) [20, 26], and agentic search [9, 17, 21, 27, 49]. In these settings, query expansion is no longer a peripheral preprocessing step but a central mechanism that drives the entire retrieval process. Deep search agents [17, 21, 27, 45], for example, may repeatedly reformulate, expand, decompose, and route queries based on intermediate retrieval results as part of multi-step research workflows. In RAG systems, query quality directly determines which documents are supplied to the generation model and thus the faithfulness of the final output. In conversational search [14, 29, 30, 43], effective reformulation must account for dialogue context and evolving user intent across turns. Across all of these paradigms, the ability to formulate effective queries is increasingly critical.

This convergence of advances in LLM-based expansion methods and new retrieval paradigms makes the present moment particularly important for revisiting query expansion. For researchers, there is a need to understand how classical principles translate to LLM-based methods, where existing approaches succeed and fail, and what open problems remain in areas such as hallucination, topic drift, and reproducibility across retrievers [1, 6]. For practitioners, the challenge lies in selecting appropriate expansion strategies for diverse system requirements and in understanding the trade-offs between effectiveness, latency, and computational cost.

This tutorial provides a comprehensive treatment of query expansion spanning its evolution, from classical statistical methods through neural and transformer-based approaches to LLM-based techniques. Rather than treating query expansion as an isolated retrieval component, the tutorial situates it within the query refinement landscape, distinguishing query formulation, expansion, rewriting, decomposition, clarification, routing, and selection. We trace the methodological progression from thesaurus-based expansion and statistical term selection through (pseudo-)relevance feedback to embedding-based, sequence-to-sequence, transformer-based, and LLM-based approaches. The LLM-based portion is organized around single-pass generation, corpus-grounded expansion, and iterative feedback, with attention to prompting strategies, integration of generated content with the original query, and failure modes such as hallucination, verbosity, and topic drift.

The tutorial further covers how query expansion operates across retrieval scenarios, including ad hoc retrieval, conversational search, RAG systems, agentic and deep search, and domain-specific retrieval, discussing how expansion interacts with user intent, task context, corpus characteristics, and downstream components. We also address evaluation and practical system design, covering query-level help/hurt analysis, robustness, reproducibility across lexical and dense retrievers, and cost considerations. Throughout, participants will engage in hands-on experimentation with QueryGym [5] (https://querygym.com/), using live demonstrations, pre-configured notebooks, and an interactive dashboard to implement and compare expansion methods. Our goal is to equip researchers and practitioners with the foundations and tools to advance query expansion research and deploy these methods in modern search systems.

2 Objectives

This tutorial aims to provide attendees with a thorough understanding of current query expansion methods, from classical techniques to LLM-based approaches, and enable them to apply these methods in their own information retrieval tasks, deep search agents, and RAG pipelines. Attendees will learn to navigate the full design space, select appropriate techniques for a given retrieval setting and system constraints, diagnose common failure modes of LLM-based expansion such as hallucination and topic drift, and design expansion strategies tailored to emerging paradigms. A central objective is to ground current research in the historical development of the field, as understanding how query expansion methods have evolved is essential for advancing them further.

Attendees will also acquire the skills to critically assess expansion methods beyond aggregate metrics, considering query-level effectiveness, robustness, and reproducibility across different retrieval methods. The tutorial will further familiarize attendees with current challenges, limitations, and open research directions in query expansion. Through hands-on experimentation, attendees will learn how to reproduce, benchmark, and compare expansion methods across retrieval settings, obtaining reusable tools and workflows for their own research.

3 Tutorial Outline

This half-day tutorial is organized into six parts totaling 3 hours of content plus short breaks.

Part I [10 Minutes]: The Query Refinement Landscape. The tutorial opens by establishing a taxonomy of query refinement tasks [3, 8] that serves as the conceptual framework for the remainder of the tutorial. We define and distinguish between query formulation, query expansion, query rewriting, query decomposition, query clarification, and query routing and selection, providing attendees with a shared vocabulary for understanding how the methods presented in subsequent parts relate to one another.

Part II [15 Minutes]: Traditional Query Expansion. Building on this taxonomy, this part covers classical query expansion approaches motivated by the vocabulary mismatch problem. We discuss statistical expansion methods based on term co-occurrence [34], relevance feedback [35, 36], and pseudo-relevance feedback [23].

Part III [35 Minutes]: Neural and Transformer-Based Query Expansion. This part covers query expansion methods that leverage neural representations and pre-trained language models. We begin with word embedding-based expansion [13, 22, 47], where locally trained embeddings are used to identify semantically related expansion terms beyond exact lexical overlap. We then cover sequence-to-sequence and neural reformulation methods [32, 46] that cast query expansion as a generation task using encoder-decoder architectures. Building on these, we discuss transformer-based query expansion approaches [11, 31, 42, 51, 52] that leverage contextualized representations to select or generate expansion terms conditioned on the full query context.

Part IV [50 Minutes]: LLM-Based Query Expansion. The central part of the tutorial provides comprehensive coverage of LLM-based query expansion methods, organized into three categories. (1) *Single-pass generation methods* [12, 16, 18, 40, 41, 48], covering pseudo-document and hypothetical document generation, prompting strategies, ensemble-based generative reformulation, question decomposition and answer-driven expansion [37], and best practices for integrating expansion text with original queries. (2) *Corpus-grounded expansion* [7, 10, 19, 24, 38], which addresses the misalignment between LLM-generated text and target corpus vocabulary through candidate-aware expansion, corpus-steered term selection, and pattern-guided reformulation. (3) *Iterative feedback* [4, 15, 25], where retrieval outcomes are turned into structured signals for successive expansion rounds. This part is accompanied by live demonstrations using QueryGym [5], allowing attendees to interactively experiment with the presented methods and observe their behavior across different retrieval settings.

Part V [25 Minutes]: Query Expansion Across Different Search Paradigms. This part covers how query expansion operates across search scenarios beyond classical ad hoc retrieval. We discuss conversational search [14, 29, 30, 43], where reformulation must account for dialogue context and evolving user intent across turns. We then cover RAG systems [20, 26, 28, 50], where query quality directly determines which documents are fed to the generation model and thus the faithfulness and accuracy of generated answers. We present agentic and deep search [9, 17, 21, 27, 49], where agents iteratively formulate, execute, and refine queries as part of multi-step research workflows, and discuss how these settings require query expansion strategies that differ fundamentally from single-shot retrieval. We also address domain-specific and structured search scenarios, where specialized terminology and vocabulary constraints impose additional requirements on expansion methods.

Part VI [45 Minutes]: Evaluation, Practical System Design, and Discussion. The final part of the tutorial addresses three interconnected themes. We begin with *multi-dimensional evaluation* of query expansion methods, moving beyond aggregate retrieval effectiveness to query-level analysis of which queries benefit from expansion and which are harmed, query difficulty prediction [2] as a tool for selective expansion, and robustness and failure analysis covering topic drift and hallucinated expansion terms [1]. We also discuss how expansion methods perform differently across retriever types [6], showing that gains under lexical retrieval do not consistently transfer to neural retrievers. We then turn to *practical system design*, discussing where query expansion fits within a retrieval pipeline and trade-offs between effectiveness, latency, and computational cost. The tutorial concludes with a *discussion and live demonstration* using QueryGym [5], where attendees learn how to implement their own reformulation methods, run established baselines across retrieval methods, and compare results on standard benchmarks through pre-configured Google Colab notebooks.

4 Relevance to the CIKM Community and Related Tutorials

Query expansion is a foundational component of information retrieval and knowledge management. The methods covered in this tutorial have direct implications for a broad range of tasks relevant to the CIKM community. Query expansion enhances first-stage retrieval in search systems, directly influences the quality of documents supplied to RAG pipelines, and plays an increasingly important role in conversational and agentic search systems that manage and synthesize knowledge at scale. Given that LLM-based query expansion is rapidly being adopted in production systems across both industry and academia, we believe the community would benefit from a systematic consolidation of methods, evaluation findings, and practical design guidance in this area.

While several recent tutorials have addressed topics adjacent to query expansion, none have provided dedicated coverage of this area. The following tutorials can be considered complementary and synergistic to the theme of our proposed tutorial:

(1) *R2LLMs: Retrieval and Ranking with LLMs* [53] (SIGIR 2025). This tutorial provides a broad overview of how generative LLMs are transforming document retrieval and ranking. While it covers retrieval and ranking with LLMs at a high level, it does not focus on query expansion as a dedicated strategy. Our tutorial provides deep, method-level coverage of the full expansion pipeline accompanied by reproducibility analysis and hands-on experimentation.
(2) *Recent Advances in Generative Information Retrieval* [39] (WWW 2024). This tutorial covers generative retrieval paradigms including document identifier generation. While both tutorials involve generative approaches to IR, the focus differs substantially: generative retrieval replaces traditional pipelines with sequence-to-sequence generation of document identifiers, whereas our tutorial focuses on using LLMs to expand queries within existing retrieval frameworks.
(3) *Query Understanding in LLM-based Conversational Information Seeking* [44] (SIGIR 2025). This tutorial explores query understanding in multi-turn conversational search, including query reformulation as one component. Our tutorial differs by providing comprehensive coverage of query expansion across retrieval paradigms for both ad hoc and cross-scenario settings, with dedicated reproducibility analysis and hands-on toolkit.

To the best of our knowledge, no existing tutorial provides a unified treatment of query expansion spanning the full pipeline from traditional methods through neural and LLM-based techniques to cross-scenario deployment, accompanied by reproducibility analysis and hands-on experimentation with a dedicated toolkit.

5 Intended Audience and Prerequisites

The target audience for this tutorial includes IR researchers, NLP practitioners, industry engineers working on search and RAG systems, and graduate students seeking an entry point into query expansion research. Familiarity with lexical and dense retrieval, standard evaluation metrics, and large language models at the level of prompting and text generation is recommended, though the tutorial provides sufficient background on foundational concepts to remain accessible to a broad audience. No deep machine learning or model training expertise is required, as the hands-on session uses pre-configured environments and API-based LLM access.

6 Presenters

Amin Bigdeli is a Ph.D. student at the University of Waterloo, focusing on ad hoc information retrieval and robustness of neural ranking models. He has published in leading venues including ACM TOIS, ACM TIST, SIGIR, CIKM, ECIR, and WWW. He has previously delivered tutorials on gender fairness in information retrieval systems at SIGIR 2022 and ECIR 2023.
Radin Hamidi Rad is a Postdoctoral Researcher at Mila – Quebec AI Institute. His research focuses on information retrieval, agentic deep search, and adversarial robustness in search systems. He received his Ph.D. in Computer Engineering from Toronto Metropolitan University, where he studied search in knowledge graphs. His work has appeared in venues including ACM TOIS, ACM TIST, IP&M, Information Sciences, SIGIR, CIKM, and WWW.
Hai Son Le is a Master's student at Toronto Metropolitan University, working on information retrieval, fair ranking systems, and query reformulation. His research spans LLM-based retrieval, retrieval-augmented generation, and neural ranking, with publications in venues such as SIGIR, WSDM, ICLR, and WWW.
Yibin Lei is a PhD student at the University of Amsterdam, working on information retrieval and search in multi-lingual and low-resource scenarios. Yibin co-organized the hackathon session of the European Summer School on IR 2024 (ESSIR 2024) and gave lectures on neural information retrieval as part of the Master's course *Information Retrieval 1* at the University of Amsterdam.
Andrew Yates is a Senior Research Scientist at Johns Hopkins University. His research focuses on developing content-based neural ranking methods and leveraging them to improve search and downstream tasks. He has co-authored a variety of papers on neural ranking methods and a book on transformer-based neural methods: "Pretrained Transformers for Text Ranking: BERT and Beyond."
Negar Arabzadeh is a Postdoctoral Researcher at the University of California, Berkeley, working on information retrieval and agentic search systems. She received her Ph.D. from the University of Waterloo. Her research has appeared in venues including SIGIR, CIKM, WWW, EMNLP, and ICML. She has contributed to tutorials at SIGIR, ECIR, and WSDM, organized workshops at NeurIPS and SIGIR, and served as a track chair for SIGIR, ECIR, and CHIIR.
Charles L. A. Clarke is a Professor in the School of Computer Science at the University of Waterloo, Canada. His research focuses on data-intensive tasks and efficiency, including search, ranking, question answering, and other problems involving human language data at scale. He has also worked on search engine technology for both Microsoft Bing and Facebook Search.
Ebrahim Bagheri is a Professor and the Director of the Laboratory for Systems, Software and Semantics (LS^3) at the University of Toronto. He formerly held a Canada Research Chair (Tier II) in Social Information Retrieval and an NSERC Industrial Research Chair in Social Media Analytics. He currently leads the NSERC Program on Responsible AI. He is an Associate Editor for ACM

Transactions on Intelligent Systems and Technology (TIST) and IEEE Transactions on Network Science and Engineering (TNSE).

7 Type of Support Materials

We will publicly share a GitHub repository prior to the conference so that participants can familiarize themselves with the content. The repository will include a comprehensive slide deck, links to code, pre-configured Google Colab notebooks for the hands-on session, and a curated reading list. In addition, we provide an interactive web-based dashboard for QueryGym, which allows participants to experiment with different query expansion methods, retrieval paradigms, and benchmark datasets in real time without any local installation. The dashboard offers a visual interface for comparing method performance across configurations, enabling attendees to engage with the hands-on component directly during the tutorial session. Attendees are encouraged but not required to bring laptops, as all hands-on activities are accessible through a web browser.

References

[1] Kenya Abe, Kunihiro Takeoka, Makoto P. Kato, and Masafumi Oyamada. 2025. LLM-Based Query Expansion Fails for Unfamiliar and Ambiguous Queries. In *Proceedings of SIGIR*.

[2] Negar Arabzadeh, Andrew Drozdov, Michael Bendersky, and Matei Zaharia. 2026. Can QPP Choose the Right Query Variant? Evaluating Query Variant Selection for RAG Pipelines. (2026).

[3] Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query Expansion Techniques for Information Retrieval: A Survey. *Information Processing & Management* (2019).

[4] Amin Bigdeli, Negar Arabzadeh, Radin Hamidi Rad, et al. 2026. ADORE: Iterative Query Expansion with Retrieval-Grounded Relevance Feedback. (2026).

[5] Amin Bigdeli, Radin Hamidi Rad, Mert Incesu, et al. 2026. QueryGym: A Toolkit for Reproducible LLM-Based Query Reformulation. In *Companion Proceedings of WWW*.

[6] Amin Bigdeli, Radin Hamidi Rad, Hai Son Le, et al. 2026. A Reproducibility Study of LLM-Based Query Reformulation. (2026).

[7] Amin Bigdeli, Mert Incesu, Negar Arabzadeh, et al. 2026. *ReFormeR: Learning and Applying Explicit Query Reformulation Patterns*.

[8] Claudio Carpineto and Giovanni Romano. 2012. A Survey of Automatic Query Expansion in Information Retrieval. *Comput. Surveys* (2012).

[9] Zijian Chen, Xueguang Ma, Shengyao Zhuang, et al. 2026. AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. (2026).

[10] Yung-Sung Chuang, Wei Fang, Shang-Wen Li, et al. 2023. Expand, Rerank, and Retrieve: Query Reranking for Open-Domain Question Answering. In *Findings of ACL*.

[11] Giulio D'Erasmo, Giovanni Trappolini, Fabrizio Silvestri, and Nicola Tonellotto. 2025. Eclipse: Contrastive Dimension Importance Estimation with Pseudo-Irrelevance Feedback for Dense Retrieval. In *Proceedings of ICTIR*.

[12] Kaustubh D. Dhole, Ramraj Chandradevan, and Eugene Agichtein. 2024. Generative Query Reformulation Using Ensemble Prompting, Document Fusion, and Relevance Feedback. (2024).

[13] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. 2016. Query Expansion with Locally-Trained Word Embeddings. In *Proceedings of ACL*.

[14] Ahmed Elgohary, Denis Peskov, and Jordan Lee Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proceedings of EMNLP-IJCNLP*.

[15] Jiazhan Feng, Chongyang Tao, Xiubo Geng, et al. 2024. Synergistic Interplay Between Search and Large Language Models for Information Retrieval. In *Proceedings of ACL*.

[16] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise Zero-Shot Dense Retrieval Without Relevance Labels. In *Proceedings of ACL*.

[17] Tz-Huan Hsu, Jheng-Hong Yang, and Jimmy Lin. 2026. Rethinking Agentic Search with Pi-Serini: Is Lexical Retrieval Sufficient? (2026).

[18] Rolf Jagerman, Honglei Zhuang, Zhen Qin, et al. 2023. Query Expansion by Prompting Large Language Models. (2023).

[19] Pengyue Jia, Yiding Liu, Xiangyu Zhao, et al. 2024. MILL: Mutual Verification with Large Language Models for Zero-Shot Query Expansion. In *Proceedings of NAACL*.

[20] Zhengbao Jiang, Frank F. Xu, Luyu Gao, et al. 2023. Active Retrieval Augmented Generation. In *Proceedings of EMNLP*.

[21] Bowen Jin, Hansi Zeng, Zhenrui Yue, et al. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. (2025).

[22] Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query Expansion Using Word Embeddings. In *Proceedings of CIKM*.

[23] Victor Lavrenko and W. Bruce Croft. 2017. Relevance-Based Language Models. In *ACM SIGIR Forum*.

[24] Yibin Lei, Yu Cao, Tianyi Zhou, et al. 2024. Corpus-Steered Query Expansion with Large Language Models. In *Proceedings of EACL*.

[25] Yibin Lei, Tao Shen, and Andrew Yates. 2025. ThinkQE: Query Expansion via an Evolving Thinking Process. (2025).

[26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems* (2020).

[27] Kuan Li, Zhongwang Zhang, Huifeng Yin, et al. 2025. WebSailor: Navigating Super-Human Reasoning for Web Agent. (2025).

[28] Xinbei Ma, Yeyun Gong, Pengcheng He, et al. 2023. Query Rewriting in Retrieval-Augmented Large Language Models. In *Proceedings of EMNLP*.

[29] Kelong Mao, Zhicheng Dou, Fengran Mo, et al. 2023. Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search. In *Findings of EMNLP*.

[30] Fengran Mo, Kelong Mao, Yutao Zhu, et al. 2023. ConvGQR: Generative Query Reformulation for Conversational Search. In *Proceedings of ACL*.

[31] Shahrzad Naseri, Jeffrey Dalton, Andrew Yates, and James Allan. 2021. CEQE: Contextualized Embeddings for Query Expansion. In *Proceedings of ECIR*.

[32] Rodrigo Nogueira and Kyunghyun Cho. 2017. Task-Oriented Query Reformulation with Reinforcement Learning. In *Proceedings of EMNLP*.

[33] Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D Manning. 2019. Answering complex open-domain questions through iterative query generation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*.

[34] Stephen E. Robertson. 1990. On Term Selection for Query Expansion. *Journal of Documentation* (1990).

[35] S. E. Robertson and K. Spärck Jones. 1976. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science* (1976).

[36] Joseph John Rocchio Jr. 1971. Relevance Feedback in Information Retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing* (1971).

[37] Wonduk Seo and Seunghyun Lee. 2025. QA-Expand: Multi-Question Answer Generation for Enhanced Query Expansion in Information Retrieval. (2025).

[38] Tao Shen, Guodong Long, Xiubo Geng, et al. 2024. Retrieval-Augmented Retrieval: Large Language Models are Strong Zero-Shot Retriever. In *Findings of ACL*.

[39] Yubao Tang, Ruqing Zhang, Weiwei Sun, et al. 2024. Recent Advances in Generative Information Retrieval. In *Companion Proceedings of WWW*.

[40] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. In *Proceedings of EMNLP*.

[41] Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2023. Generative Query Reformulation for Effective Adhoc Search. (2023).

[42] Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2023. ColBERT-PRF: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval. *ACM Transactions on the Web* (2023).

[43] Shi Yu, Jiahua Liu, Jingqin Yang, et al. 2020. Few-Shot Generative Conversational Query Rewriting. In *Proceedings of SIGIR*.

[44] Yifei Yuan, Zahra Abbasiantaeb, Mohammad Aliannejadi, and Yang Deng. 2025. Query Understanding in LLM-Based Conversational Information Seeking. In *Proceedings of SIGIR*.

[45] Xinyu Gao et al. Yunfan Gao, Yun Xiong. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. (2023).

[46] Salah Zaiem and Fatiha Sadat. 2019. Sequence to sequence learning for query expansion. In *proceedings of the AAAI conference on artificial intelligence*.

[47] Hamed Zamani and W. Bruce Croft. 2016. Embedding-Based Query Language Models. In *Proceedings of ICTIR*.

[48] Le Zhang, Yihong Wu, Qian Yang, and Jian-Yun Nie. 2024. Exploring the Best Practices of Query Expansion with Large Language Models. In *Findings of EMNLP*.

[49] Weinan Zhang, Junwei Liao, Ning Li, et al. 2024. Agentic Information Retrieval. (2024).

[50] 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 *International Conference on Learning Representations*.

[51] Zhi Zheng, Kai Hui, Ben He, et al. 2020. BERT-QE: Contextualized Query Expansion for Document Re-ranking. In *Findings of EMNLP*.

[52] Zhi Zheng, Kai Hui, Ben He, et al. 2021. Contextualized Query Expansion via Unsupervised Chunk Selection for Text Retrieval. *Information Processing & Management* (2021).

[53] Guido Zuccon, Shengyao Zhuang, and Xueguang Ma. 2025. R2LLMs: Retrieval and Ranking with LLMs. In *Proceedings of SIGIR*.

Statement of Generative AI Usage

In preparing this manuscript, the authors used Claude only as a writing assistant to improve the grammar, wording, and readability of the manuscript. The authors critically evaluated, edited, and approved all suggestions and take full responsibility for the content of the final manuscript.