Query Understanding in LLM-based Conversational Information Seeking

Yifei Yuan
University of Copenhagen
Denmark
yiya@di.ku.dk

Zahra Abbasiantaeb
University of Amsterdam
The Netherlands
z.abbasiantaeb@uva.nl

Yang Deng
Singapore Management University
Singapore
ydeng@smu.edu.sg

Mohammad Aliannejadi
University of Amsterdam
The Netherlands
m.aliannejadi@uva.nl

Abstract

Query understanding in Conversational Information Seeking (CIS) involves accurately interpreting user intent through context-aware interactions. This includes resolving ambiguities, refining queries, and adapting to evolving information needs. Large Language Models (LLMs) enhance this process by interpreting nuanced language and adapting dynamically, improving the relevance and precision of search results in real-time. In this tutorial, we explore advanced techniques to enhance query understanding in LLM-based CIS systems. We delve into LLM-driven methods for developing robust evaluation metrics to assess query understanding quality in multi-turn interactions, strategies for building more interactive systems, and applications like proactive query management and query reformulation. We also discuss key challenges in integrating LLMs for query understanding in conversational search systems and outline future research directions. Our goal is to deepen the audience's understanding of LLM-based conversational query understanding and inspire discussions to drive ongoing advancements in this field.

CCS Concepts

• **Information systems** → *Query suggestion*; *Query reformulation*; *Query intent*.

Keywords

Query understanding, Large language models, Conversational search

ACM Reference Format:
Yifei Yuan, Zahra Abbasiantaeb, Yang Deng, and Mohammad Aliannejadi. 2025. Query Understanding in LLM-based Conversational Information Seeking. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3701716.3715869

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3715869

1 Introduction

Query understanding refers to a system's ability to accurately interpret a user's intent, even when expressed through incomplete, vague, or ambiguous queries [17]. This becomes especially critical in CIS, where users tend to express their needs in less direct or structured ways, compared to ad-hoc retrieval. In typical CIS scenarios, users often begin with vague or imprecise queries and progressively refine their queries, ask follow-up questions, or even shift contexts mid-conversation [38]. These dynamics pose significant challenges for the system, as it must engage in a dialogue to clarify and refine the user's intent, ensuring that the responses are both accurate and relevant throughout the interaction.

In recent years, LLMs, such as GPT-4, have demonstrated remarkable capabilities in tasks far beyond Natural Language Understanding (NLU), revolutionizing the way users interact with Information Retrieval (IR) systems. These models excel at handling natural language queries with exceptional accuracy and contextual awareness, greatly enhancing the overall user experience in retrieving relevant information. In the context of CIS systems, LLMs significantly impact query understanding in several key areas, such as conversational context understanding [31], query clarification [20], user simulation [36], query reformulation [34].

Despite the advancements LLMs have brought, several open challenges remain: (i) developing robust evaluation metrics, as it remains challenging to establish effective ways to measure how well a system understands and addresses user intent across dynamic, multi-turn conversations [42]; (ii) Improving conversational interaction by making exchanges smoother and more natural [29]; (iii) increasing user proactivity, encouraging users to take a more active role in refining and clarifying their searches [20]; and (iv) handling ambiguity in user queries, as users frequently submit vague or incomplete queries, requiring LLMs to strike a balance between generating appropriate responses and requesting clarifications [26]. This tutorial addresses these challenges by exploring advanced techniques to improve query understanding in LLM-based CIS.

1.1 Query Understanding Evaluation

Query understanding refers to the process where a system interprets the intent and context of a user's query to deliver more accurate and relevant search results [19]. Evaluating query understanding involves assessing how accurately a system interprets and responds to user queries in alignment with their intent [54]. We discuss two main sets of works:

- **End-to-end evaluation** utilizes human-judged benchmarks to assess the relevance of query-passage pairs. Among these benchmarks, QReCC [10] and TopioCQA [4] are two large-scale open-domain conversational question-answering datasets. TREC CAsT 19-22 [19, 35] and TREC iKAT 23 [5] benchmarks feature complex, knowledge-intensive conversations.
- **LLM-based relevance assessment** leverages LLMs to evaluate the relevance of retrieved information to a user's query [2, 27, 33]. However, using LLMs comes with challenges such as non-reproducibility, unpredictable outputs, and potential data leakage between fine-tuning and inference stages [37].

1.2 LLM-based Conversational Interaction

LLM-based conversational interactions improve query understanding through dynamic, back-and-forth exchanges that clarify user intent and enhance search precision. Unlike static searches, conversational LLMs capture nuances and progressively build context. We focus on two key aspects:

- **LLM-based user simulation.** Simulating diverse user behaviors, intents, and query patterns helps LLMs learn to anticipate real-world conversational scenarios, preparing them to handle complex queries and varied user needs effectively [52, 53]. This approach has become essential in evaluating systems across domains such as information-seeking dialogues [40, 43], conversational question-answering [3], and task-oriented dialogues [44].
- **Multimodal conversational interactions.** Integrating beyond-text content (e.g., images, audio) into conversations enables multimodal interactions, allowing LLMs to interpret and respond across diverse media types. This capability has expanded applications in areas like e-commerce, healthcare, and spatial analysis, enhancing tasks such as LLM-powered multimodal fashion search [24, 58], medical image retrieval [48], and beyond.

1.3 LLM-based Proactive Query Management

Conventional CIS systems passively respond to user queries. For example, current CIS systems may refuse to answer or provide low-quality answers when encountering unanswerable user queries. Here, we will introduce recent advances in developing LLM-based proactive CIS systems that can further provide useful information to unanswerable queries, or clarify the uncertainty of the query for more efficient and precise information seeking. In particular, we will cover the following:

- **Unanswerable query mitigation.** Typically, the system handles unanswerable queries passively by responding with `No Answer` [16] if there is no direct information that matches the query. This undesired result will downgrade the user experience when interacting with the CIS systems. Researchers investigate various proactive behaviors to mitigate this issue, including providing relevant information that can partially satisfy the user's information needs [55] or explanations on why the query is unanswerable [22], and suggesting other useful queries [41, 47].
- **Uncertain query clarification.** Asking clarifying questions allows the users to further clarify their queries in case the model is uncertain about their intent [8, 9, 61]. Recent studies develop various training paradigms to teach LLMs to ask clarifying questions, such as in-context learning [20], self-learning [11], reinforcement learning [15], and contrastive learning [14], as well as in multimodal scenarios [60].
- **Balancing user and system initiatives.** Taking the conversation initiative by the system introduces a great risk of harming user experience [62], while not necessarily leading to improved retrieval [28]. Therefore, it is of utmost importance to learn "when" to take the initiative in a conversation. Recent work argues that LLMs are not capable of effective planning for taking system-initiative actions [7, 45]. A solution is to predict for the system when to take the initiative in a conversation [32, 50] while simulating user–system interactions [6] is used to understand the dynamics of system initiative better.

1.4 LLM-based Query Enhancement

Query enhancement is the process of modifying a user's original query to improve retrieval performance and enhance the accuracy of search performance. By reformulating the query, systems can better match the user's intent, leading to more relevant and precise results. We cover several interactive query enhancement techniques where LLMs interpret and refine queries to capture deeper semantic nuances and better understand user intent, namely:

- **Resolving ambiguity in queries.** Query ambiguity has been investigated in many studies from various aspects such as automatic ambiguous query detection and introducing taxonomy of queries [26]. To resolve ambiguous queries, LLM-based techniques such as query expansion [25, 51], query refinement [23], and follow-up question suggestion [12] have proven effective. These approaches help clarify intent and guide users toward more precise search results.
- **Conversational query rewriting** is the process of rephrasing or modifying a user's query within a conversational context to improve retrieval accuracy and relevance [39, 49]. LLMs enhance query rewrite performance in several ways: (i) handling low-resource (few-shot or zero-shot) scenarios [31, 56, 57]; (ii) incorporating multimodal contents to improve the rewrites [59]; and (iii) generating LLM-based answer for better retrieval [1].

1.5 Open Challenges and Beyond

In the final part, we will explore key open challenges in integrating LLMs for query understanding within conversational search systems and outline research directions for future investigation.

- **Multilingual and cross-cultural query understanding.** While LLMs perform reasonably well in understanding English queries, challenges persist in handling queries across diverse languages

and cultural contexts. Expanding LLM capabilities to better support multilingual and culturally nuanced queries is essential for fostering more inclusive and accurate search experiences.
- **Real-time adaptation to evolving user intent.** Developing models that can dynamically adjust search strategies in response to evolving user intent throughout a conversation remains a significant challenge for CIS systems [46]. Under this context, instructing LLMs to accurately detect and adapt to shifts in user intent remains an important future direction.

2 Relevance to the Conference

This tutorial is highly relevant to The Web Conference as it addresses critical challenges in advancing conversational AI and IR systems — two fields that are integral to the future of web interactions. By exploring the latest techniques to improve LLM-based conversational systems, this tutorial aligns with the conference's mission to drive innovations in web technologies and enhance user experiences on the web.

Related tutorials in recent years include: (i) *Conversational Information Seeking: Theory and Application (SIGIR22)* [18]; (ii) *Proactive Conversational Agents in the Post-ChatGPT World (SIGIR23)* [30]; (iii) *Large Language Model Powered Agents in the Web (WWW24)* [21]; (iv) *Tutorial on User Simulation for Evaluating Information Access Systems on the Web (WWW24)* [13]. However, these tutorials mainly introduce applications of conversational IR and agent-based interactions. Our tutorial mainly focuses on enhancing query understanding within LLM-based conversational IR systems and beyond.

3 Detailed Schedule

This tutorial will be a **lecture-style** tutorial focusing on the latest advancements in query understanding based on LLM-powered CIS systems. The outline of this tutorial is summarized as follows:
- **Introduction** (20 min): ad-hoc search; preliminary of query understanding; adapting LLMs in query understanding.
- **Part I: conversational query understanding evaluation** (30 min): end-to-end evaluation; LLM-based relevance assessment.
- **Part II: LLM-based conversational interaction** (30 min): LLM-based user simulation; multimodal conversational interaction.
- **Part III: LLM-based proactive query management** (40 min): unanswerable query mitigation; ambiguous query clarification; balancing user and system initiatives.
- **Part IV: LLM-based query enhancement** (30 min): resolving ambiguity in queries; conversational query rewrite techniques.
- **Summary and outlook** (30 min): open challenges and beyond.

4 Target Audience and Materials

This tutorial is designed for researchers, students, and anyone interested in LLM-based conversational search, query understanding, IR, data mining, and NLP. The target audience includes **NLP and IR Researchers**: those exploring how LLMs enhance conversational query understanding and search. **Conversational AI Practitioners**: professionals developing AI-driven chatbots, virtual assistants, and support systems. **Graduate Students and Academics**: early-career researchers and students looking to apply LLMs in CIS.

We will create a website for all relevant materials: (i) **a presentation slide** covering the background, technique, and future directions discussed in the tutorial; (ii) **a video teaser** for public promotion [1]; and (iii) **annotated reference** to enable further study.

5 BIOGRAPHY OF PRESENTERS

Yifei Yuan is a Postdoctoral Research Fellow at the University of Copenhagen. She received her Ph.D. degree from the Chinese University of Hong Kong in 2023 and B.Eng. degree from Harbin Institute of Technology in 2019. Her research interests lie in Natural Language Processing (NLP) and IR, especially for conversational interactive search systems and image-text-based multimodal learning. She has published more than 15 papers on relevant topics at top conferences in NLP and Data Mining. She has been serving as a reviewer or program committee member of mainstream machine learning venues such as ICLR, ACL, SIGIR, and WWW.

Zahra Abbasiantaeb is a second-year Ph.D. student at the Information Retrieval Lab (IRLab), University of Amsterdam (UvA). She received her master's degree from Amirkabir University (Tehran), on Artificial Intelligence in 2021. Her research interests lie in IR and CIS systems. She has published several papers at top conferences including SIGIR and WSDM. She is co-organizing the interactive Knowledge Assistant Track (iKAT) at the Text REtrieval Conference (TREC), aiming to advance the development of personalized conversational search systems.

Yang Deng is an Assistant Professor at Singapore Management University. His research lies in NLP and IR, especially for conversational and interactive systems. He has published over 50 papers on relevant topics at top venues such as WWW, SIGIR, ACL, EMNLP, and ICLR, and serves as Area Chair for ACL, EMNLP, and NAACL. He has rich experience in organizing tutorials at top conferences, including WWW 2024, SIGIR 2024, and ACL 2023.

Mohammad Aliannejadi is an Assistant Professor at IRLab, University of Amsterdam. His research interests include conversational information access, recommender systems, and LLM-based data augmentation and evaluation. Mohammad has co-organized various evaluation campaigns such as TREC CAsT, TREC iKAT, CLEF Touché, ConvAI3, and IGLU, focusing on different aspects of user interaction with conversational agents. Moreover, Mohammad has held multiple tutorials and lectures on CIS, such as ECIR, SIGIR-AP, WSDM, CHIIR, SIKS, and ASIRF.

References

[1] Zahra Abbasiantaeb and Mohammad Aliannejadi. 2024. Generate then Retrieve: Conversational Response Retrieval Using LLMs as Answer and Query Generators. *arXiv preprint arXiv:2403.19302* (2024).
[2] Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. 2024. Can We Use Large Language Models to Fill Relevance Judgment Holes? arXiv:2405.05600 [cs.IR]
[3] Zahra Abbasiantaeb, Yifei Yuan, E. Kanoulas, and Mohammad Aliannejadi. 2023. Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions. *WSDM* (2023).
[4] Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *TACL* 10 (2022), 468–483.
[5] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. TREC iKAT 2023: A Test Collection for Evaluating Conversational and Interactive Knowledge Assistants. In *SIGIR*. ACM, 819–829.

[1] https://drive.google.com/file/d/1UMki52eKDXMnph3ifl9bM0lhV55wFYc9/view?usp=sharing