Rethinking the Human-Centered Evaluation of Conversational Systems

Clemencia Nyamisa Siro

Rethinking the Human-Centered Evaluation of Conversational Systems

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 8 juli 2025, te 10.00 uur

door

Clemencia Nyamisa Siro

geboren te Kisii

Promotiecommissie

Promotor:	prof. dr. Maarten de Rijke	Universiteit van Amsterdam
Co-promotor:	dr. Mohammad Alian Nejadi	Universiteit van Amsterdam
Overige leden:	prof. dr. Theo Araujo	Universiteit van Amsterdam
	prof. dr. Fabio Crestani	University of Lugano
	prof. dr. Natali Helberger	Universiteit van Amsterdam
	prof. dr. Evangelos Kanoulas	Universiteit van Amsterdam
	dr. Maria Soledad Pera	Technische Universiteit Delft
	dr. Siva Reddy	McGill University

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research was carried out at the Information Retrieval Lab of the University of Amsterdam, with support from DreamsLab.

Cover by Dennis Muthenya
Printed by Ridderprint, Amsterdam

ISBN: 978-94-6522-346-9

Acknowledgements

They say it takes a village to raise a child. It turns out that it also takes a global community to shape a PhD candidate. Embarking on a PhD is often described as a journey, and mine indeed was. It began with a move to Amsterdam during the COVID-19 pandemic, when offices were closed and most people were working from home. Starting a new life and research path in such uncertain times was not easy, but it marked the beginning of a transformative chapter, made possible by those who helped me get there.

My academic path would not have been possible without Siva, my Master's thesis advisor. Thank you for replying to a message from a stranger and agreeing to supervise my thesis. You guided me with care when I was new to research and encouraged me throughout the PhD application process. Having you on my committee now, witnessing the culmination of what you helped initiate, brings this journey full circle.

I would like to thank Maarten for the supervision. You possess the rare ability to make my wildest ideas and impossible deadlines seem entirely reasonable. You encouraged me to be independent in my research and supported my passion for community-driven work, even when it was outside the direct scope of my PhD. Your belief in me, your patience, and your willingness to push me to be the best version of myself made all the difference. Thank you also for accepting my opinionated nature and for learning to recognise my silently disagreeing face.

To my co-supervisor, Mohammad, thank you for being an incredible mentor. Thank you for your guidance and for trusting me to take the lead in my research. You were consistently present and engaged in all aspects of my PhD. I am especially thankful for your support with the various organisational roles I took on and for always being available to provide feedback, even during your holidays. More than a supervisor, you have also been a friend and gym buddy. Thank you for introducing me to Zumba, Pilates, and more recently, Strength and Conditioning, where we are both learning what it means to be strong people.

To the members of my PhD committee, Theo, Fabio, Natali, Evangelos, Maria, and Siva, thank you for taking the time to engage with my work and offering your thoughtful feedback. I appreciate your contribution to this final stage of my journey.

Thank you to Philipp and Beth for being my paranymphs and friends who have made this journey brighter.

The IRLab became my intellectual home, filled with colleagues who made this experience remarkable. The lunches, dinners, and hotpots created a sense of belonging that sustained me through the inevitable challenges of doctoral research. I would like to extend my thanks to: Alessio, Ali A, Ali V, Amin, Ana, Andrew, Antonis, Arezoo, Arian, Barrie, Catherine, Chen, Chuan, Clara, Cosimo, Dan, Daniel, David, Dylan, Gabriel, Gabrielle, Georgios, Hongyi, Ivana, Jasmin, Jia-Hong, Jiahuan, Jie, Jin, Jingfen, Jingwei, Julien, Kidist, Kiki, Lu, Maarten M, Maartje, Maria, Mariya, Maryam, Masoud, Maurits, Maxime, Ming, Mohanna, Mozhdeh, Olivier, Pablo, Panagiotis, Petra, Philipp, Pooya, Romain, Roxana, Ruben, Ruqing, Saedeh, Sam, Sami, Sebastian, Shaojie, Shashank, Sid, Siddharth, Simon, Songgaojun, Svitlana, Teng, Thilina, Thong, Vaishali, Vera, Weijia, Xinyi, Yangjun, Yibin, Yifei, Yixing, Yougang, Yongkang, Yuanna, Yuanxing, Yubao, Yuyue, Zahra, Zhirui, Zihan, and Ziyi.

My DReaMS lab colleagues, thank you for being part of this journey and for all the shared laughs and support. Thilina, you have been my PhD buddy from day one. Shashank, thank you for being a solid friend and for always having the information that Thilina and I somehow managed to miss every time.

Some connections seem written by fate. Kidist, our paths first crossed in Rwanda in 2019, during our master's, where we went our separate ways, only to converge again in Amsterdam. Thank you for your support, encouragement, and for introducing me to Ethiopian dances that I never quite mastered despite my best efforts.

I also want to thank the Masakhane community for leading the way in improving research for low-resource African languages. Being part of this community has been a truly meaningful experience. Thank you, David, for welcoming me as a new member, guiding me through my first community project, and making me feel like I was part of something bigger than myself.

I also want to thank the students I had the chance to supervise: AmirAhmad, Bram, and Liz. I learned just as much from you as I hope you did from me.

I would like to thank my friends who provided necessary distractions and helped me take my mind off research when needed: Ben, Beth, Ivona, Ivy, Helen, Jacky, Joan, Kups, Nathalie, Syrenna, Theo, Tom, and Xiaoyi. Your friendship maintained my balance and perspective.

Lastly, I want to thank my family for always cheering me on. Being a first-generation student, this path came with unique challenges, but my family never wavered in their support. To my parents, thank you for encouraging my academic pursuits even when they took me far from home. To my siblings, your weekly updates made the distance between us feel insignificant, keeping me connected to family life across continents.

Clemencia
Kisii, April 2025

Contents

1
Introduction

The landscape of information access has fundamentally transformed, driven by a shift in how people interact with information systems [25, 177]. While users once carefully constructed search queries, they now engage in natural conversations, expecting systems to adapt to human communication patterns [16, 147]. Consider a user planning a vacation, instead of crafting separate queries about flights, timing, and accommodations, they might simply begin: "I am thinking about visiting Japan." This evolution from structured queries to dialogue reflects some challenges. (1) How can we understand user expectations and interaction patterns with conversational systems? and; (2) How can we reliably evaluate their effectiveness in understanding user intent and preference, delivering satisfying interaction experiences, and assisting users in achieving their goals?

This shift toward more conversational interactions, while it reduces user effort, fundamentally challenges traditional approaches to understanding system effectiveness [118, 138]. Search evaluation frameworks rely on implicit signals such as clicks, dwell time, and mouse hovers [72, 90], but conversational interactions present more subtle indicators of user satisfaction: rephrasing requests, engaging with system suggestions, or abandoning lines of inquiry [12, 154]. Understanding these signals requires considering several factors, such as user expertise, task complexity, and dialogue state [77, 208]. Moreover, these signals vary significantly across user groups [4, 84], making evaluation particularly challenging for conversational systems.

The evaluation of conversational systems has evolved to address these challenges with three approaches. Automatic evaluation methods, leveraging computational metrics, offer scalable and reproducible assessments of response quality [156]. These range from traditional lexical similarity measures to more sophisticated neural approaches that attempt to capture semantic and pragmatic aspects of conversations [81, 170, 235]. While efficient, these metrics often fail to capture the nuanced aspects of conversational quality that humans can readily assess [77]. Interactive evaluation, through direct user engagement, provides perhaps the most authentic insights into system performance [188]. However, interactive evaluation faces several significant limitations. Live system tests can be disrupted by technical issues, while soliciting user feedback during natural interactions can be intrusive and negatively impact user experience. The timing of such assessments presents additional challenges – frequent interruptions may annoy users, while delayed feedback might miss crucial interaction details. Controlled laboratory studies attempt to address these issues through structured environments but, in doing so, sacrifice the diversity and authenticity of real-world interactions [118]. Furthermore, participants in laboratory settings often exhibit different behavioral patterns compared to natural usage contexts, where users have genuine tasks and motivations [60].

To complement interactive evaluation and overcome its scalability challenges, crowdsourced evaluation has increasingly been adopted, where trained annotators assess system responses across diverse scenarios [96]. This approach enables large-scale and cost-effective evaluation of conversational systems. However, despite these advantages, crowdsourcing presents challenges that can affect assessment quality [37, 105]. Unlike actual users, crowdworkers lack authentic task context and motivation, potentially leading to superficial or inconsistent evaluations [60, 117]. Assessors must simultaneously track dialogue context, interpret implicit user feedback, and judge response appropriateness while managing their cognitive load. As conversations grow longer, assessors may struggle to maintain attention to contextual details and subtle interaction patterns [119]. The variability in evaluator expertise, domain knowledge, and quality criteria interpretation further challenges the reliability of crowdsourced assessments [101, 115, 116, 173].

Recent advances in large language models (LLMs) have sparked interest in their potential to address the evaluation challenges for conversational systems [63, 222]. These models can efficiently analyze conversations at scale, assessing aspects such as coherence, relevance, and contextual consistency [41]. Large language models offer advantages in systematic evaluation, maintaining consistent criteria across extended dialogues where human assessors might struggle with fatigue or varying standards. While capable of mitigating some of the crowdsourcing challenges, the LLM-as-a-Judge approach also has significant limitations. LLMs exhibit self-bias, favoring their own-generated content [169, 233], and demonstrate cognitive biases, including anchoring effects, position bias, and length bias [128, 135]. Furthemore, they struggle with subjective assessments requiring an understanding of user preferences [202] and face challenges with context sensitivity and cultural nuance [210].

As the field of information retrieval advances, there has been a growing recognition of the need to move beyond text-only conversational systems. By integrating visual elements alongside the text, conversational systems offer richer ways to convey information and support user tasks [23]. A single image of a neighborhood can instantly communicate what "less touristy" means more effectively than paragraphs of text. City maps can clarify transit options, while seasonal tourism charts can inform travel timing. However, the impact of these visual elements varies significantly across users and tasks [205]. Some users may find visual information on the perceived success of a conversation clarifying, while others might find it distracting. This variability demands evaluation approaches that can assess whether visual elements are relevant and when and how they genuinely enhance the interaction.

Building on the challenges and advancements in conversational systems, this thesis focuses on the evaluation and enhancement of conversational systems. We investigate two key system actions that shape the user experience: *response generation* and *asking clarifying questions*. When users interact with a system, the system is expected to interpret their intent and provide relevant information, such as recommendations, explanations, or alternative suggestions. However, ambiguity in natural language and the complexity of maintaining context across dialogue often lead to misinterpretations. To address this, asking clarifying questions enables the system to manage uncertainty in user requests. Instead of making assumptions when faced with ambiguity or incomplete information, the system can engage users in a dialogue to refine their intent, resolve ambiguity, and improve response accuracy. This shifts the system from a passive responder to an active conversational partner, facilitating a collaborative exchange.

In this thesis, we explore the theoretical foundations, implementation strategies, and evaluation challenges associated with these system actions. We examine how dialogue aspects influence user satisfaction and develop predictive models to capture these effects. We investigate the influence of dialogue context on crowdsourced evaluations, assessing the impact of contextual information on annotation quality. Additionally, we analyze how user feedback in follow-up utterances shapes evaluation judgments, identifying differences in human and LLM-based assessments. Beyond evaluation, we introduce a framework for automating the generation and evaluation of clarifying questions. Recognizing the growing role of multimodal interactions, we further explore how integrating visual elements into clarifying questions enhances user understanding and retrieval effectiveness in conversational search.

By integrating scalable evaluation techniques with multimodal strategies, we advance conversational AI, ensuring that systems not only generate meaningful responses but also effectively manage uncertainty. These contributions support the development of adaptive, user-centered, and context-aware conversational systems.

1.1 Research Outline and Questions

The actions taken by conversational systems, as discussed above, form the foundation of this thesis. We investigate these actions through two complementary themes by answering the following research questions in this thesis:

RQ1 Which dialogue aspects influence user satisfaction in a conversational recommender system, and can we effectively predict user satisfaction using these dialogue aspects?

RQ2 What is the effect of dialogue context on crowdsourced evaluation labels in task-oriented dialogue systems?

RQ3 How does incorporating user feedback through follow-up utterances affect evaluation judgments by humans and LLMs, and what does this reveal about their respective strengths as annotators?

RQ4 How effectively can large language models generate and evaluate clarifying questions for conversational search systems?

RQ5 How do images in clarifying questions affect user performance and preferences in conversational search across different tasks and user expertise levels?

We address these research questions under the following two themes:

Theme 1: Understanding the Evaluation of Task-Based Conversational Systems. This theme examines how to evaluate systems designed to help users accomplish specific tasks through conversations. We focus on understanding effective evaluation methodologies for task-based conversational systems, referred to variously as conversational recommender system (CRS) or task-oriented dialogue system (TDS) across different chapters. Through the examination of user satisfaction, dialogue context, and user feedback patterns, we aim to align evaluation approaches with actual user needs. Within this theme, we address the following research questions:

RQ1 Which dialogue aspects influence user satisfaction in a conversational recommender system, and can we effectively predict user satisfaction using these dialogue aspects?

To understand reliable evaluation methods, we first investigate which aspects of a conversation most influence user satisfaction and how they contribute to user satisfaction. User satisfaction depicts the effectiveness of a system from the user's perspective. Understanding and predicting user satisfaction is vital for the design of user-oriented evaluation methods for conversational recommender system. Current approaches rely on turn-level satisfaction ratings to predict a user's overall satisfaction with a conversational recommender system. These methods assume that all users perceive satisfaction similarly, failing to capture the broader dialogue aspects that influence overall user satisfaction. To answer **RQ1**, we propose and investigate the effect of several dialogue aspects on user satisfaction when interacting with a conversational recommender system. To this end, we annotate dialogues based on six aspects (i.e., relevance, interestingness, understanding, task completion, interest-arousal, and efficiency) at the turn and dialogue levels. We then adopt these aspects as features for predicting response quality and user satisfaction, demonstrating the effectiveness of the proposed dialogue aspects in predicting user satisfaction and identifying dialogues where the system is failing. Our analysis reveals that different aspects affect user satisfaction at different conversation levels. Relevance influences turn-level satisfaction, while, overall satisfaction depends more on interest arousal and task completion.

Having identified the key dialogue aspects influencing user satisfaction, we now need to understand how evaluators can reliably assess them, leading to our investigation of the effect of dialogue context:

RQ2 What is the effect of dialogue context on crowdsourced evaluation labels in task-oriented dialogue systems?

Crowdsourced labels play a crucial role in evaluating task-oriented dialogue systems. Obtaining high-quality and consistent ground-truth labels from annotators presents challenges. When evaluating a task-oriented dialogue system, annotators must fully comprehend the dialogue before providing judgments. Previous studies suggest using only a portion of the dialogue context in the annotation process. However, the impact of this limitation on label quality remains unexplored. For **RQ2**, we investigate the influence of dialogue context on annotation quality, considering the truncated context for relevance and usefulness labeling. We further propose to use LLMs as an annotation assistant to summarize the dialogue context to provide a rich and short description of the dialogue context and study the impact of doing so on the annotator's performance. Our investigation shows that dialogue context significantly influences evaluation quality. Too little context leads to unreliable judgments, while too much context overwhelms evaluators. However, we also discovered that even with optimal context, assessors sometimes miss important signals about response quality that are only revealed through users' follow-up reactions.

This finding leads us to examine how user feedback influences the reliability of evaluation labels:

RQ3 How does incorporating user feedback through follow-up utterances affect evaluation judgments by humans and LLMs, and what does this reveal about their respective strengths as annotators?

In ad-hoc retrieval, evaluation relies heavily on user actions, including implicit feedback. In a conversational setting, such signals are usually unavailable due to the nature of the interactions, and, instead, the evaluation often relies on crowdsourced evaluation labels. The role of user feedback in annotators' assessment of turns in a conversational perception has been little studied. We focus on how the evaluation of task-oriented dialogue systems is affected by considering user feedback, explicit or implicit, as provided through the follow-up utterance of a turn being evaluated to answer **RQ3**. We explore and compare two methodologies for assessing task-oriented dialogue systems: one includes the user's follow-up utterance, and one does not. We use crowdworkers and LLMs as annotators to assess system responses across four aspects: relevance, usefulness, interestingness, and explanation quality.

The findings of the first theme on the effect of context and feedback interpretation reveal both the challenges and opportunities in automated evaluation. Particularly promising is the ability of LLMs to understand dialogue context and assess response quality. Building on this potential, our second theme explores practical applications: how can we use LLMs to generate effective clarifying questions, and how can visual elements enhance clarification strategies, thereby improving the effectiveness of conversational search systems?

Theme 2: Advancing Clarification in Conversational Search. The second theme of this thesis focuses on how conversational search systems can better understand user needs through clarifying questions. We examine how to use LLMs to automatically generate and evaluate these questions at scale and investigate whether adding images to clarifying questions helps users better express their information needs:

RQ4 How effectively can large language models generate and evaluate clarifying questions for conversational search systems?

Generating diverse and effective clarifying questions is crucial for improving query understanding and retrieval performance in open-domain conversational search systems. To answer **RQ4**, we propose AGENT-CQ (Automatic GENeration, and evaluaTion of Clarifying Questions), an end-to-end LLM-based framework addressing the challenges of scalability and adaptability faced by existing methods that rely on manual curation or template-based approaches. AGENT-CQ consists of two stages: a generation stage employing LLM prompting strategies to generate clarifying questions and an evaluation stage (CrowdLLM) that simulates human crowdsourcing judgments using multiple LLM instances to assess generated questions and answers based on comprehensive quality metrics. Extensive experiments on the ClariQ dataset [8] demonstrate CrowdLLM's effectiveness in evaluating question and answer quality.

While our work thus far has advanced our understanding of evaluation methodology and demonstrated the potential of automated approaches, it has primarily focused on text-based interactions. However, modern conversational systems increasingly incorporate multiple modalities to enhance user interaction. This raises important questions about how different modalities affect user behavior and system effectiveness:

RQ5 How do images in clarifying questions affect user performance and preferences in conversational search across different tasks and user expertise levels?

Conversational search systems increasingly employ clarifying questions to refine user queries and improve the search experience. Previous studies have demonstrated the usefulness of text-based clarifying questions in enhancing both retrieval performance and user experience. While images have been shown to improve retrieval performance in various contexts, their impact on user performance, when incorporated into clarifying questions, remains largely unexplored. To answer **RQ5**, we conduct a user study with 73 participants to investigate the role of images in conversational search, specifically examining their effects on two search-related tasks: (i) answering clarifying questions and (ii) query reformulation. We compare the effect of multimodal and text-only clarifying questions in both tasks within a conversational search context from various perspectives. We also conduct retrieval experiments for the two tasks with clarifying questions from the two setups to show the effectiveness of multimodal clarifying questions.

1.2 Main Contributions

This thesis makes several contributions including methodological, empirical, and resource.

Methodological Contributions

- A framework of six dialogue aspects for modeling user satisfaction in conversational recommender systems, demonstrating how different aspects influence satisfaction at turn and dialogue levels (Chapter 2)
- A systematic approach for predicting user satisfaction in conversational recommender systems using dialogue aspects (Chapter 2)
- An LLM-based method for generating concise dialogue context summaries, improving the consistency and efficiency of crowdsourced dialogue annotations (Chapter 3)
- A framework for integrating and evaluating the role of user feedback in the evaluation of task-oriented dialogue systems (Chapter 4)
- AGENT-CQ: An end-to-end framework for generating and evaluating clarifying questions in conversational search, enhancing retrieval effectiveness (Chapter 5)
- CrowdLLM: A novel evaluation system that simulates human crowdsourcing judgments using multiple LLM instances, reducing annotation cost and variance (Chapter 5)

– A methodology for comparing the effectiveness of multimodal versus text-only clarifying questions in conversational search (Chapter 6)

Empirical Contributions

– An analysis of how dialogue context affects annotation quality and reliability in evaluation tasks (Chapter 3)

– Comparative analysis of evaluation methodologies with and without user feedback using both crowdworkers and LLMs (Chapter 4)

– Analysis on the generation capabilities of different prompting strategies and LLMs (Chapter 5)

– Quantitative analysis of how visual elements in clarifying questions affect user performance in search tasks (Chapter 6)

Resource Contributions

– An annotated dialogue dataset with satisfaction ratings and aspect labels at turn and dialogue levels (Chapter 2)

– A dataset of task-oriented dialogue system evaluations across different context conditions (Chapter 3)

– A comparative dataset of annotations with and without user feedback (Chapter 4)

– A synthetic dataset of LLM generated clarifying questions with simulated user answers (Chapter 5)

– A dataset comparing user interactions with multimodal and text-only clarifying questions (Chapter 6)

1.3 Thesis Overview

This thesis consists of seven chapters, an introductory chapter, five research chapters, and a concluding chapter. In each of the next five chapters, one of the main research questions (defined in Section 1.1) is discussed. Additionally, Chapters 2–6 answer more fine-grained research questions that concern chapter-specific contributions. The current chapter, Chapter 1, introduces the research problem of evaluating and advancing conversational systems, presents the motivation and context for this work, outlines the research questions, and summarizes the main contributions and the origins of each chapter. Chapter 2 investigates what makes conversational recommender systems effective from the user's perspective. Chapters 3 and 4 address the crucial challenge of obtaining reliable evaluation labels through crowdsourcing. Chapter 5 presents AGENT-CQ, an end-to-end framework that addresses scalability challenges in generating and evaluating clarifying questions. Chapter 6 investigates how visual elements affect user interaction with clarifying questions. The thesis concludes with a final chapter that synthesizes the findings, explores broader implications, and outlines future research directions.

Each chapter in this thesis is based on a published paper, is self-contained, and can be read independently. To maintain the integrity of the original publications, we avoid creating alternative versions of the work. As a result, there is some unavoidable overlap between chapters, particularly in the background information, related work, and methodology sections. Additionally, this structure may lead to referring to or describing the same dataset differently across chapters, reflecting the context of each paper. Readers may approach the chapters sequentially or focus on specific areas of interest.

1.4 Origins

This thesis is based on the following publications:

Chapter 2 is based on the following paper:

- **C. Siro**, M. Aliannejadi, and M. de Rijke. Understanding and predicting user satisfaction with conversational recommender systems. *ACM Transactions on Information Systems, 42(2):Article 55, Sep 2023.* ACM 2023.

Author contributions:

- **CS**: Conceptualization, Methodology, Data Curation, Investigation, Formal Analysis, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing.
- **MA, MdR**: Supervision, Methodology, Validation, Writing – Review & Editing.

Chapter 3 is based on the following paper:

- **C. Siro**, M. Aliannejadi, and M. de Rijke. Context does matter: Implications for crowdsourced evaluation labels in task-oriented dialogue systems. *In Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, pages 1258–1273.* Association for Computational Linguistics, 2024.

Author contributions:

- **CS**: Conceptualization, Methodology, Data Curation, Investigation, Formal Analysis, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing.
- **MA, MdR**: Supervision, Methodology, Validation, Writing – Review & Editing.

Chapter 4 is based on the following paper:

- **C. Siro**, M. Aliannejadi, and M. de Rijke. Rethinking the evaluation of dialogue systems: Effects of user feedback on crowdworkers and LLMs. *In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, pages 1952–1962.* ACM, 2024.

Author contributions:

- **CS**: Conceptualization, Methodology, Data Curation, Investigation, Formal Analysis, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing.
- **MA, MdR**: Supervision, Methodology, Validation, Writing – Review & Editing.

Chapter 5 is based on the following paper:

– **C. Siro**, Y. Yuan, M. Aliannejadi, and M. de Rijke. AGENT-CQ: Automatic Generation and Evaluation of Clarifying Questions for Conversational Search with LLMs. *Under submission.*

Author contributions:

– **CS**: Conceptualization, Methodology, Data Curation, Investigation, Formal Analysis, Software, Validation, Visualization, Project Administration, Writing – Original Draft, Writing – Review & Editing.

– **YY**: Methodology, Investigation, Software, Validation, Visualization, Writing – Review & Editing.

– **MA, MdR**: Supervision, Methodology, Validation, Writing – Review & Editing.

Chapter 6 is based on the following paper:

– **C. Siro**, Z. Abbasiantaeb, Y. Yuan, M. Aliannejadi, and M. de Rijke. Do images clarify? A study on the effect of images on clarifying questions in conversational search. *In CHIIR '25: ACM SIGIR Conference on Human Information Interaction and Retrieval, Melbourne, Australia.* ACM, 2025.

Author contributions:

– **CS**: Conceptualization, Methodology, Data Curation, Investigation, Formal Analysis, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing.

– **ZA**: Methodology, Data Curation, Investigation, Formal Analysis, Software, Validation, Visualization, Writing – Review & Editing.

– **YY**: Methodology, Data Curation, Investigation, Validation, Writing – Review & Editing.

– **MA, MdR**: Supervision, Methodology, Validation, Writing – Review & Editing.

The writing of the thesis also benefited from work on the following publications:

– J. Wang, D. I. Adelani, ..., **C. Siro**, ..., S. T. Sari, and P. Stenetorp. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced african languages. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2024, Mexico City, Mexico, pages 5997–6023.* Association for Computational Linguistics, 2024.

– Y. Yuan, **C. Siro**, M. Aliannejadi, M. de Rijke, and W. Lam. Asking multimodal clarifying questions in mixed-initiative conversational search. *In Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, pages 1474–1485.* ACM, 2024.

– D. I. Adelani, M. Masiak, ..., **C. Siro**, ..., I. Ssenkungu, and P. Stenetorp. MasakhaNEWS: News topic classification for african languages. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational*

Linguistics, IJCNLP 2023, Nusa Dua, Bali, pages 144–159. Association for Computational Linguistics, 2023.

- O. Ogundepo, T. R. Gwadabe, ..., **C. Siro**, ..., R. N. Iro, and S. Adhiambo. Cross-lingual open-retrieval question answering for african languages. *Findings of the Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, pages 14957–14972.* Association for Computational Linguistics, 2023.
- A. Srivastava, A. Rastogi, ..., **C. Siro**, ..., Z. Wang, and Z. Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research (TMLR), 2023, 2023.* TMLR, 2023.
- **C. Siro** and T. O. Ajayi. Evaluating the robustness of machine reading comprehension Models to low resource entity renaming. *In Proceedings of the 4th Workshop on African Natural Language Processing, co-located with ICLR 2023, AfricaNLP@ICLR 2023, Kigali, Rwanda.*
- **C. Siro,** M. Aliannejadi, and M. de Rijke. Understanding user satisfaction with task-oriented dialogue systems. *In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, pages 2018–2023.* ACM, 2022.
- D. I. Adelani, J. Z. Abbott, ..., **C. Siro**, ..., T. Marengereke, and S. Osei. 2021. MasakhaNER: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics, TACL 2021, 9:1116–1131, 2021.* Association for Computational Linguistics, 2021.

2
Understanding User Satisfaction

To gauge the effectiveness of conversational systems, we need to understand how users perceive and experience interactions with these systems. Unlike traditional information systems, which rely on explicit signals such as clicks and ratings, conversational systems present a more complex challenge: different qualities of a system may have different impacts on user satisfaction and this differs significantly among users and contexts. Therefore, in this chapter, we investigate:

RQ1: Which dialogue aspects influence user satisfaction in a conversational recommender system, and can we effectively predict user satisfaction using these dialogue aspects?

Through crowdsourcing, we perform an in-depth analysis of the annotated turns and dialogues to understand how the proposed dialogue aspects influence a user's overall satisfaction. This investigation lays the foundation for the thesis by establishing what fine-grained dialogue aspects need to be evaluated in conversational systems and informing the development of evaluation methodology in subsequent chapters.

2.1 Introduction

Evaluation is a major concern when developing information retrieval (IR) systems, and it can be conducted based on measures of result relevance or user experience, such as user satisfaction, which focuses on the user's perspective. While relevance metrics such as nDCG or average precision [103] have been commonly used, are re-usable and allow for system comparison, they often demonstrate poor correlation with the user's actual interaction experience [5, 215]. As a result, in recent years, there has been a growing interest in user-oriented evaluation approaches that rely on various user interaction signals, in contrast to system-oriented evaluation methodologies, i.e., the Cranfield paradigm [51, 52].

In traditional recommender systems (RSs), user-oriented evaluation strategies often rely on implicit user feedback such as user clicks and mouse scroll events to assess whether a user finds the recommended item appealing or not. However, such interaction signals are not available for conversational recommender systems (CRSs) whose

This chapter was published as C. Siro, M. Aliannejadi, and M. de Rijke. Understanding and predicting user satisfaction with conversational recommender systems. *ACM Transactions on Information Systems, 42(2):Article 55, Sep 2023.* ACM 2023.

main interaction with users is in natural language, either by text or speech [79]. In CRSs, users interact with the system through natural language with utterances such as "I like the movie, I will watch it," expressing their preference in more detail [178]. This distinction in user interaction poses unique challenges in evaluating CRSs, both in terms of design and deployment, to ensure that these systems effectively cater to the user's needs.

User satisfaction. CRSs are recommender systems designed to provide recommendations that address the specific needs of users. As such, they fall under the category of task-oriented dialogue systems task-oriented dialogue systems (TDSs). Standard automatic evaluation metrics such as BLEU [170], ROUGE [136] and METEOR [59] have shown poor correlation with human judgment [139], thus making them unsuitable for the evaluation of TDSs. In recent years, the research community has shown significant interest in developing new automatic evaluation metrics tailored to dialogue systems. These metrics not only exhibit a stronger correlation with human judgment but also consider various aspects of dialogues, such as relevance, interestingness, and understanding, without relying solely on word overlap [80, 100, 157, 218, 244]. While these metrics perform well during system design, their efficacy during system deployment is still a subject of ongoing investigation.

As a consequence, a significant number of TDSs rely on human evaluation to measure the system's effectiveness [92, 126]. An emerging approach for evaluating TDSs is to estimate a user's overall satisfaction with the system from explicit and implicit user interaction signals [92, 126]. While this approach is valuable and effective, it does not provide insights into the specific aspects or dimensions in which the CRS is performing well. Understanding the reasons behind a user's satisfaction or dissatisfaction is crucial for the CRS to learn from errors and optimize its performance in individual aspects, thereby avoiding complete dissatisfaction during an interaction session.

Understanding user satisfaction in a task-oriented setting. Understanding user satisfaction with CRSs is crucial, mainly for two reasons. Firstly, it allows system designers to understand different user perceptions regarding satisfaction, which in turn leads to better user personalization. Secondly, it helps prevent total dialogue failure by enabling the deployment of adaptive conversational approaches, such as failure recovery or topic switching. By conducting fine-grained evaluations of CRSs, the system can learn an individual user's interaction preferences, leading to a more successful fulfillment of the user's goal.

Various metrics, including engagement, relevance, and interestingness, have been investigated to understand fine-grained user satisfaction and their correlation with overall user satisfaction in different scenarios and applications [83, 194, 218]. While recent research has seen a surge in fine-grained evaluation for dialogue systems, most of these studies have focused on open-domain dialogue systems that are non-task-oriented [74, 80, 157]. On the other hand, conventionally, TDSs such as CRSs are evaluated based on task success and overall user satisfaction. In CRSs, user satisfaction is modeled as an evaluation metric for measuring the ability of the system to achieve a pre-defined goal with high accuracy, that is to make the most relevant recommendations [182]. In contrast, for non-task-based dialogue systems (i.e., chat-bots), the evaluation focus is primarily on the user experience during interaction (i.e., how engaging or interesting the system is) [131].

Evaluating user satisfaction. Recent studies have examined user satisfaction in dialogue systems, particularly in the context of CRSs [207]. These studies typically estimate user satisfaction by collecting overall turn-level satisfaction ratings from users during system interactions or by leveraging external assessors through platforms like Amazon mechanical turk (MTurk).[1] In these evaluations, users[2] are typically asked to provide ratings for each dialogue turn by answering questions such as, *Are you/Is the user satisfied with the system response?* While overall turn-level satisfaction ratings provide a measure of user satisfaction, they may not capture the broader aspects that contribute to a user's satisfaction [200]. When humans are asked to evaluate a dialogue system, they often consider multiple aspects of the system [74]. Therefore, the satisfaction label aims to summarize the user's opinion into one single measure. Venkatesh et al. [218] argue that user satisfaction is subjective due to its reliance on the user's emotional and intellectual state. They also demonstrate that different dialogue systems exhibit varying performance when evaluated across different dialogue aspects, indicating the absence of a one-size-fits-all metric.

Previous studies have proposed metrics that offer a a granular analysis of how various aspects influence user satisfaction in chat-bot systems [83, 218]. However, it is unclear how these aspects specifically influence user satisfaction in the context of TDSs [see, e.g., 125, 245]. With most aspect-based evaluations focusing on chat-bot systems [156, 157], only a few studies have so far investigated the influence of dialogue aspects for TDSs [109, 200]. Jin et al. [109] present a model that explores the relationship between different conversational characteristics (e.g., adaptability and understanding) and the user experience in a CRS. Their findings demonstrate how conversational constructs interact with recommendation constructs to influence the overall user experience of a CRS. However, they do not specifically examine how individual aspects impact a user's satisfaction with the CRS. In [200], we proposed several dialogue aspects that could influence a user's satisfaction with TDSs. We found that, in terms of turn-level aspects, *relevance* strongly influenced a user's overall satisfaction rating (Spearman's ρ of 0.5199). Additionally, we introduced a newly defined aspect, *interest arousal*, which exhibited a high correlation with overall user satisfaction (Spearman's ρ of 0.7903). However, we did not establish a direct relationship between turn-level aspects and turn-level user satisfaction in our previous study.

Research questions. In this study, we seek to extend the study we carried out in [200]. We aim to understand a user's satisfaction with CRSs by focusing on the dialogue aspects of both the response and the entire dialogue. We intend to establish the relationship between individual dialogue aspects and overall user satisfaction to understand how they relate with satisfactory (Sat) and dissatisfactory (DSat) dialogues.

In addition, we aim to evaluate how effective the proposed aspects are in estimating a user's satisfaction at the turn and dialogue levels. To this aim, we carry out a crowdsourcing study with workers from MTurk on recommendation dialogue data, viz. the ReDial dataset [133]. The ReDial dataset provides a high-quality resource to investigate how several dialogue aspects affect a user's satisfaction during interaction with a

[1] `https://www.mturk.com`

[2] Here, *users* represent both actual users and external assessors.

CRS. We ask workers to annotate 600 dialogue turns and 200 dialogues on six dialogue aspects following [200]: *relevance*, *interestingness*, *understanding*, *task completion*, *interest arousal*, and *efficiency*. The dialogue aspects are grouped into utility and user experience (UX) dimensions of a TDS. Different from [200], we also ask workers to give their turn-level overall satisfaction rating and use it to establish a relationship between turn-level aspects and turn-level user satisfaction.

In this chapter, we answer the following chapter-level research questions:

RQ1.1 How do the proposed dialogue aspects influence overall user satisfaction with a CRS?

RQ1.2 Can we estimate user satisfaction at each turn from turn-level aspects?

RQ1.3 How effective are the dialogue-level aspects in estimating user satisfaction compared to turn-level satisfaction ratings on CRSs?

Main findings. To address our research questions, we analyze the annotated turns and dialogues in-depth to understand how the proposed dialogue aspects influence a user's overall satisfaction. We note that for most annotators, at the turn level, the ability of a CRS to make relevant recommendations has a high influence on their turn-level satisfaction rating with a Spearman's ρ of 0.6104. In contrast, at the dialogue level, arousing a user's interest in watching a novel recommendation along with completing a task are the most influential determinants for overall satisfaction ratings from annotators with a Spearman's ρ of 0.6219 and 0.5987, respectively.

To evaluate the effectiveness of the proposed dialogue aspects, we experimented with several machine learning models on user satisfaction estimation and compared their performance using the annotated data. At the turn-level user satisfaction estimation task, we achieve a Spearman's ρ of 0.7337 between a random forest regressor model's prediction and the ground truth ratings. We achieve a correlation score of 0.7956 for predicting user satisfaction at the dialogue level. These results show the efficacy of the proposed dialogue aspects in estimating user satisfaction. Additionally, these results also demonstrate the significance of assessing the performance of a CRS at the aspect level; they can help system designers to identify on what dialogue quality a CRS is not performing as expected and optimize it.

Contributions. Our contributions in this chapter can be summarized as follows.

(C1) In [200], we conducted a study on 40 dialogues and 120 responses. To gain more insights, we extend that study with an extra 160 dialogues and 480 responses. In total, we conducted our investigations on 200 dialogues and 600 responses.

(C2) We ask annotators to assess dialogues on six dialogue aspects and overall user satisfaction. In addition, they provide judgments on turn-level satisfaction. User satisfaction ratings at the turn level allow us to establish the relationship between turn-level aspects and not only overall dialogue satisfaction but also turn-level satisfaction.

(C3) We carry out an in-depth feature analysis on individual dialogue aspects and at the class level (i.e., Sat and DSat classes) to understand which dialogue aspects correlate highly with each of the classes.

(C4) Leveraging the annotated data, we experiment with several classical machine learning models and compare their performance in estimating user satisfaction at the turn and dialogue levels.

(C5) Our findings indicate that predictive models perform better at estimating user satisfaction based on the proposed dialogue aspects than based on turn-level satisfaction ratings.

To the best of our knowledge, our work is the first attempt to establish a relationship between the proposed dialogue aspects and user satisfaction at both the turn and dialogue levels and to evaluate their effectiveness in estimating user satisfaction with CRSs.

Organization of the chapter. The rest of this chapter is organized as follows. In Section 2.2, we discuss related work. We describe the dialogue aspects investigated in this study in Section 2.3. In Section 2.4, we detail our annotation process and the instructions given to the annotators. In Section 2.5, we analyze the annotated data to answer **RQ1.1**. Section 2.6 discusses our problem formulation and predictive models used to estimate turn- and dialogue-level user satisfaction, while Section 2.7 presents the results of our experiments and answers **RQ1.2** and **RQ1.3**. We discuss our results and the limitations of this study in Section 2.8 and make our conclusions, implications, and future work in Section 2.9.

2.2 Related Work

2.2.1 Conversational recommender systems

Research on recommendation via conversational interactions with IR systems is increasingly receiving attention from both industry and academia. With multi-turn interactions, a CRS enables users to find their most relevant recommendations [77]. The CRS can interactively elicit users' current preferences from their feedback and build a more complete user model to make better recommendations. Conventional recommender systems, on the other hand, only support a single interaction mode, i.e., displaying a set of suggestions depending on users' historical activities [184]. Some older CRSs took advantage of user interface elements, such as critiquing-based systems [46], where users give input on suggestions by picking from a list of pre-defined criticisms [102]. Nonetheless, recent developments in natural language technology have led to more interest in developing a CRS based on conversational user interface (CUI), where users can converse with the recommender system [112]. Several other approaches have been explored to enhance the effectiveness of recommendations, such as knowledge graph integration [253], prompt learning [224], and topic guidance [254].

The evaluation of CRSs is based on offline experiments that try to simulate a user's behavior relying on their past interaction data. One line of research evaluates the performance of a CRS based on how well it accomplishes the user's goal by making relevant recommendations using metrics such as task success and recommendation accuracy. Another line of work focuses on dialogue generation aspects, assessing the quality of the responses using word-overlap metrics such as the ROUGE score [60]. However, as argued by Deriu et al. [60], such individual measures do not reflect the overall quality of the system. Thus, current evaluation metrics that rely heavily on the system's utility do not provide us with information about the evaluation findings in practical settings. On the other hand, research shows that empirical studies conducted using user-centric approaches can accurately assess the system's performance in actual scenarios [16]. Ideally, a system should be assessed separately on each specific dialogue-level aspect to capture its performance on individual aspects [200].

So far, little work has been done to establish the relationship between dialogue aspects and overall response and system quality [200].

2.2.2 User satisfaction

Kelly [118] defines *user satisfaction* as the fulfillment of a user's specified desire or goal. User satisfaction has gained popularity as an evaluation metric of IR systems based on implicit signals [92, 106, 123, 125, 126]. In IR, user satisfaction is usually estimated based on the user's interaction experience and goal fulfillment [118]. Factors such as system effectiveness, user effort, characteristics, and expectations influence a user's satisfaction rating in IR systems [4]. Dialogue systems are often evaluated on their overall satisfaction [60], where users give their satisfaction rating at the turn and dialogue levels [38, 207]. Though subjective, user satisfaction provides valuable insights into users' perceptions, preferences, and overall evaluation of a system's performance. Additionally, it is a widely used and accepted metric in user experience research [see, e.g., 28, 92, 126].

However, for task-based conversational systems such as CRS, which should optimize towards recommendation and user experience, overall satisfaction does not capture the broad and diverse aspects influencing a user's satisfaction [200]. Thus in this research, we seek to investigate this concept.

2.2.3 Fine-grained evaluation

Due to the poor correlation between automatic metrics such as BLEU and human judgment, accurate evaluation of dialogue systems relies on human evaluation [139]. Non-task-oriented dialogue systems are evaluated on specific aspects such as relevance and engagingness [157, 218]. However, task-oriented dialogue systems are often limited to estimating the user's overall satisfaction [125, 207]. Recent research suggests that user satisfaction is multifaceted and subjective and thus should not be reduced to a single label [218].

Several recent studies have proposed to evaluate dialogue systems at an aspect level. For example, one would measure the performance of a system in making *relevant* or *understandable* responses instead of the overall quality of the response. PARADISE [219] is one of the first popular evaluation frameworks that decoupled a dialogue system's task requirements from its behavior. With predictive factors such as task success, dialogue efficiency, and dialogue quality, a system's effectiveness can be measured without having to collect user satisfaction ratings. Walker et al. [219] propose a framework for evaluating dialogues in a multi-faceted manner. They measure several dialogue aspects and combine them to estimate user satisfaction [219]. Mehri and Eskenazi [157] developed an automatic metric for evaluating dialogue systems at a fine-grained level, including interestingness, engagingness, diversity, understanding, specificity, and inquisitiveness. In their study, Venkatesh et al. [218] investigates the performance of multiple dialogue systems involved in the Alexa competition on several dialogue aspects and shows that different systems perform well in specific dialogue aspects. Moreover, they show that no single measurement can be used to evaluate the overall performance of a system accurately. Several other studies have been carried out on human evaluation of multiple dialogue aspects [see, e.g., 67, 157, 194, 243, 246].

2.2.4 Predicting user satisfaction

Predicting user satisfaction is critical in capturing whether a user's goal has been fulfilled or not. In web search, user satisfaction is viewed as a subjective measure of a user's experience during search [118]. Different from traditional IR relevance measures, such as precision and recall, user satisfaction takes into account both task success and user interaction experience [92, 93, 125]. For search systems, rich user interaction signals, such as clicks, dwell time, and mouse scroll events, are used to predict a user's satisfaction [106, 123]. Such interaction signals cannot be collected from dialogue-based systems whose main interaction is through natural language, either in text or spoken. Research on spoken dialogue systems, such as intelligent assistants, has addressed this challenge by suggesting the use of features such as spoken implicit features, intent-sensitive query embeddings, and touch-related features, showing their effectiveness in predicting user satisfaction [92, 125]. Several other features have been suggested in line with text-based dialogue systems, including implicit dialogue features, user intent, utterance length, and user-system actions, and proven to be effective [38, 207]. Bodigutla et al. [27] demonstrates the effectiveness of traditional machine learning models in predicting user satisfaction. Using predicted turn-level ratings with implicit dialogue features, models such as gradient boosting classifiers demonstrate competitive performance [27]. In task-oriented systems, several publications predict user satisfaction from turn-level overall quality user judgment ratings [207], user intents [38, 175], and implicit features such as utterance length and sentiment analysis.

Despite the success of related work in predicting user satisfaction with task-oriented systems, there has been less focus on trying to understand which dialogue aspects affect a user's satisfaction with these systems. In [200] we established the relationship between several dialogue aspects and overall user satisfaction in a TDS, especially at the dialogue level. However, compared to related work, our work in this chapter is different in several ways:

1. Unlike in our previous study [200], where we focused on dialogue-level user satisfaction, in this chapter we establish the relationship between turn- and dialogue-level user satisfaction;
2. We show the effectiveness of the dialogue aspects in estimating user satisfaction by experimenting with several classical machine learning models; and
3. We increase the data sample size by re-annotating data from our previous work [200] with one more aspect (turn-level satisfaction) and annotating an additional 160 dialogues and 480 turns. Thus, in total, we have 200 dialogues and 600 turns annotated.

2.3 Aspects Influencing User Satisfaction

In this section, we discuss the dialogue aspects we use in our crowdsourcing study. We map the qualities from prior work [109, 157, 200, 218], highlighting their definitions in different settings and defining them in our work. These qualities are derived from two TDS dimensions defined in [200]; the *utility* and *user experience* dimensions.

2.3.1 Utility

The *utility* dimension focuses on the objective nature of a CRS, that is to make relevant recommendations and accomplish a user's goal. In this dimension, we investigate two qualities, namely, *relevance* measured at the turn level and *task completion* measured at the dialogue level.

Relevance. *Relevance* is a central concept in the field of IR and plays an important role in the evaluation of conversational systems [191]. In essence, *relevance* is logically defined in the relationship between the information at hand and the user's information need [56]. In the field of conversational agents, it is used as a criterion for assessing the effectiveness of a dialogue system to potentially convey a piece of information that meets the user's needs. Ideally, relevance judgment labels should be collected from actual users to reflect their opinions (i.e., whether the suggested responses meet their information needs or not). However, it is hard to collect relevance judgments from actual users during an interaction, especially for conversational systems. This approach can be intrusive and may negatively impact the user's overall interaction experience with the system. In recent work, crowdsourcing has emerged as a reliable platform for collecting relevance labels for web search and conversational systems [13].

In our work, we employ crowdsourcing to collect relevance labels for dialogue responses. To assess the relevance of a response, we instruct annotators to rely solely on the user's explicit feedback provided in the current user's utterance. For instance, expressions such as "I don't think that is a horror movie," "I like it," "I have seen that one," "Could you recommend more like that one?" following a system's recommendation indicate whether the items suggested are relevant to the user's needs. In contrast to web search, where assessors judge the relevance of a query-document pair, relevance assessment for dialogue systems focuses on the appropriateness of the response [156]. In this study, we primarily evaluate the relevance of recommended movies rather than the appropriateness of the dialogue response itself. Therefore, we first ask annotators to determine if a movie is recommended in the response or not. If a response does not include a movie recommendation, we skip the relevance assessment. However, if a movie is recommended, we ask the annotators to determine a three-level relevance label (see Section 2.4 for more details). We adopted this definition because of the nature of our study, which is task-oriented, where our focus is on the utility of the system. Hence, relevance indicates how well the recommendations provided by the system align with the user's needs and preferences in the given conversational context. Assessing relevance at the turn level allows us to evaluate the immediate impact of a recommendation on the ongoing conversation and its ability to address the user's current needs.

Task completion. Task completion is a crucial aspect of task-oriented conversational recommendation systems (CRS), as they are designed with a predefined goal in mind. Traditionally, the main evaluation metric for task-oriented systems has been Task Success (TS), which measures the system's ability to fulfill a user's goal [229]. However, in the case of interactive CRS, TS alone may not capture the overall satisfaction of the user with the dialogue. This is due to the interactive nature of the system and the fact that task success can vary depending on individual users and task complexity [141]. Simply relying on system logs to infer user search success is inadequate because task complexity and individual user needs cannot be accurately depicted in the logs.

To address this limitation, recent research has proposed using additional interaction cues, such as self-reported user task success or expert-annotator labels on task success [60]. In our work, we investigate how the system's ability to accomplish a user's goal influences the overall impression of the dialogue for the user with a CRS. We assess the system's capability to understand the user's intent and provide recommendations that satisfy their needs. To measure the quality of task completion, we rely on the user's acknowledgment within the conversation. Utterances such as "I like it, I will watch it tonight" and "I think I will add that to my watching list" serve as signals indicating the successful accomplishment of the task from the user's perspective. By considering these explicit expressions of satisfaction or intent to engage with the recommended items, we can assess how effectively the CRS understands and addresses the user's needs. By incorporating task completion as an evaluation metric, we aim to capture the system's ability to achieve the user's desired outcome and provide recommendations that align with their preferences. This approach allows us to evaluate the CRS beyond the traditional notion of task success and consider the overall dialogue satisfaction from the user's point of view.

2.3.2 User experience

We assess how different dialogue aspects of a CRS during interaction could affect a user's satisfaction. The ideal requirement would be a system that interacts naturally with the user, making the interaction experience pleasing. Thus, inspired by related work [157, 200, 218, 246], we investigate the *interestingness*, *understanding*, *interest arousal*, and *efficiency* aspects, detailed below:

Interestingness. Due to recent advances in machine learning and natural language understanding, conversational agents such as Alexa and Siri have become increasingly common. While these agents are classified as task-oriented, there is an emerging interest in building dialogue systems that can socially engage with users while accomplishing a task [200, 206]. This quality has been used as a metric for evaluating non-goal-oriented dialogue systems in recent work [157, 218, 246]. Several proxies have been suggested for measuring interestingness, such as the number of dialogue turns and the total duration of a conversation [108, 218]. Though useful, these proxies assume the dialogue is non-goal-oriented. For goal-oriented systems, a dialogue is often supposed to be as short as possible so that the user's needs can be met quickly. Therefore, conversation length is not an accurate proxy for measuring interestingness in task-oriented systems. In our work, *interestingness* is the ability of the system to chit-chat while making relevant recommendations, that is, a system making a recommendation naturally as found in casual human conversations. It reflects the system's ability to suggest items that pique the user's curiosity or meet their interests naturally, thus enhancing their overall conversational experience. By annotating interestingness at the turn level, we aim to assess the immediate impact of a recommendation on the user's level of interest and engagement.

Understanding. The aspect of "understanding" has been investigated at both the system response and dialogue level. A system's response is said to be understandable if it makes sense in the provided context history [157]. For instance, a system is not supposed to make an utterance about racing car movies when the context is on religion (such a response will be rated as not understandable). At the dialogue level, a system is said to be understanding if it can track the user's preference and intent along the whole dialogue [200]. An understanding system is expected to conform its dialogue style to the user's preference to make sensible utterances. We show that for a dialogue system to meet a user's needs, it should be able to understand the user's preference and intent of interaction, thus, this quality is crucial in a CRS's ability to accomplish a user's task.

Interest arousal. We introduced interest arousal in [200], as an aspect highly correlated with overall user impression at the dialogue level. The ability of a task-oriented dialogue system to arouse a user's interest is significant enough to determine satisfactory dialogues [200]. This quality can be seen as a merge of two qualities: *novelty* and *explainability*. To measure the two together, we define interest arousal as "the ability of the system to suggest novel items to the user and give a brief explanation in the form of synopsis or main actors to attract the user's interest to accept the item." We rely mostly on the user's immediate utterance to capture this quality. User utterances such as, "I do not know that movie" or "Who's the main actor?" indicate that the suggested movie is not known by the user and the CRS's next action should be to give a brief explanation. Note that we do not measure this quality at the response level because annotators require at least two turns to determine user interest arousal as it is measured after a novel suggestion has been made. In this work, we are interested in quantifying the relationship between interest arousal and user satisfaction.

Efficiency. Task-oriented systems are expected to be efficient, i.e., accomplish a specified task within a minimal number of turns of interactions. In web search, a system's efficiency is measured by considering how many comparisons a user has to make before getting the needed results (number of documents examined before getting the relevant one). Various interaction signals are used to measure this aspect, including conversation length, conversation duration for spoken dialogue systems, and search session length in web search systems. Since ReDial is a text-based dataset, we use conversation length to measure a system's efficiency, that is, the ability of the system to make suggestions that meet the user's needs within minimal turns. From our analysis, we note that in most conversations, a user acknowledges a recommendation within the first three turns, and thus we conform to our previously proposed definition [200].

2.4 Data Annotation

To establish how the dialogue aspects in Section 2.3 affect a user's overall satisfaction, we create an additional annotation layer for the ReDial [133] dataset. We set up an annotation experiment on MTurk using the so-called master workers to assess:

1. Three randomly selected responses from each dialogue on two aspects, namely, *relevance* and *interestingness*;
2. The quality of the system at the dialogue level on the following aspects: *understanding*, *task completion*, *interest arousal*, and *efficiency*; and
3. The overall satisfaction of the system response and the entire dialogue.

The complete instructions and definitions given to the assessors are provided in Table 2.A.1(see the appendix). We display all three turns on a single page and instruct the annotators to answer questions for each turn as shown in Figure 2.1. After completing the turn-level annotation, the same annotators are taken to a new page where they provide dialogue-level annotations on the same dialogue (see Figure 2.A.1 in the appendix). We do not allow the annotators to return to the turn-level annotation page. This restriction is based on two considerations: (i) to avoid bias of annotators on the turn-level labels when making decisions on the dialogue-level annotations; and (ii) to prevent annotators from going back to change their turn-level ratings. With this, we aim to capture how well an annotator's turn-level ratings correlate with their dialogue-level ratings and the overall satisfaction ratings.

2.4.1 Recommendation dialogue dataset

The ReDial dataset [133] is a conversational movie recommendation dataset. It consists of $11,348$ dialogues, and the dataset is collected using crowdworkers, i.e., one person acts as the movie seeker, while the other is the recommender. The dialogues are both system and user-initiated. The movie seeker should explain their movie preferences based on the genre, actor, and movie title and ask for suggestions. The recommender's role is to understand the seeker's movie taste and intent and make the right suggestions to the user. Due to this back-and-forth process of eliciting a user's preference, which mostly involves chit-chat, this dataset is categorized as both chit-chat and goal-oriented, thus allowing us to investigate dialogue aspects from both the utility and UX dimensions of a CRS.

2.4.2 Turn-level annotation

Unlike previous work [156, 157, 207], the annotators in our study have access to the user's current utterance. We treat the response quality annotation as a turn-level task. Considering the interactive nature of a CRS, a turn is defined as a single exchange between the user and the system [207]. A turn, in this case, consists of two exchanges between the user and the system. Therefore, we define a turn in this work as:

$$T_i = S_{i-1}U_{i-1}, S_iU_i,$$

where U is the user utterance, S the system utterance and i is the current response position. In a recent study [207], turn-level annotation is conducted with workers having access to all previous system and user utterances up to the current system utterance as context and their role is to assess if the user would be satisfied with the current system response given the context without viewing the user utterance at position i. This approach requires annotators to understand the user's intent during the interaction and make judgments based on previous interactions. We argue that a user has a dynamic preference and intent during dialogue interactions, and this can change from turn to turn, thus affecting their overall satisfaction of the system. To remedy this, we ask annotators to rely exclusively on the user's current utterance while making judgments on the dialogue aspects. That is, for each system response S_i to be annotated, the annotator has access to the previous user (U_{i-1}) and system (S_{i-1}) utterances as context and the current user utterance U_i from which they should make their judgment. In this way, we aim to limit annotators' bias, in that instead of annotators making judgments influenced solely by their own opinions, they reflect the opinions of the actual user as closely as possible.

Following Mehri and Eskenazi [157], we hand-selected three system responses from each conversation for turn-level annotation. To ensure three responses cover most of the dialogue, we only select dialogues with at most fifteen turns. We limit the context window to two such that each annotated response (S_i) has two previous utterances from the system (S_{i-1}) and the user (U_{i-1}) as context plus the current user utterance (U_i). This way, we ensure that an annotator does not have to keep track of a long conversation context when annotating a single response, and each response has a reasonably long context during annotation.

For each response, we ask the annotators to assess them on relevance and interestingness and, based on their ratings for the two aspects, give their turn-level overall impression (satisfaction) rating as shown in Figure 2.1. The questions the annotators answered in this subtask are:

– *Is the system response relevant?*

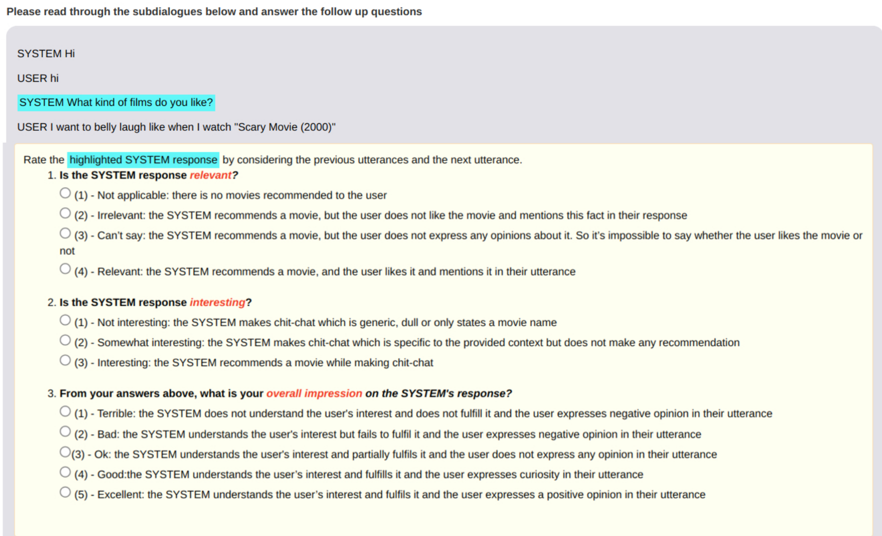
– *Is the system response interesting?*

– *Based on your ratings above what is your overall impression of the system response?*

As our annotators are not actual system users, we ask them to base their judgments solely on the next user's utterance to make the label judgment. For example, if the user states, "I don't like that movie," an annotator should be able to judge the system's response and recommendation as irrelevant since the suggested movie does not meet the user's needs. For "I have seen that and like it" the response should be rated as relevant. For the overall impression rating, we ask the annotators to base their judgment on the relevance and interestingness aspects. Each aspect comes with three options, namely, *No*, *Somewhat*, and *Yes*. For *relevance*, we also provide a *Not applicable* option when a system response does not contain a movie suggestion (e.g., if the system chit-chats or tries to elicit a user's preference). Due to limited annotation resources, we chose to focus on *relevance* and *interestingness* as the primary aspects for turn-level annotation, as they provide a strong foundation for evaluating the quality of recommendations in CRSs.

2.4.3 Dialogue-level annotation

At the dialogue level, we ask the annotators to assess the quality of the entire dialogue based on four aspects: understanding, task completion, efficiency, and interest arousal. We instructed the annotators to answer the following questions:


Please read through the subdialogues below and answer the follow up questions

SYSTEM Hi

USER hi

SYSTEM What kind of films do you like?

USER I want to belly laugh like when I watch "Scary Movie (2000)"

Rate the highlighted SYSTEM response by considering the previous utterances and the next utterance.

1. **Is the SYSTEM response *relevant*?**
   - (1) - Not applicable: there is no movies recommended to the user
   - (2) - Irrelevant: the SYSTEM recommends a movie, but the user does not like the movie and mentions this fact in their response
   - (3) - Can't say: the SYSTEM recommends a movie, but the user does not express any opinions about it. So it's impossible to say whether the user likes the movie or not
   - (4) - Relevant: the SYSTEM recommends a movie, and the user likes it and mentions it in their utterance
2. **Is the SYSTEM response *interesting*?**
   - (1) - Not interesting: the SYSTEM makes chit-chat which is generic, dull or only states a movie name
   - (2) - Somewhat interesting: the SYSTEM makes chit-chat which is specific to the provided context but does not make any recommendation
   - (3) - Interesting: the SYSTEM recommends a movie while making chit-chat
3. **From your answers above, what is your *overall impression on the SYSTEM's response?***
   - (1) - Terrible: the SYSTEM does not understand the user's interest and does not fulfill it and the user expresses negative opinion in their utterance
   - (2) - Bad: the SYSTEM understands the user's interest but fails to fulfil it and the user expresses negative opinion in their utterance
   - (3) - Ok: the SYSTEM understands the user's interest and partially fulfils it and the user does not express any opinion in their utterance
   - (4) - Good:the SYSTEM understands the user's interest and fulfills it and the user expresses curiosity in their utterance
   - (5) - Excellent: the SYSTEM understands the user's interest and fulfils it and the user expresses a positive opinion in their utterance

Figure 2.1: Turn-level annotation interface. A turn comprises two user and system utterances with three follow-up questions regarding the highlighted system utterance.

- *Is the system understanding the user's request?*
- *Did the system manage to complete the task?*
- *Is the system efficient?*
- *Does the system arouse the user's interest?*

Understanding and task completion are rated on a scale of 1–3 with the options of *No*, *Somewhat*, and *Yes*. Interest arousal is judged on a 4-point scale with a *Not Applicable* option for when no novel movie is recommended to the user or a novel movie is recommended, but the user does not follow up about it. Lastly, efficiency is assessed on a binary scale [68, 126] where the system has either made a recommendation meeting a user's request within the first three turns or not. Following [157, 200], we also ask annotators to rate the entire dialogue on *overall impression* using a 5-point Likert scale based on their turn and dialogue level aspects' ratings. Finally, we ask the workers to justify their rating on *overall impression* in a few words. We use the justifications to contextualize the given ratings and analyze and discover additional aspects that affect the quality of dialogue, as shown in Table 2.4.

2.4.4 Quality control and filtering

Here, we describe the demographics of our participants, followed by more details on the collected data and the measures we took to ensure the high quality of the data.

Participants. A total of 70 unique workers participated in the annotation. 56% male and 44% female, their age ranges from 18–40, with the majority aged between 24–35. A large number of the workers report not having experience with dialogue systems — 78% have no experience vs. 22% who do have experience. To ensure quality annotations, we filter workers based on their MTurk approval rate. We recruit workers located in the United States to ensure they are all English-proficient, with an approval rate of 95% for more than 1000 hits.

Data. The number of turns in each dialogue used in this study ranges between 12 and 13. From the analysis we carried out on the dataset, we note that most of the long dialogues with more than 20 turns tend to deviate from the movie recommendation subject into other subjects, such as politics. Each dialogue is initially annotated with at least three annotators. We always use an odd number of workers to allow for majority voting. If we lack a single agreed-upon label, an additional assessment is made with two more workers (mostly for the overall impression aspect). For the rest of the dialogue aspects, we use the labels as they are from the annotation scale to cater to the subjectivity of users in annotating the aspects. It is worth noting that we collected a set of additional annotation labels for a subset of 40 dialogues. To get to a single label for each dialogue, we treat as outliers all labels different by more than 1.5 from the mean label. In case we do not achieve a single majority label after the additional annotation, the authors re-annotate the dialogues themselves and agree with a single label.

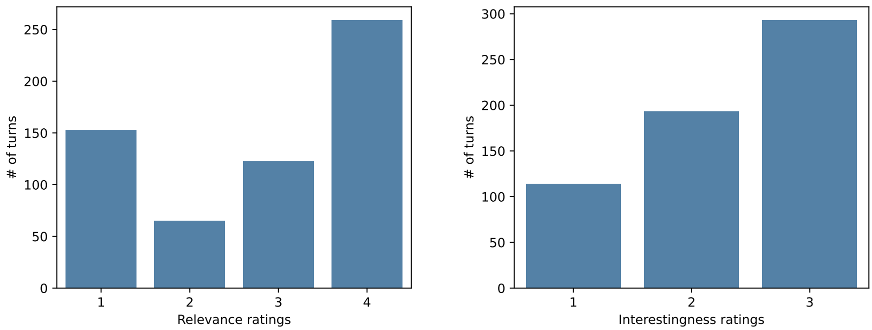
2.5 Dialogue Dataset Analysis

Using the annotated data, we first investigate **RQ1.1**: *How do the proposed dialogue aspects influence user satisfaction with a CRS?* To answer this question, we conduct several analyses to study the relationship between overall user satisfaction and both turn- and dialogue-level aspects. In addition, we identify essential aspects for the Sat and DSat classes.

2.5.1 Turn-level analysis

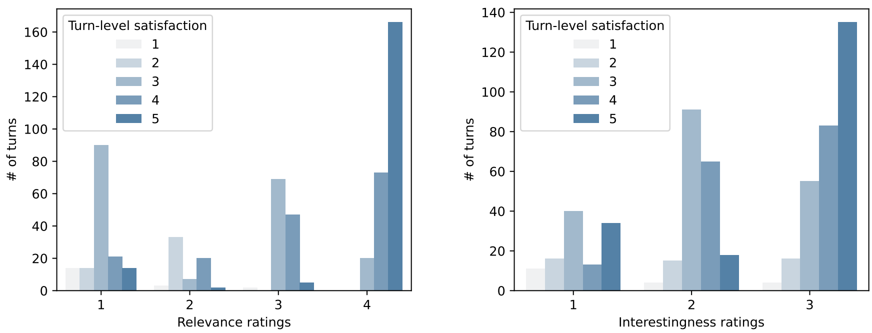
At each turn, the aspects *relevance*, *interestingness*, and *overall turn quality* are rated. We show the distribution of the ratings for these aspects in Figures 2.2a, 2.2b, and 2.3 for *relevance*, *interestingness*, and turn-level satisfaction, respectively. Note that the distributions in Figure 2.2 are computed over the three annotated turns in each dialogue. We can see that around 25% of the turns were annotated as not containing any movie recommendation ($R = 1$), while over 40% are annotated as very relevant. This result is not surprising because of the nature of the ReDial dataset, where a recommender system needs to elicit a user's preference before making a suggestion, thus having multiple chit-chat turns. Meanwhile, turning to Figure 2.3, we observe that turns rated as very relevant and interesting at the same time overall led to a satisfactory turn, showing that CRS, though goal-oriented, should not only focus on making relevant recommendations but also in a natural and interesting manner.

Figure 2.4 shows Pearson's r between turn-level user satisfaction and (i)) relevance (annotators assess if the recommended movie meets the user's preference), (ii)) interestingness of system's response. Also, we report the correlation between relevance and interestingness in the figure. We note that the relevance and interestingness aspects have a moderate positive correlation with each other ($\sim$0.4). However, we see that relevance exhibits a higher correlation with overall turn impression than interestingness. Our analysis indicates that when a turn is rated as relevant, the turn's



(a) (b)

Figure 2.2: Marginal distribution of (a) *relevance* annotations and (b) *interestingness* annotations. The values 1–3 mean not relevant/interesting, somewhat relevant/interesting, and very relevant/interesting, respectively, and with 1 for relevance meaning no movie is recommended.



(a) (b)

Figure 2.3: Distribution of (a) relevance ratings, (b) interestingness ratings against turn level satisfaction, showing how assessors rated each response based on individual dialogue aspect.

overall impression is more likely to be satisfactory (96% of the relevant turns).[3] On the other hand, the same does not hold for turns rated as irrelevant (43% of the irrelevant turns led to a satisfactory dialogue), suggesting that in this case, the user's overall impression depends not only on *relevance* but on other dialogue aspects too such as response interestingness.

In summary, we note that at the turn level, the relevance and interestingness aspects are important in understanding a user's satisfaction. Specifically, we can rely on the relevance aspect to identify Sat responses, while interestingness can be used to identify DSat responses. Characterizing the relationship between these two classes could be

[3] We use "overall impression" and "overall user satisfaction" interchangeably; both refer to overall user satisfaction.

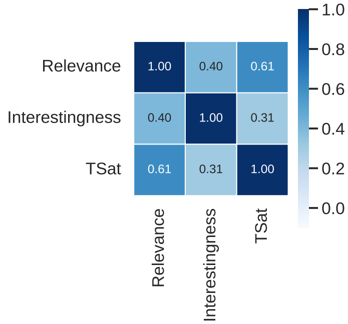


Figure 2.4: Correlation of turn-level aspects to each other and to turn-level user satisfaction.

Table 2.1: Correlation of dialogue-level overall impression with turn-level and dialogue-level aspects' ratings. All correlations in this table are statistically significant ($p < 0.01$).

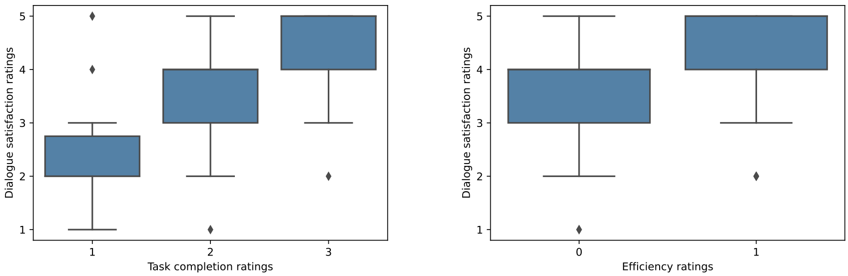
Level	Aspect	Spearman's ρ	Pearson's r
Turn	Relevance	0.3756	0.3935
	Interestingness	*0.1710*	0.2061
	Turn-level satisfaction (TSat)	**0.5397**	**0.5774**
Dialogue	Understanding	0.4929	0.5940
	Task completion	0.5987	**0.6429**
	Interest arousal	**0.6219**	0.6038
	Efficiency	0.3653	0.4004

useful in the automatic estimation of response quality.

2.5.2 Dialogue-level analysis

Table 2.1 reports Spearman's ρ and Pearson's r correlation coefficients of all six quality aspects, including *turn-level satisfaction* (TSat), with the overall dialogue satisfaction rating. Since three turns were annotated for each dialogue, we report the average results over all three turns for the three aspects. Note that both *relevance* and *turn-level satisfaction* have a moderate correlation (second row) with the overall dialogue satisfaction ratings. Compared to *interestingness*, *relevance* has a higher correlation, confirming our previous findings [200].

Notice that the *turn-level satisfaction* rating exhibits a high correlation with dialogue-level user satisfaction. This indicates that one can use a single overall turn-level quality metric to estimate a user's overall dialogue satisfaction, which has been used in previous studies [207]. We also do a correlation analysis on each turn separately and note that both *relevance* and *turn-level satisfaction* achieve a high correlation in their third and last interaction turn compared to the other two previous turns. This



(a) (b)

Figure 2.5: Box plots showing distribution of the (a) *task completion* and (b) *efficiency* aspects ratings against overall impression ratings.

shows that a system's success in making a successful suggestion[4] in the final turn has more weight on the overall impression than the preceding turns. This conforms to the findings of [126, 140, 200], showing that the latest interactions with a system have more influence on the overall satisfaction of users.

At the dialogue level, interest arousal achieves a high Spearman's ρ coefficient while task completion achieves a high Pearson's r coefficient, as shown in Table 2.1 (third row). *Efficiency* is the least correlating aspect for both scores. In our study, this aspect captures the system's ability to make relevant recommendations meeting the user's need within the first three exchanges. Unlike chatbots, which are meant to engage with a user for a long period, TDS dialogues should be concise and efficient [77].

In Figure 2.5, we plot the distribution of the ratings for the dialogue-level aspects against the overall impression. We see a clear dependency of the *overall impression* on the *task completion* aspect; out of the dialogues classified as satisfactory, 68% were rated high in terms of task completion (see Figure 2.5a). We also notice that most dialogues rated low ($= 1$) in terms of task completion are unsatisfactory overall, with a few outliers. Thus, we conclude that the ability of a CRS to complete a user's specified task can be the determinant of the overall impression.

We see in Figure 2.5b that more dialogues are rated efficient than inefficient (72.5% vs. 27.5%). We note that an efficient system, making suggestions meeting a user's need within three turns, leads to a satisfactory dialogue. Our analysis, however, indicates that the opposite cannot be said for inefficient dialogues: most of them were rated satisfactory (61.5%). We note from the annotators' open comments that even though a system took extra turns to make a relevant suggestion, as long as the user got a suggestion, they rated the system as satisfactory. This indicates that a system that fails to satisfy the user's need in the first three interactions is less likely to do so in further interactions.

To understand the significance of the investigated dialogue aspects to the *overall impression*, we train various regression models considering different aspect combinations (both single and multiple aspects) and report their R^2; see Tables 2.2 and 2.3 for the results. R^2 represents the coefficient of determination for the regression model,

[4] A successful suggestion is a movie suggestion that the user accepts.

Table 2.2: Determinant coefficients computed with regression showing the effect size of turn-level aspects to turn-level satisfaction. All results except the italicized results are significantly significant to ($p < 0.05$).

Aspect	Utility	User experience	R^2
Relevance (R)	+		0.377
Interestingness (I)		+	*0.092*
R + I	+	+	**0.382**

Table 2.3: Determinant coefficients computed with regression showing the effect size of both turn and dialogue levels aspects to overall dialogue satisfaction. All results except the italicized results are significantly significant to ($p < 0.05$).

	Aspect	Utility	User experience	R^2
Turn (T)	Relevance (R)	+		0.186
	Interestingness (I)		+	*0.036*
	Turn-level satisfaction (TSat)		+	0.290
	R + I + TSat	+	+	**0.310**
Dialogue (D)	Understanding (U)		+	0.353
	Task completion (TC)	+		0.413
	Interest arousal (IA)		+	0.365
	Efficiency (E)		+	0.160
	IA + TC + U + E	+	+	**0.559**
D + T	R + TC	+		0.452
	IA + U + I + E + TSat		+	0.572
	IA + TC + U + I + E + R + TSat	+	+	**0.607**

which indicates the proportion of the variance in turn and dialogue level satisfaction that is explained by the independent and combined aspects [40]. At the turn level, an approach that combines both aspects outperforms the best turn-level single aspect (*relevance*). As for the dialogue-level aspects, *interest arousal* exhibits the highest significance among all other aspects, taken individually. The combination of dialogue-level aspects clearly shows a stronger relationship to the overall rating model than individual aspects. Unsurprisingly, combining all aspects performs better than individual aspects or different levels.

Tables 2.1 and 2.3 show that dialogue-level aspects have a bigger influence on the *overall impression* than turn-level aspects. This suggests that turn-level aspects cannot be used solely to estimate the user's overall satisfaction effectively. This is attributed to cases where a system's response at a turn is sub-optimal, thus not representing the entire dialogue impression. The turn and dialogue aspects concern two evaluation dimensions: utility and user experience. *Relevance* and *task completion* measure the utility of a TDS, i.e., its ability to accomplish a task by making relevant suggestions. The user experience dimensions (*understanding*, *interest arousal*, *efficiency*, and *interestingness*) focus on the user's interaction experience. The combination of dialogue aspects from both dimensions has a strong relationship with the *overall impression*,

Table 2.4: Additional aspects captured from the open comments. The % shows how often the aspect was stated.

Aspect	Definition	Annotator comment
Opinion (2.4%)	System expresses general opinions on a generic topic or expressing strong personal opinion	"I don't think that the system should be providing its own opinions on the movies"
Naturalness (5.42%)	The flow of the conversation is good and fluent	"The conversation flow naturally from one exchange to the next"
Success on the last interaction (10.8%)	System gets better as time goes by	"The system finally recommends a good movie at the very end"
Repetition (1.8%)	The system repeats itself or suggestions	"The system has good suggestions, but it repeats itself over and over which is strange"
User (4.21%)	User's actions influencing the overall impression	"The system was being helpful but the user was difficult in answering preference questions"

unlike the individual aspects. In Table 2.3, the columns Utility and User experience show the two dimensions: combining both dimensions (the last row in each section in Table 2.3) leads to the best performance. The combination of turn and dialogue level aspects (D+T, third group) achieves the highest R^2. In summary, leveraging aspects from both dimensions (utility and user experience) is essential when designing a TDS that is meant to achieve a high overall impression.

Analyzing annotators' open-comments. To identify additional dialogue aspects that influence a user's satisfaction with a CRS, we conduct a manual inspection of the workers' open comments. We only report aspects based on dialogue-level user satisfaction.

We go through the comments and assign them to evaluation aspects based on the worker's perspective. For example, a comment that mentions "the system kept recommending the same movie" signals the existence of a novel aspect that concerns repeated recommendations in a dialogue. Table 2.4 lists the (dominant) novel categories discovered from the comments, together with a gloss and example. Several notable aspects are observed by the annotators. For instance, most annotators dislike the fact that the system expresses its opinion on a genre or movie. In cases where the system is repetitive (in terms of language use or recommended items), the annotators' assessments are negatively impacted. This observation is in line with [78], where they show that overexposure of an item to a user in a short time leads to a drop in user satisfaction. Some annotators note the positive impact of dialogue being natural and human-like or that the system makes a good recommendation after several failed suggestions (i.e., success on the last interaction). There are some examples where all annotators agree that the suggestions are good, but the user does not react rationally.

To summarize, in this section, we first established the relationship between several dialogue aspects with user satisfaction. We then analyzed the annotators' open comments to identify additional aspects that could influence a user's satisfaction. We conclude that at the turn-level, relevance is the most important aspect, whereas, at the dialogue level, the ability of the system to generate a user's interest and accomplish a task is significant in determining a user's overall satisfaction with a CRS. Therefore, we notice that the proposed dialogue aspects influence users' interaction with CRS differently. For some, a relevant recommendation has more effect on their overall rating, whereas others consider the ability of the system to make relevant recommendations naturally as the most important factor influencing their overall satisfaction. Thus, user satisfaction is subjective to individual users, and the design and development of CRS should cater to personalization for individual users.

2.6 Predicting User Satisfaction

In this section, we present our approach to predicting user satisfaction in CRS. We discuss the problem formulation, models used, and the evaluation metrics for both turn and dialogue level user satisfaction.

2.6.1 Turn-level satisfaction estimation

Task success [193] is a measure used in the evaluation of dialogue systems. This metric evaluates the quality of a dialogue with the assumption that users only care about their tasks being accomplished at the expense of interaction quality (IQ). Since an annotator has to accurately determine a user's intended task, the metric is not accurate enough to estimate the quality of a dialogue response. Differently, in this work, we choose turn-level satisfaction (TSat) to determine the overall quality of a response in a dialogue. TSat estimation requires each turn to be annotated at a 5-point Likert scale. Unlike Sun et al. [207], who obtain the overall response quality at each turn, our annotation scheme requires annotators to rate three randomly sampled responses from each dialogue on two dialogue aspects, namely, *relevance and interestingness*. Then, we ask them to provide their overall quality rating. Response quality estimation could be used to identify the effect of a certain response on overall user satisfaction from a user's perspective.

Problem definition. To answer **RQ1.2**: *Can we estimate user satisfaction at each turn from turn-level aspects?*, we formulate turn-level user satisfaction estimation as a regression problem. That is, given a randomly sampled turn T_i, with ratings for both the relevance (R_i) and interestingness (I_i) aspects, can we estimate a user's overall quality (O_i) rating for the given response? For example, a dialogue response rated 4 and 3 for the relevance and interestingness aspects, respectively, our task is to predict the user's overall response rating from these dialogue aspects. Using these turn-level aspects alleviates the need to manually craft features to predict turn-level satisfaction since our results show a comparative performance of simple machine learning models in estimating the quality at the response level.

Regression methods. We consider various regression models similar to [27] for predicting overall response quality rating on a continuous scale of 1–5. We experiment with five popularly used models for regression, including linear regression (LR) [230], linear support vector machine (SVM) [66], decision tree regressor (DTR) [33], random forest regressor (RFR) [32], and gradient boosting regressor (GBR) [75] which ranks features by their importance.

Evaluation criteria. For regression tasks, we use mean-squared error (MSE), root-mean-squared error (RMSE), and mean-absolute error (MAE). Following Bodigutla et al. [28], we also report Pearson's r correlation coefficient for the performance of each model's 1–5 predictions compared to the ground-truth human labels.

We implement the regression models (mentioned in Section 2.6.1) using scikit-learn [5]. We use five-fold cross-validation to tune the hyper-parameters and select the best values based on the MSE on the validation set.

2.6.2 Dialogue-level satisfaction estimation

We now investigate **RQ1.3** in this section: *How effective are dialogue aspects in estimating user satisfaction compared to turn-level satisfaction ratings?* In recent studies, dialogue-level user satisfaction for task-oriented systems has been estimated leveraging rich signals such as user intents, dialogue acts, turn-level satisfaction ratings, and implicit turn and dialogue features [125, 207]. One major limitation of estimating overall user satisfaction using turn-level satisfaction ratings is the inability to capture specific aspects influencing a user's overall impression with a dialogue system. In this work, we propose to estimate overall user satisfaction from several dialogue aspects annotated in Section 2.4. We report on a performance comparison between the two approaches and show that estimating user satisfaction from dialogue aspects leads to a better-performing model.

Problem definition. We formulate the overall user satisfaction estimation problem as a supervised binary classification task. Given the dialogue aspects' ratings, the goal is to classify the dialogue as either Sat or DSat. Due to label imbalance, we split the classes with dialogues (rating > 3) representing the satisfactory class and dissatisfactory class for dialogues (rating ≤ 3).

Classification methods. To estimate the overall quality of a dialogue system, we consider several classification models: logistic regression (Lr), a support vector machine (SVM) [66], a decision tree classifier (DTC) [33], a random forest classifier (RFC) [32], and a gradient boosting classifier (GBC) [75].

Evaluation criteria. As evaluation metrics, we adopt four commonly used metrics for binary-classification task: precision (Prec) measures the proportion of correct predicted dialogue labels to the number of predicted dialogue labels, recall (Rec) refers to the percentage of correct predicted dialogue labels to the actual number of dialogue labels, and F1-score (F1) is the harmonic mean of precision and recall. Due to the high label imbalance for the Sat class (the Sat class is three times the size of the DSat class), we do not use the accuracy metric. To understand how each model is performing, we report results for each class separately.

As with the models in Section 2.6.1, we implement the classification models with scikit-learn. For each model, we use five-fold cross-validation. To search for optimal

[5] `https://scikit-learn.org/`

hyper-parameters, we use grid-search. The best values were selected based on F1-DSat. We train our predictors based on several aspects of combination variants.

2.7 Results

In this section, we present our prediction results for both turn- and dialogue-level user satisfaction. turn-level satisfaction (TSat) is predicted with ratings from turn-level aspects (i.e., *relevance and interestingess*) whereas dialogue level user satisfaction is predicted from three types of ratings: First from the TSat ratings, second, dialogue-level aspects' ratings and finally ratings combined from both the dialogue level aspects and TSat ratings.

2.7.1 Turn-level satisfaction

Figure 2.6 shows the distribution of human-annotated response quality ratings. We note that 62% of the turns are Sat (rating > 3) compared to DSat (38%) (rating ≤ 3). Turn-level satisfaction prediction is very useful in online evaluation for identifying a problematic turn in a dialogue, thus allowing the system to adjust its recommendation or dialogue policy to avoid total dissatisfaction of the user by recovering from errors during the conversation.

At the turn level, the aim is to estimate the quality of the response from the annotated turn-level dialogue aspects, thus, we utilize graded satisfaction prediction in this task. We compare the performance of various regression models in estimating a user's response quality rating, given the relevance and interestingness ratings for the current turn, and report the results in Table 2.5. All models perform comparatively well in estimating the user rating of each response. We note that ensemble models seem to learn a good representation of the aspects and improve their predictive performance compared to single models. The performance of traditional machine learning models is a clear indication that turn-level aspects can be used to estimate the quality of response in cases where we do not have the user's turn-level satisfaction rating.

We also report the correlation coefficient between the predicted labels and the ground truth labels for each model. Among the six models we experimented with, RFR achieves the highest correlation coefficient (0.7337), followed closely by DTR at 0.7234. Our analysis of the predicted labels reveals that in most cases, the models predict accurately or close to the ground truth label for satisfactory dialogues compared to dissatisfactory dialogues. Identifying turns where the system fails is a difficult task due to label imbalance, as the majority of the turns are rated as satisfactory. It is worth noting that identifying dissatisfactory turns is more important for CRSs to adjust their interaction policy and avoid total user dissatisfaction.

In summary, extensively experimenting with the dialogue aspects as features, we conclude that both relevance and interestingness are important in predicting the quality of a response with CRS. We note that the random forest regressor achieves a high correlation coefficient of 0.7337 compared to other models. Thus, in cases where we do not have access to the user's response quality ratings, we can rely on dialogue aspects such as relevance and interestingness to estimate the quality of a response.

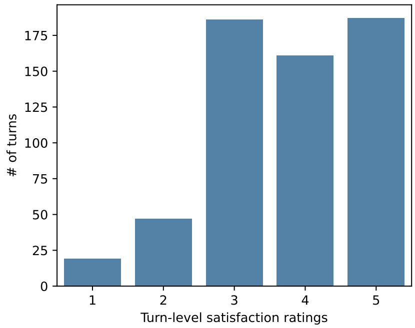


Figure 2.6: Distribution of turn-level overall quality ratings.

Table 2.5: Comparison of the performance of regression models in estimating response quality measured using MSE, and correlation between the predicted and ground truth labels. All correlations in this table are statistically significant ($p < 0.01$).

Model	MSE	Pearson's r
Linear regression	0.7762	0.6017
Support vector machine	0.8723	0.5526
Decision tree regressor	0.6089	0.7234
Random forest regressor	0.5901	0.7337
Gradient boosting regressor	0.6181	0.7197

Table 2.6: Performance of machine learning methods with a variant, predicting user satisfaction using turn-level satisfaction ratings, where the best precision (Prec), recall (Rec) and F1-score (F1) for both the satisfactory (Sat) and dissatisfactory (DSat) class are in bold. All correlations in this table are statistically significant ($p < 0.01$).

	Prec		Rec		F1		
Models	Sat	DSat	Sat	DSat	Sat	DSat	Spearman's ρ
Lr	**0.93**	**1.00**	**1.00**	**0.56**	**0.96**	**0.71**	**0.7177**
SVM	0.91	0.67	0.96	0.44	0.93	0.53	0.4823
DTC	0.92	0.42	0.86	**0.56**	0.89	0.48	0.3734
RFC	0.92	0.50	0.90	**0.56**	0.91	0.53	0.4383
GBC	0.92	0.50	0.90	**0.56**	0.91	0.53	0.4383

2.7.2 Dialogue-level satisfaction

To show how effective the proposed dialogue aspects are in predicting user satisfaction, we report the results for several classical machine learning models on user satisfaction prediction. First, we predict overall user satisfaction from turn-level satisfaction ratings (see Table 2.6). Second, we experiment with a combination of turn- and dialogue-level aspects separately (see Table 2.7). Finally, to show the effectiveness of our proposed dialogue aspects, we predict user satisfaction from all the proposed dialogue aspects (see Table 2.8).

Table 2.7: Performance comparison of machine learning methods with a variant predicting user satisfaction from turn-level dialogue aspects vs dialogue-level aspects, where the best Prec, Rec and F1 for both Sat and DSat class are in bold. All correlations in this table are statistically significant ($p < 0.01$).

	Prec		Rec		F1		
Models	Sat	DSat	Sat	DSat	Sat	DSat	Spearman's ρ
Turn-level Aspects							
SVM	0.86	**0.75**	**0.96**	0.38	0.91	0.50	0.4583
RFC	0.90	0.56	0.88	0.62	0.89	0.59	0.4789
GBC	0.94	0.67	0.91	0.75	0.92	**0.71**	**0.6286**
Dialogue-level Aspects							
SVM	0.91	0.67	**0.96**	0.44	**0.93**	0.53	0.4823
RFC	**0.96**	0.58	0.90	**0.78**	**0.93**	0.67	0.6067
GBC	**0.96**	0.58	0.90	**0.78**	**0.93**	0.67	0.6067

Table 2.6 shows the performance of several machine learning models in predicting user satisfaction from turn-level satisfaction ratings. We report the evaluation metrics for both the Sat and DSat classes, except for the correlation coefficient to capture the performance of the models in predicting a dissatisfactory dialogue. This is because identifying a problematic dialogue is of more importance for system designers to improve the model's performance for the next interaction. We note that for the Sat class, the models perform better in Prec, Rec, and F1 metrics than the DSat class. In terms of F1-DSat and Spearman's ρ, logistic regression is the best-performing model. This model classifies all the predicted satisfactory dialogues accurately as it achieves a recall score of 1.00 compared to 0.56 for dissatisfactory dialogues. Apart from having the limitation of dataset size representing dissatisfactory dialogues, it indicates that it is challenging for the model to identify dialogues where the user is dissatisfied since most of the data represents positive dialogues. Thus, understanding dialogue aspects that can easily be used to identify problematic dialogues is useful.

Additionally, we note that predicting user satisfaction from turn-level satisfaction ratings does not lead to a good performance for the DSat class. This demonstrates that user satisfaction ratings at each turn are not optimal in estimating whether a whole dialogue is satisfactory or not. We hypothesize that all turns are not equally weighted by the users when determining their overall satisfaction. Our experiments on predicting user satisfaction from individual turns reveal that the last turn is more important compared to the other turns in predicting user satisfaction. This indicates that the ability of the system to have a successful last interaction impacts a user's overall impression.

In Table 2.8, we observe an increase in the performance of F1-DSat when we predict user satisfaction from all the annotated dialogue aspects. For precision, random forest performs better in the DSat class, and decision tree in terms of recall, with random forest and SVM scoring a high F1-DSat. The predictions of the random forest model have a high correlation score with the ground truth labels, followed closely by SVM predictions. Although we do not experiment with neural architectures to allow

Table 2.8: Performance of machine learning methods with a variant predicting user satisfaction using ratings from all the proposed dialogue aspects where the best Prec, Rec and F1 for both Sat and DSat class are in bold. All correlations in this table are statistically significant ($p < 0.01$).

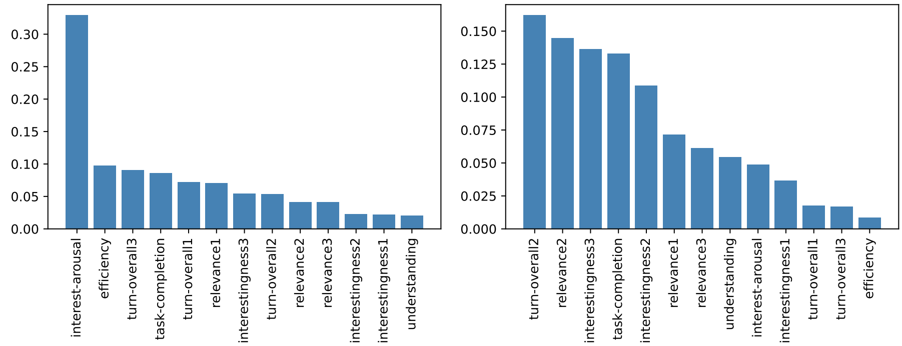
	Prec		Rec		F1		
Models	Sat	DSat	Sat	DSat	Sat	DSat	Spearman's ρ
Lr	0.93	0.83	0.98	0.56	0.95	0.67	0.6379
SVM	**0.96**	0.87	0.98	0.67	0.96	**0.80**	0.7934
DTC	0.94	0.67	0.91	**0.78**	0.92	0.71	0.6067
RFC	0.94	**1.00**	**1.00**	0.67	0.97	**0.80**	**0.7956**
GBC	**0.96**	0.79	0.93	**0.78**	0.94	0.78	0.7385

us to model the dialogue context, all models indicate a comparative performance in predicting user satisfaction from dialogue aspects with moderate correlation scores. Thus, this implies that traditional machine learning approaches can be leveraged in user satisfaction prediction, and we can rely on dialogue aspect ratings to predict user satisfaction and get comparative results without context modeling and additional implicit features.

Taking the three best-performing models from Table 2.8 (SVM, RFC, and GBC), we experiment with predicting user satisfaction using turn- and dialogue-level aspects and report the results in Table 2.7. GBC performs better in terms of F1-DSat for both the turn and dialogue levels, 0.71 and 0.67, respectively. All models perform better for precision, recall, and F1 for the Sat class. We note a superior performance when predicting user satisfaction with the dialogue level aspects compared to the turn level aspects, suggesting dialogue level aspects benefit the models more in identifying satisfactory dialogues. The DSat class seems to benefit more from the turn-level aspects when combined with turn-level satisfaction as we observe a high F1-DSat from this level. It is worth noting that, though we observe a high F1-DSat when predicting user satisfaction from the turn-level aspects, GBC and RFC from the dialogue-level aspects (see Table 2.7 row 5) achieve a high recall score for the DSat class showing their capability to accurately classify the predicted dialogues as dissatisfactory compared to the methods using turn-level features. We also report the correlation coefficients in Table 2.7 and note a comparative performance for GBC in both turn and dialogue level aspects.

Feature importance analysis. Since we experiment with several combinations of the aspects, we treat the aspects as our input features and conduct a feature importance analysis using RFC. As we report our result per class (i.e., Sat and DSat), we also report the importance of each feature based on each class, in addition to overall satisfaction prediction.

Figure 2.7 shows the significant percentage of features for (a) the Sat class and (b) the DSat class. The ability of the system to arouse a user's interest to watch an unseen movie is the most significant feature for the Sat class. We note a five percent gap between the most significant feature (*Interest arousal*- 16%) and the second most important feature (*turn-overall3* at 11%). Closely followed by *turn-overall1, task-completion* and *relevance1*. This indicates that for a CRS to improve a user's in-



(a) Satisfactory (Sat) (b) Dissatisfactory (DSat)

Figure 2.7: Bar plots showing the importance of the dialogue aspects as input features for predicting, (a) satisfactory (Sat) class and (b) dissatisfactory (DSat) class using RFC model. For turn-level aspects, we represent the performance of the three turns towards user satisfaction prediction where *relevance* $1-3$,*interestingness* $1-3$, and *turn-overall* $1-3$ are labels at turn $1-3$.

teraction experience, it should create a good impression to the user at the start and end of a conversation.

We see that a user's overall impression in turn two is the most significant feature in predicting user dissatisfaction for the entire dialogue, as shown in Figure 2.7b. Followed closely by *relevance2*, *interestingness3*, *task-completion* and *interestingness2*. Out of the top five features, we note that 3 of them are rated at the second turn, that is *turn-overall2*, *relevance2*, *interestingness2*. In most dialogues we examined, recommendations start at this turn after preference elicitation in turn one. If a system fails to capture a user's preference in the first turn, in most cases, it leads to an irrelevant recommendation being made, resulting in overall dissatisfaction. To improve the performance of the system at turn two, the system should be more understanding towards a user's request and preference. The features, *efficiency*, *turn-overall3*, *turn-overall1*, and *interestingness1* are the least significant in the prediction of the DSat class.

In general, we note that combining features from both the utility and user experience dimensions leads to a better user satisfaction measurement. In both the Sat and DSat classes, turn- and dialogue-level aspects are important. For Sat, the strong signal is interest arousal, which is measured at the dialogue level, whereas turn-level satisfaction at response two (*turn-overall2*) is the strongest in the DSat class. Evidently, we can conclude that features from both the turn and dialogue levels are important in determining satisfactory and dissatisfactory dialogues with CRSs. Therefore, based on results from Tables 2.6 and 2.8, we show that relying on only dialogue-level aspects to predict user satisfaction is as effective as using turn-level satisfaction ratings.

2.8 Discussion and Limitations

In this section, we present an analysis of our key findings and their significance on understanding and predicting user satisfaction in CRS motivated by our experimental results. Furthermore, we examine the limitations of our research, primarily based on the methodology employed throughout this study. We delve into more details below.

2.8.1 Discussion

In this chapter, we focused on understanding user satisfaction with CRSs, generally categorized as a goal-oriented dialogue system. Although goal-oriented dialogue systems are ideally expected to optimize towards task accomplishment, in this study we show that a system's behavior during interaction has an influence on their overall satisfaction during interactions at both the turn and dialogue levels. The interestingness aspect, however, does not show a high correlation with turn-level satisfaction. We hypothesize that when asked to scrutinize a CRS response explicitly on interestingness, annotators tend to rate such responses less favorably than they would if they were rating the overall experience according to the established rating process. Though this aspect is highly researched for non-task-oriented dialogue systems [28, 156, 157], from both the annotations and open-ended comments, we find that engaging with users in the form of chit-chat has both positive and negative effects on their overall satisfaction. If a user is already happy with a provided recommendation, more engagement can lead to further *interest arousal*, and hence more satisfaction; however, if the system fails to meet the user's expectations, it can have a negative effect. This is in line with [206], who stress the importance of finding the right amount of chit-chat in a goal-oriented dialogue.

Providing relevant recommendations throughout a dialogue is crucial for user satisfaction, but it does not tell the whole story. When a system makes relevant recommendations, they certainly lead to a satisfactory dialogue, but when the responses are both relevant and interesting, most users tend to rate their experience as very satisfactory for both levels. This indicates that a CRS that can make relevant recommendations alongside generating natural responses that are interesting is more likely to result in an improved user's overall interaction experience. Thus, system designers and dataset creators should consider optimizing these two aspects during the design and development of CRS systems and datasets.

Our analysis of the justifications that support a user's overall satisfaction rating reveals new aspects that can affect users' satisfaction. In line with our quantitative analysis and related work [126, 140], many annotators mention the importance of a good user experience in the final turns of a conversation. Success in the last interaction has an implication on task completion, interest arousal, and overall user satisfaction. When a system accomplishes its predefined goal, users tend to utter responses such as "Thank you for the suggestion!" "It was nice chatting with you." While utterances such as "But you did not get me something to watch" and "Such a waste of my time" indicate an inability of the system to fulfill a user's need. Therefore, in various cases, we can rely on the last user interaction to assess the system's ability to fulfill or not fulfill a user's need. It is also worth mentioning that other aspects, such as repeated utterances and recommendations, negatively impacted the user experience.

In general, we note that the UX dimension (*interestingness*, *understanding*, *interest arousal*, and *efficiency*) of a CRS plays a very important role in user satisfaction. The ability of a CRS to make relevant recommendations and accomplish a user's goal could lead to overall satisfaction, however, a system that demonstrates to be more engaging and understanding has a higher chance of satisfying users. This indicates the need to jointly optimize turn- and dialogue-level metrics and for a fine-grained model of user satisfaction that incorporates multiple aspects.

2.8.2 Limitations

In this work, we relied on external assessors to judge user satisfaction based on the user's utterances and reactions to the system's responses. While we observed a high level of agreement for most dialogues, we also noticed disagreement between annotators on some. This limitation could introduce a potential gap between the assessors' ratings and the subjective satisfaction levels of users in real-world scenarios. Additionally, interpretation biases among assessors can affect the reliability of turn and dialogue-level ratings. Therefore, it is essential to conduct this study with actual users to collect a set of fine-grained annotations from real users [145].

At the turn level, due to the substantial annotation effort required, following Mehri and Eskenazi [157], we sampled three responses from each dialogue for annotation. While this approach may not capture the full picture in a dialogue, we show that our sampling strategy provides meaningful insights into what aspects influence turn-level satisfaction. Investigating the optimal way of selecting responses to annotate from each dialogue may provide additional useful findings, but this was not our concern in this study. Therefore, we think there is a rich research gap to solve the significant annotation effort required in dialogue annotations when all the turns are annotated.

2.9 Conclusion and Future Work

In this chapter, we answered **RQ1:** *Which dialogue aspects influence user satisfaction in a conversational recommender system, and can we effectively predict user satisfaction using these dialogue aspects?* by taking a user-oriented approach to understanding user satisfaction in conversational recommendations. We conducted a study to assess the influence of multiple dialogue aspects on overall user satisfaction. Through a carefully designed annotation process, we collected external assessors' feedback ratings on six dialogue aspects (*relevance*, *interestingness*, *understanding*, *task completion*, *interest arousal*, and *efficiency*) and user satisfaction at the turn and dialogue level. With this data, we investigated the relationship between several dialogue aspects and user satisfaction. Furthermore, we adopted several machine learning methods to predict response quality and overall user satisfaction with different feature combinations.

Combining both the qualitative and quantitative methods, our results indicate that: (i) Relevant recommendations are necessary but not sufficient for high user satisfaction feedback. Therefore, several aspects should be considered in estimating a user's overall satisfaction with a CRS. (ii) In the absence of response quality ratings, we can rely on turn-level aspects to estimate the user's rating for each response. And (iii) user satisfaction can be predicted more accurately with combined dialogue aspects as features, unlike only using turn-level satisfaction ratings.

In addition to understanding how several dialogue aspects influence a user's overall satisfaction with a CRS, our findings also have implications for the design and evaluation of CRSs. Our results show that predicting user satisfaction with aspects representing the utility of a CRS (relevance and task completion) performs poorly compared to predicting with a combination of all aspects. Thus, to achieve high user satisfaction, the design of CRSs should not only be optimized towards goal accomplishment but also a good user interaction experience.

Our experimental results with traditional machine learning methods indicate a strong performance. We did not experiment with neural network architectures in this study as it is not the main focus of our work, and we leave this to future work. Furthermore, other dialogue features, such as dialogue context, intent, and system-user action, could be modeled in a neural architecture as they have proven to improve user satisfaction prediction. Since our study involves a small sample dataset, we plan to verify our findings on a larger scale and with diverse data collected from actual users interacting with the system. Collecting a large-scale dataset can be achieved automatically by leveraging existing predictive models to capture key patterns by training them with explicit ratings or in an unsupervised way. Apart from that, techniques such as user simulation can be used to provide annotated user feedback within dialogues, thus increasing the amount of data to be annotated [22], where this feedback can include explicit ratings on the dialogue aspects allowing for the collection of ground truth data for training and automatic evaluation at scale.

Though the focus of our study is to uncover the relationship between various dialogue aspects and user satisfaction, we believe our findings can provide insights into the factors that contribute to increased user satisfaction in CRS and can serve as a basis for future research and system development. We, therefore, encourage future research to investigate the practical implications of our findings by looking at the impact of increasing dialogue aspects on user satisfaction through experimental studies or user-centered evaluations using tools such as CRSLab [255] to compare different CRS methods.

For future work, we are interested in integrating large language models (LLMs) in the annotation process to further enhance the accuracy, richness, and scale of the annotated dataset. We hypothesize that their advanced contextual understanding and semantic analysis capabilities will benefit the annotations. In particular, following [70], we expect that the annotations on the recommended items will more closely align with user preferences and intents expressed in the conversation.

In this chapter, we made practical choices to manage the annotation effort by sampling three responses per dialogue and providing limited dialogue context to annotators. This raises important questions about how these methodological choices affect evaluation quality, which we investigate in Chapter 3.

Chapter Appendices

2.A Instructions for Assessors

Table 2.A.1 shows the annotation instructions given to the assessors during the human quality annotation process. Figure 2.A.1 shows a sample interface that was used for dialogue-level annotation. In Table 2.A.2, we show a dialogue example with assessors' annotations. These instructions and examples are a sample of what was shown to the assessors.

Table 2.A.1: Annotation instructions given to the crowdworkers.

<table>
<tr><td colspan="2">In this task, your goal is to rate how well an intelligent SYSTEM (like Siri or Alexa) converses with a USER. The USER is looking for some movies and the SYSTEM tries to understand what the USER likes to finally give some suggestions to the USER. You will rate the quality of the provided SYSTEM responses and the overall dialogue.</td></tr>
<tr><td colspan="2">Turn-level annotation</td></tr>
<tr><td>Relevance (1–4)</td><td>This means the response is appropriate to the previous utterance and a movie was mentioned that fulfills a user's goal, that is the user liked it, has seen it, or agreed to watch it.
1. Not applicable: there is no movie recommended to the user in the response
2. Irrelevant: the SYSTEM recommends a movie, but the user does not like the movie and mentions this fact in their response
3. Can't say: the SYSTEM recommends a movie, but the user does not express any opinions. So it's impossible to say whether the user likes the movie or not
4. Relevant: the SYSTEM recommends a movie and the user expresses a positive opinion in their utterance</td></tr>
<tr><td>Interestingness (1–3)</td><td>This means: the SYSTEM suggested a movie in the response accompanied by some small talk which would make a user want to continue interacting with the SYSTEM.
1. Not interesting: the SYSTEM makes small talk that is generic, dull, or only states a movie name
2. Somewhat interesting: the SYSTEM makes small talk that is specific to the provided context but does not make any recommendation
3. Interesting: the SYSTEM recommends a movie while making small talk</td></tr>
</table>

Continued on next page

Table 2.A.1 – continued from previous page

Turn-overall (1–5)	What is your overall impression of the system response? 1. Terrible: the SYSTEM does not understand the user's interest and does not fulfill it and the user expresses a negative opinion in their utterance 2. Bad: the SYSTEM understands the user's interest but fails to fulfill it and the user expresses a negative opinion in their utterance 3. Ok: the SYSTEM understands the user's interest and partially fulfills it and the user does not express any opinion in their utterance 4. Good: the SYSTEM understands the user's interest and fulfills it and the user expresses curiosity in their utterance 5. Excellent: the SYSTEM understands the user's interest and fulfills it and the user expresses a positive opinion in their utterance
Dialogue-level annotation	
Understanding (1–3)	This means: the SYSTEM understands the user's request and makes a recommendation meeting their interest. 1. Not understanding: the SYSTEM does not understand the user's request and makes recommendations that the user did not like 2. Somewhat understanding: means the SYSTEM understands the user's request but did not make recommendations liked by the user 3. Understanding: the SYSTEM understands the user's request and makes recommendations that the user liked
Task completion (1–3)	This means: the SYSTEM makes recommendations that either the user 'likes' or 'has seen' and agrees to watch one of the recommendations by the end of the conversation. 1. Not complete: the SYSTEM makes recommendations the USER does not like and the user ends up with no movie to watch 2. Somewhat complete: the SYSTEM makes recommendations that the USER likes but the user does not state if they will watch any of them 3. Complete: the SYSTEM makes recommendations that the USER likes and will watch

Continued on next page

Table 2.A.1 – continued from previous page

<table>
<tr><td>Interest arousal (1–4)</td><td>This means: the SYSTEM makes a novel recommendation and tries to encourage the user to like and watch it by giving more explanation about the movie.
1. Not applicable: no novel recommendation is made, that is the user does not state they don't know any of the recommended movies
2. No interest arousal: a novel recommendation is made but the SYSTEM does not try to encourage the user to accept the movie
3. Somewhat interest arousal: a novel recommendation is made, and the system tries to encourage the user to accept the movie but the user does not like or state if they will watch it
4. Full Interest arousal: a novel recommendation is made and the system tries to encourage the user to accept it and the user agrees to watch it</td></tr>
<tr><td>Efficiency (0–1)</td><td>This means: the SYSTEM makes recommendations that meet the user's interest within the first three turns.
1. Not efficient: the SYSTEM takes more than three turns to make a recommendation that the user likes
2. Efficient: the SYSTEM makes recommendations that the user likes within the first three turns.
A turn: in this case are two consecutive utterances from the SYSTEM and the USER.</td></tr>
<tr><td colspan="2">Dialogue-level annotation</td></tr>
<tr><td>Dialogue-overall</td><td>What is your overall impression of the system?
1. Terrible: the SYSTEM does not understand the user and meet their interest and the user does not like the recommendations.
2. Bad: the SYSTEM understands the user but fails to meet their interest and the user does not like the recommendations
3. Ok: the SYSTEM understands the user and partially meets their interest but the user does not express whether they like or dislike the recommendations
4. Good: the SYSTEM understands the user and meets their interest but takes extra turns to do so and the user likes the recommendations
5. Excellent: the SYSTEM understands the user and meets their interest efficiently and the user likes the recommendations</td></tr>
</table>

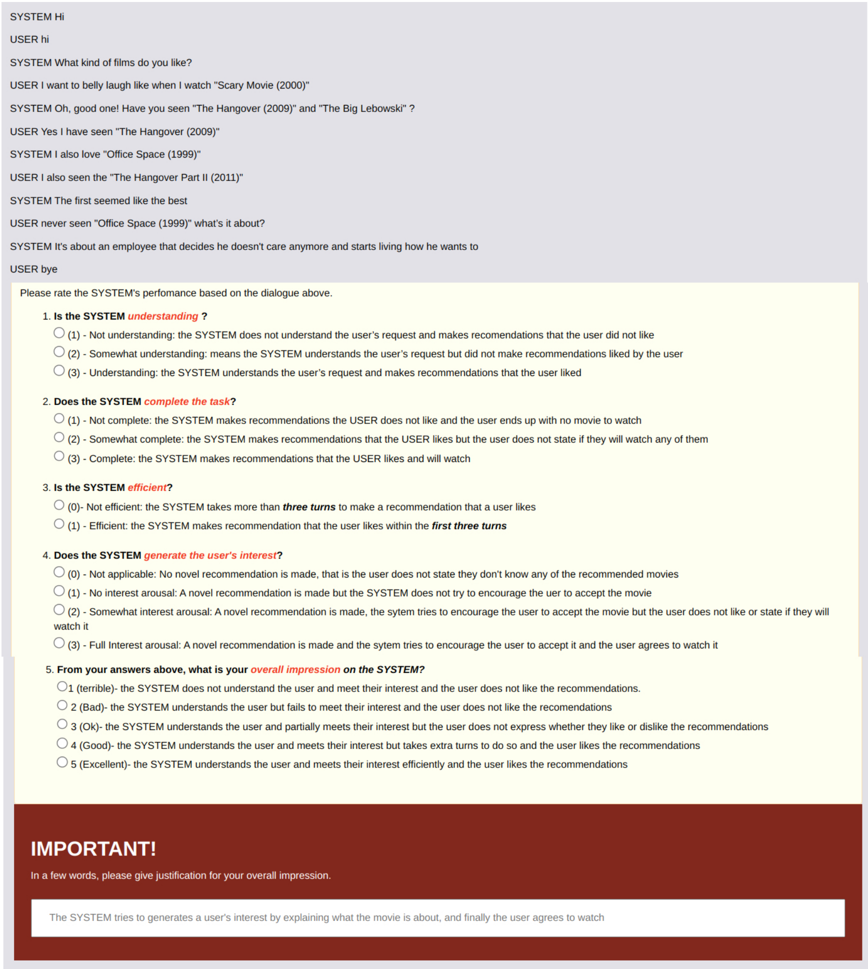


Figure 2.A.1: Interface used by annotators to annotate dialogues at the dialogue level showing the annotation guidelines for each of the four dialogue aspects, overall user satisfaction, and annotators open comment box

Table 2.A.2: Sample turns annotated in the task with worker ratings for relevance, interestingness, and turn-overall impression.

Turns	Utterances
1	SYSTEM hello how can I help you this morning USER hello I'm looking for some newer movies no preference SYSTEM Okay how about "Thor: Ragnarok (2017)"? Or "Justice League (2017)"? USER I haven't seen either one of those
	relevance: [3, 3, 4]; interestingness: [1, 1, 3]; TSat: [3, 3, 4]
2	SYSTEM They are both great films USER they sound great I will try those out on our movie night" SYSTEM you should check them out on DVD's And how about "Split (2017)"? USER I haven't seen that one either
	relevance: [3, 3, 4]; interestingness: [3, 3, 3]; TSat: [3, 4, 5]
3	SYSTEM good film! USER sounds good SYSTEM "REC (2007)"? Its about a zombie apocalypse you should check it out! USER sound really good I will check them out! thanks for the recommendations
	relevance: [4, 4, 4]; interestingness: [3, 3, 3]; TSat: [5, 5, 5]

3

Effect of Dialogue Context on Crowdsourced Evaluation Labels

In Chapter 2, we identified key aspects influencing user satisfaction through dialogue annotation. To manage the assessors' workload, we made practical choices to randomly sample three turns per dialogue and provide evaluators with a window of three prior turns as context. Although we assumed this context window would suffice, it raises questions about our evaluation methodology: How does the amount of dialogue history available affect evaluation quality? Could providing more context improve reliability, or would it overwhelm annotators? In this chapter, we explore these questions in depth, examining the impact of varying context lengths on evaluation outcomes to answer the following research question:

RQ2: What is the effect of dialogue context on crowdsourced evaluation labels in task-oriented dialogue systems?

We examine different approaches to context presentation and their impact on evaluation quality and introduce an LLM-based summarization method to improve consistency while managing annotation efforts.

3.1 Introduction

With recent advances in pre-trained language models and large language models (LLMs), task-oriented dialogue systems (TDSs) have redefined how people seek information, presenting a more natural approach for users to engage with information sources [36, 231]. As TDSs become increasingly integral to information-seeking processes, the question of how to accurately and effectively evaluate their performance becomes critical. Due to the poor correlation of automatic metrics with human-generated labels [60], evaluation of TDSs has shifted towards relying on user ratings or crowdsourced labels as ground-truth measures [131].

Various crowdsourcing techniques have been employed to collect ground-truth labels, such as sequential labeling [207], where the annotators go through each utterance

This chapter was published as C. Siro, M. Aliannejadi, and M. de Rijke. Context does matter: Implications for crowdsourced evaluation labels in task-oriented dialogue systems. *In Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, pages 1258–1273*. Association for Computational Linguistics, 2024.

and annotate them one by one. This approach introduces certain risks in the annotation process, such as annotators' fatigue and high cognitive load in extra-long dialogues, requiring them to remember and track the state of the dialogue as they annotate the utterances [200]. While following and understanding the dialogue context is crucial and can influence the annotators' ratings, reading and understanding very long dialogues can lead to degraded performance.

To address this issue, another line of research proposes to randomly sample only a few utterances in each dialogue to be annotated [157, 200, 201]. While addressing the high cognitive load and fatigue, limiting annotators' understanding of the dialogue poses obvious risks, such as unreliable and biased labels [192, 200]. In particular, the amount of dialogue context can lead to biases. For example, annotators who lack rich context may unintentionally lean towards positive or negative ratings, neglecting the broader quality of the response. Thus, offering annotators too little context risks misleading judgments, potentially leading to inaccurate or inconsistent labels. Conversely, flooding annotators with excessive information can overwhelm them, which can lead to lower returns in terms of label quality.

Prior work has investigated factors that affect the quality and consistency of crowdsourced evaluation labels, including annotator characteristics, task design, cognitive load, and evaluation protocols [see, e.g., 172, 186, 187, 190]. However, no previous work studies the effect of random sampling and the number of sampled utterances on the annotation quality.

In this chapter, we aim to address this research gap by investigating how different amounts of contextual information impact the quality and consistency of crowdsourced labels for TDSs, contributing to the understanding of the impact of such design choices. We experiment with crowdsourcing labels for two major evaluation aspects, namely, *relevance* and *usefulness,* under different conditions, where we compare the annotation quality under different dialogue context truncation strategies.

Addressing the challenge of insufficient context at the turn level, we propose to use heuristic methods and LLMs to generate the user's information need and dialogue summary. LLMs can play the role of annotation assistants [70] by summarizing the dialogue history, facilitating a more efficient and effective understanding of the dialogue context before annotating an utterance. To this aim, we use GPT-4 for dialogue context summarization and compare the performance of annotators under different conditions, as well as different context sizes. Through these experiments in this chapter, we answer two chapter-level research questions:

RQ2.1 How does varying the amount of dialogue context affect the crowdsourced evaluation of TDSs?

RQ2.2 Can the consistency of crowdsourced labels be improved with automatically generated supplementary context?

Our findings reveal that the availability of previous dialogue context significantly influences annotators' ratings, with a noticeable impact on their quality. Without prior context, annotators tend to assign more positive ratings to system responses, possibly due to insufficient evidence for penalization, introducing a positivity bias. In contrast, presenting the entire dialogue context yields higher relevance ratings. As for usefulness, presenting the entire dialogue context introduces ambiguity and slightly lowers annotator agreement. This highlights the delicate balance in contextual information provided for evaluations. The inclusion of automatically generated dialogue context enhances annotator agreement in the no-context (C_0) condition while reducing annotation time compared to the full-context (C_7) condition, presenting an ideal balance between annotator effort and performance.

Our findings extend to other task-oriented conversational tasks like conversational search and preference elicitation, both relying on crowdsourced experiments to assess system performance.

3.2 Related Work

Several user-centric dialogue evaluation metrics [80, 100, 157] have been proposed. For TDSs, high-level dimensions such as user satisfaction [5, 126] and fine-grained metrics such as relevance and interestingness [200] have gained interest. Due to the ineffectiveness of standard evaluation metrics such as ROUGE [136] and BLEU [170], which show poor correlation with human judgments [60], a significant amount of research on these metrics relies on crowdsourcing dialogue evaluation labels to improve correlation with actual user ratings. Crowdsourcing ground-truth labels has gained momentum in information retrieval (IR) for tasks like search relevance evaluation [14] and measuring user satisfaction in TDS. A major challenge is ensuring the quality and consistency of crowdsourced labels. Task design and annotators' behavioral features and demographics can affect the quality of the collected labels [101, 116, 173]. Kazai et al. [117] examine how effort and incentive influence the quality of labels provided by assessors when making relevance judgments. Other factors such as judgment scale [161, 187], annotator background [115, 186], and annotators' demographics [62] have also been studied. Most studies focus on search systems, not dialogue systems. Closer to our work, Santhanam et al. [190] study the effect of cognitive bias in the evaluation of dialogue systems. Providing an anchor to annotators introduces anchoring bias, where annotators' ratings are close to the anchor's numerical value. Like Santhanam et al. [190], we focus on the effect of task design on the evaluation of TDSs. In particular, we investigate how the amount and type of dialogue context provided to annotators affect the quality and consistency of evaluation labels and the annotator experience during the evaluation task.

3.3 Methodology

We examine how contextual information about a dialogue affects the consistency of crowdsourced judgments regarding *relevance* and *usefulness* of a dialogue response. Here, contextual information refers to the information or conversation that precedes a specific response. We carry out experiments in two phases. **Phase 1** involves varying the *amount* of dialogue context for annotators to answer **RQ2.1**. In **Phase 2**, we vary the *type* of previous contextual information available to annotators to address **RQ2.2**.

3.3.1 Experimental data and tasks

We use the recommendation dialogue (ReDial) dataset [133], a conversational movie recommendation dataset, comprising of over 11K dialogues. The dataset is collected using a human-human approach, i.e., one person acts as the movie seeker, while the other is the recommender with the goal of recommending a suitable movie to the seeker, thus making the dataset goal-oriented. We randomly select system responses from 40 dialogues for the assignment of relevance and usefulness labels. These dialogues typically consist of 10 to 11 utterances each, with an average utterance length of 14 words. We evaluate the same system responses across all experimental conditions.

The annotation task for the annotators involves two dimensions: (i) *relevance*: Is the system response relevant to the user's request, considering the context of the dialogue? And (ii) *usefulness*: How useful is the system's response given the user's information need? For the *relevance task*, we ask annotators to judge how relevant the system's recommendations are to the user's request [14]. First, the annotator has to judge whether the system response includes a movie recommendation or not; if yes, the annotator assesses whether the movie meets the user's preference; if not, we ask them to note that the utterance does not recommend a movie. The judgment is on a binary scale for the latter case, where the movie is either relevant (1) or not (0). For each experimental condition (see below), annotators only assess the system response with access to the previous context. Note that we forego the user's feedback on the evaluated response (next user utterance) to focus on the topical relevance of the recommended movie, that is, if the movie meets the user request and preference in terms of the genre, actor, director, etc. For the *usefulness task*, annotators assess a response with or without a movie recommendation to determine how useful the system's response is to the user [150]. The judgment is done on a three-point scale (i.e., very, somewhat, and not useful). Unlike the relevance task, annotators have access to the user's next utterance for the usefulness task; usefulness is personalized to the user, in that even though a movie may be in the same genre, sometimes a user may not like it (e.g., does not like the main actor), thus making the system response relevant but not useful to the user.

3.3.2 Automatic generation of diverse dialogue contexts

User information need. The user's information need plays a significant role when assessing or improving the quality of the data collected in IR systems [150]. It refers to *the specific requirement or query made by a user, which guides the system in understanding their preferences and retrieving relevant information to fulfill that need*. For TDSs, understanding the user's intent is crucial for annotators participating in the evaluation, as they are not the actual end users. This understanding improves the alignment of evaluation labels with the actual user's requirements. We define the user's information need as their movie recommendation preference. Given the consistency of user preferences in the ReDial dataset, where users tend to maintain a single preference throughout a conversation, providing the user's initial information need aids annotators in evaluating the current turn for relevance or usefulness.

We adopt two approaches to generate the user's information need. One is to heuristically extract the first user utterance that either requests a movie recommendation or expresses a movie preference, based on phrases such as "looking for," "recommend me," and "prefer." These phrases are extracted from the first three user utterances in a dialogue, with the top 10 most common phrases selected. The second approach relies on LLMs to generate the user's information need. We hypothesize that LLMs can identify pertinent user utterances in a dialogue and generate the corresponding information need. We use GPT-4 [165] in a zero-shot setting; with the dialogue context up to the current turn as input, we prompt the model to generate the user's information need.

Generating dialogue summaries. Dialogue summarization is beneficial for providing a quick context to new participants of a conversation and helping people understand the main ideas or search for key content after the conversation, which can increase efficiency and productivity [71]. We use dialogue summaries to provide annotators with quick prior context of a dialogue. We use GPT-4 [165] in a zero-shot setting, as in the case of user information needs, but we vary the prompt. We instruct GPT-4 to generate a summary that is both concise and informative, constituting less than half the length of the input dialogue. Both the generated user information needs and summaries are incorporated in Phase 2 of the crowdsourcing experiments.

Due to LLM 's potential for hallucination [30, 43], we evaluate the generated summaries and user information need to ensure factuality and coherence. We elaborate on the steps we took in Section 3.B.

3.3.3 Crowdsource experiments

Following [113, 117, 186], we design human intelligence task (HIT) templates to collect relevance and usefulness labels. We deploy the HITs in variable conditions to understand how contextual information affects annotators' judgments. Our study has two phases: in Phase 1, we vary the *amount* of contextual information; in Phase 2, we vary the *type* of contextual information. In each phase and condition, the annotators were paid the same amount as this study is not focused on understanding how incentive influences the quality of crowdsourced labels. Like [117], we refrain from disclosing the research angle to the annotators in both phases; this helps prevent potential biases during the completion of the HIT.

Phase 1. In Phase 1, the focus is on understanding how the *amount* of dialogue context impacts the quality and consistency of relevance and usefulness labels. We vary the length of the dialogue context to address RQ2.1. Thus, we design our experiment with three variations: C_0, C_3, and C_7 (see Section 3.3.4). The HIT consists of a general task description, instructions, examples, and the main task part. For each variation, we gather labels for two main dimensions (relevance and usefulness) and include an open-ended question to solicit annotators' feedback on the task. Each dimension is assessed with 3 annotators in a separate HIT, with the same system response evaluated by each. This ensures a consistent evaluation process for both relevance and usefulness.

Phase 2. In Phase 2, the focus shifts to the *type* of contextual information, to answer RQ2.2. We take the approach of a machine in the loop for crowdsourcing. We restrict our experiments to experimental variation C_0 (defined below), where no previous dialogue context is available to the annotators. We aim to enhance the quality of crowdsourced labels for C_0 by including additional contextual information alongside the turn being evaluated. We hypothesize that without prior context, annotators may face challenges in providing accurate and consistent labels. By introducing additional context, like the user's information need or a dialogue summary, we expect an increase in the accuracy of evaluations. Through this, we aim to approach a level of performance similar to when annotators have access to the entire dialogue context while minimizing the annotation effort required. We enhance the 40 dialogues from Phase 1 with the user's information need or a dialogue summary, as detailed in Section 3.3.2. Thus, in Phase 2, we have three experimental setups: C_0-llm, C_0-heu, and C_0-sum. Table 3.A.1 in Section 3.A summarizes the setups.

The HIT design closely mirrors that of Phase 1. The main task remains unchanged, except for the inclusion of the user's information need or a dialogue summary. Annotators answer the same two questions on relevance and usefulness in separate HITs. While we do not strictly enforce reliance on the additional information provided, annotators are encouraged to use it when they perceive that the current response lacks sufficient information for an informed judgment.

3.3.4 Experimental conditions

We focus on two key attributes: the *amount* and *type* of dialogue context. For both attributes, we explore three distinct settings, resulting in 6 variations for both relevance and usefulness; each was applied to the same 40 dialogues:

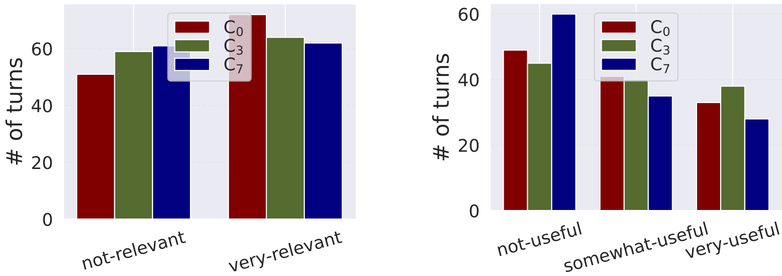
- *Amount of context*. We explore three truncation strategies: no-context (C_0), partial context (C_3), and full context (C_7), designed to encompass scenarios where no previous dialogue context is accessible to the annotator (C_0), where some previous dialogue context is available but not comprehensively (C_3), and when annotators have access to the complete previous dialogue context (C_7).
- *Type of context*. Using the contexts generated in Section 3.3.2, we experiment with three variations of context type: heuristically generated information need (C_0-heu), an LLM-generated information need (C_0-llm), and dialogue summary (C_0-sum).

Table 3.A.1 in Appendix 3.A summarizes the experimental conditions.

3.3.5 Participants

We enlisted master workers from the US on Amazon mechanical turk (MTurk) [15] to ensure proficient language understanding. Annotators were filtered based on platform qualifications, requiring a minimum accuracy of 97% across 5000 HITs. To mitigate any learning bias from the task, each annotator was limited to completing 10 HITs per batch and participating in a maximum of 3 experimental conditions. A total of 78 unique annotators took part in Phases 1 and 2, and each worker was paid \$0.4 per HIT, an average of \$14 per hour. Their average age range was 35–44 years. The gender distribution was 46% female and 54% male. The majority held a four-year undergraduate degree (48%), followed by two-year and master's degrees (15% and 14%, respectively).

We conduct quality control on the crowdsourced labels to ensure reliability as de-



(a) Relevance (b) Usefulness

Figure 3.1: Distribution of (a) relevance and (b) usefulness labels for dialogue annotations in Phase 1.

scribed in Section 3.B in the appendix.

3.4 Results and Analysis

We address **RQ2.1** and **RQ2.2** by providing an overview of the results and an in-depth analysis of our crowdsourcing experiments. We first describe the key data statistics.

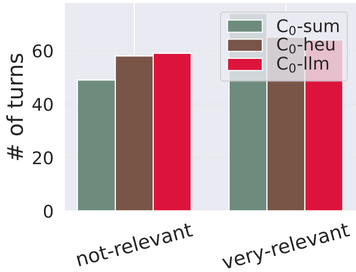
3.4.1 Data statistics

Phase 1. Figure 3.1 presents the distributions of relevance and usefulness ratings across the three variations, C_0, C_3, and C_7. Figure 3.1a indicates a larger number of dialogues rated as relevant when annotators had no prior context (C_0), compared to instances of C_3 and C_7, where a lower number of dialogues received such ratings. This suggests that in the absence of prior context, annotators are more inclined to perceive the system's response as relevant, as they lack evidence to assert otherwise. This trend is particularly prevalent when user utterances lean towards casual conversations, such as inquiring about a previously mentioned movie or requesting a similar recommendation to their initial query, aspects to which the annotators have no access. Consequently, this suggests that annotators rely on assumptions regarding the user's previous inquiries, leading to higher ratings for system response relevance.

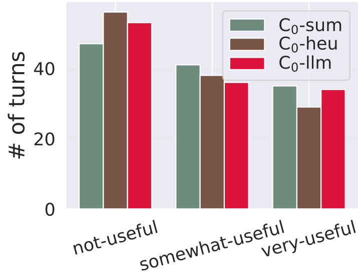
We observe a similar trend for usefulness (Figure 3.1b), compared to C_3 and C_7, C_0 has more dialogues rated as useful. The introduction of the user's next utterance introduced some level of ambiguity to the annotators. Evident in instances where the user introduced a new item not mentioned in the system's response and expressed an intention to watch it, the usefulness of the system's response became uncertain. This ambiguity arises particularly when annotators lack access to prior context, making it challenging to tell if the movie was mentioned before in the preceding context.

These observations highlight the impact of the amount of dialogue context on the annotators' perceptions of relevance and usefulness in Phase 1. This emphasizes the significance of taking contextual factors into account when evaluating TDSs.

Phase 2. In Phase 2, we present findings on how different types of dialogue contexts influence the annotation of relevance and usefulness labels. When the dialogue sum-



(a) Relevance



(b) Usefulness

Figure 3.2: Distribution of (a) relevance and (b) usefulness ratings when annotators have access to additional context in C_0 Phase 2.

mary is included as supplementary information for the turn under evaluation (C_0-sum), a higher proportion of dialogues are annotated as relevant compared to C_0-llm for relevance (60% vs. 52.5%, respectively); see Figure 3.2a.

In contrast to the observations made for relevance, we see in Figure 3.2b that a higher percentage of dialogues are predominantly labeled as not useful when additional information is provided to the annotators. This accounts for 60% in C_0-heu, 47.5% in C_0-llm, and 45% in C_0-sum. This trend is consistent with our observations from Phase 1, highlighting that while system responses may be relevant, they do not always align with the user's actual information need. We find that C_0-sum exhibits the highest number of dialogues rated as useful, indicating its effectiveness in providing pertinent information to aid annotators in making informed judgments regarding usefulness.

3.4.2 RQ2.1: Effect of varying amount of dialogue context

Label quality. To gauge the quality of the crowdsourced labels, we rely on inter-annotator agreement [29, 39]. To understand how the amount of dialogue context influences the quality of ratings by annotators, we calculate the agreement between annotators for both relevance and usefulness across the three variations; see Table 3.1. To address potential randomness in relevance ratings, given the binary scale, we randomly drop one rating from each dialogue and compute the agreement. We repeat this process for each annotator and calculate an average Cohen's Kappa score. For usefulness, we compute Kappa for each pair of annotators and then calculate the average. We assess the significance of the agreement using the Chi-squared method. All Kappa scores are statistically significant ($p \leq 0.05$).

We observe an increase in the Kappa and Tau score as the dialogue context increases from C_0 to C_7. Despite the lack of context in C_0, there is a moderate level of agreement regarding the relevance of the current turn. With the introduction of more context in C_3 and C_7, comes an increase in agreement regarding the relevance of the current turn (see Table 3.1). Providing additional dialogue context seems to lead to higher levels of consensus among annotators. This is likely due to dataset characteristics: users tend to express their preferences early in the dialogue, rather than in subsequent exchanges. Hence, in the case of C_0, which only includes the current

Table 3.1: Inter annotator agreement (Cohen's Kappa) and Tau correlation for relevance and usefulness across the three experimental setups in Phase 1.

Aspect	Variation	Kappa	Tau
Relevance	C_0	0.53	0.47
	C_3	0.61	0.49
	C_7	0.70	0.61
Usefulness	C_0	0.64	0.54
	C_3	0.68	0.60
	C_7	0.56	0.41

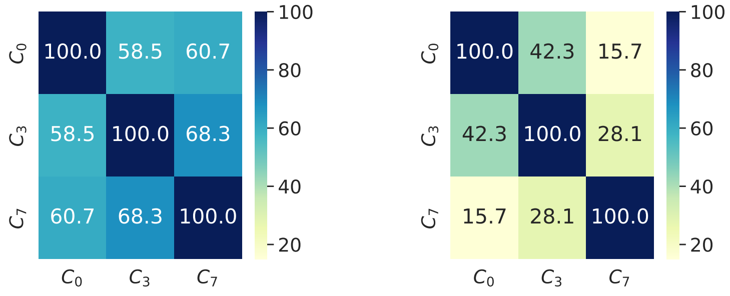
turn, when the user's utterance is incomplete, lacking an explicit expression of their preference, annotators rate more dialogues as relevant compared to C_3 and C_7. Overall, we conclude that when annotators have insufficient information to come up with a judgment, they tend to judge the system positively, introducing a positivity bias [171].

We see in Table 3.1 (row 3) that despite the lack of context in C_0, there is substantial agreement regarding the usefulness of the current turn. This is due to the availability of the user's next utterance, which serves as direct feedback on the system's response, resulting in higher agreement than for relevance assessment. As more context is provided, there is an even higher level of agreement among annotators regarding the usefulness of the current turn. Access to a short conversation history significantly improves agreement on usefulness.

Surprisingly, despite having access to the entire conversation history in C_7, there is a slightly lower level of agreement than in C_3. The complete dialogue context may introduce additional complexity or ambiguity in determining the usefulness of the current turn. This occurs when conflicting feedback arises from the user's next utterance compared to the previous dialogue context. For example, when the system repeats a recommendation that the user has already watched or stated before, and the user expresses their intent to watch the movie in the next utterance, it leads to divergent labels. Similar trend is observed with the Tau correlations though the values are lower compared to the Kappa scores.

Label consistency across conditions. We examine the impact of varying amounts of dialogue context on the consistency of crowdsourced labels across the three variations for relevance and usefulness and report the percentage of agreement in Figure 3.3. We observe moderate agreement (58.54%) between annotations of C_0 and C_3, suggesting that annotators demonstrate a degree of consistency in their assessments when provided with different amounts of context. This trend continues with C_0 and C_7, where the agreement increases slightly to 60.98%. The most notable increase is between C_3 and C_7 (68.29%). As annotators were exposed to progressively broader contextual information, their assessments became more consistent.

Usefulness behaves differently. We observe moderate agreement (41.71%) between C_0 and C_3, indicating a degree of consistency in annotator assessments within this range of context. A notable decrease in agreement is evident when comparing C_3 and C_7, down to 28.3% agreement. The most substantial drop is observed between C_0 and C_7, yielding a mere 14.63% agreement. These findings emphasize the significant



(a) Relevance (b) Usefulness

Figure 3.3: The percentage of agreement in (a) relevance and (b) usefulness labels across the three experimental setups in Phase 1.

Table 3.2: Inter annotator agreement (Cohen's Kappa) and Tau correlation for relevance and usefulness across the three experimental setups in Phase 2.

Aspect	Variation	Kappa	Tau
Relevance	C_0-heu	0.75	0.54
	C_0-sum	0.60	0.45
	C_0-llm	0.51	0.44
Usefulness	C_0-heu	0.71	0.59
	C_0-sum	0.63	0.49
	C_0-llm	0.53	0.44

impact of context on the consistency of usefulness annotations. For usefulness assessment, providing annotators with a more focused context improves their agreement.

With respect to **RQ2.1**, we note considerable differences in the labels assigned by annotators as we vary the amount of dialogue context. As the context expands, annotators incorporate more information into their assessments, resulting in context-specific labels. Annotator judgments are shaped not only by response quality but also by the broader conversation. This highlights the complexity of the task and the need for a carefully designed annotation methodology that considers contextual variations. These findings emphasize the significance of dialogue context in annotator decision-making.

3.4.3 RQ2.2: Effect of automatically generated dialogue context

Label quality. In Phase 2, our experiments aim to establish the impact of presenting annotators with different types of context during crowdsourcing. Different from conventional dialogue context, we provide the annotators with the dialogue summary (C_0-sum), the user's information need in the dialogue (C_0-heu and C_0-llm). We also aim to uncover if we can improve the quality of the crowdsourced labels in C_0 to match those in C_7. We calculate the Cohen's Kappa similar to Section 3.4.2; see Table 3.2.

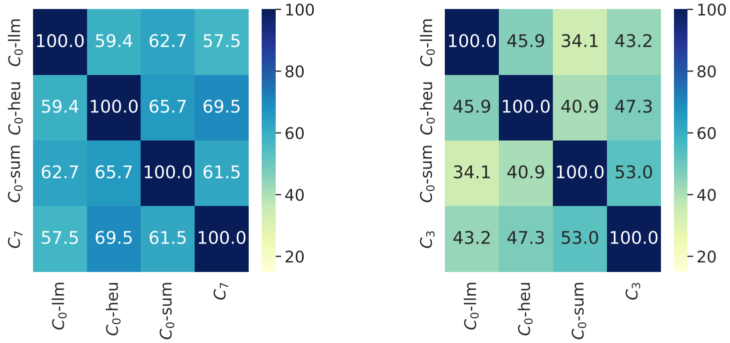
The heuristic approach (C_0-heu) yields the highest agreement (Kappa and Tau), indicating a noteworthy degree of agreement in relevance assessments. The LLM-generated context (C_0-llm and C_0-sum) results in a moderate to substantial level of agreement, signifying a reasonable level of agreement regarding the relevance of the system response. We observe similar results for usefulness. The heuristic approach (C_0-heu) again leads with the highest level of agreement (0.71 and 0.59), C_0-sum follows with a kappa score of 0.63, while C_0-llm has a kappa score of 0.53. This high level of agreement (Kappa) for the two aspects indicates the quality of the labels; the additional context provided, generated either heuristically or with LLMs, is effective in conveying relevant information to annotators, leading to more consistent assessments.

For both relevance and usefulness, C_0-heu consistently improves agreement among annotators, while the LLM-generated context (C_0-llm and C_0-sum) has a substantially lower agreement than C_7. This difference reflects the limitations of LLMs in capturing context and generating a factual summary. While they generate coherent text, LLMs sometimes fail to correctly represent the sequential order of the dialogue and users' language patterns.

Label consistency across conditions. In Figure 3.4a we report the agreement between the setups in Phase 2 and compare them to C_7 (relevance) and C_3 (usefulness) due to their high inter-annotator agreement (IAA) and label consistency. For the relevance annotations, varying levels of agreement emerge. There is substantial agreement between C_0-heu and C_0-llm (59.36%), showing a significant overlap in the labels assigned using both methods, although there are instances where annotators differ in their assessments of relevance. C_0-sum exhibits moderate label agreement with C_0-llm (62.74%) and C_0-heu (65.67%), pointing to relatively similar label assignments across the setups.

We observe similar results for usefulness in Figure 3.4b. While the heuristically generated approach achieves high IAA, the C_0-sum method demonstrates greater consistency with all other setups in terms of usefulness. This suggests that while annotators using the C_0-heu approach often agreed on a single label, the chosen label may not have always been the most accurate. We note slightly low agreement levels for a similar label between the three setups, consistent with results in Phase 1. Unlike relevance, which used a binary scale, usefulness was rated on a 1–3 scale. This finer-grained scale may explain the lower agreement compared to relevance, as different types of contextual information can influence usefulness scores.

Regarding **RQ2.2**, we show that we can improve the consistency of the labels assigned by crowdworkers in the C_0 condition by augmenting the current turn with automatically generated supplementary dialogue context. The heuristic approach demonstrates higher consistency in both IAA and label consistency for relevance and usefulness compared to C_0 and C_7. Providing annotators with the user's initial utterance expressing their preference, particularly in scenarios lacking context, can significantly enhance the quality and consistency of crowdsourced labels. This approach can yield performance comparable to a setup involving the entire dialogue C_7, without imposing the cognitive load of reading an entire conversation on annotators. This streamlines the annotation process and maintains high-quality results, offering a practical strategy for obtaining reliable labels for dialogue evaluation.



(a) Relevance (b) Usefulness

Figure 3.4: The percentage of agreement in (a) relevance and (b) usefulness labels across the three experimental setups in Phase 2.

3.5 Discussion and Implications

Our findings reveal intriguing insights into the impact of context size and type on crowdsourced relevance and usefulness labels for TDS. Expanding the dialogue context from C_0 to C_7 significantly improves agreement among annotators, indicating that annotators rely on comprehensive context to make more accurate assessments. This trend does not hold for usefulness, where we notice a decrease in agreement when all previous dialogue context is available. The optimal amount of context required for reliable labels relies on the aspect evaluated.

Consistent with prior work [69, 114], we observe an inconsistency in relevance labels across variations, with the same system response being rated differently depending on the context provided. Given the lack of label consistency across variations, future studies should carefully tailor their annotation task design and test various settings to ensure high-quality and consistent labels. Additionally, much care should be taken when comparing the performance of a system across several datasets when labels are crowdsourced with a different strategy to ensure a fair comparison, as models similar to humans can be sensitive to the annotation strategy [111, 119].

We also analyzed data from the open-ended question asking annotators about their experience with the annotation task. Annotators note that dialogue summaries fail to convey a user's emotion, limiting their annotation process. Additionally, lower accuracy of the context generated by an LLM may lead to low agreement among annotators. This signifies the importance of carefully considering the quality and accuracy of generated content in the evaluation process. We provide examples in Section 3.E in the appendix. While there may be constraints in presenting user information need and dialogue summary as dialogue context, one key consideration to take into account is the cognitive load of annotators. Providing a shorter, focused context reduces the cognitive burden on annotators, allowing them to devote more attention to actually evaluating a response. This not only streamlines the annotation process but also helps maintain high-quality results. Reducing the amount of content to be assessed may lead to faster annotation times without compromising the quality of ratings [190]. Another approach to using LLMs in annotation is for researchers to consider co-annotation [132] between humans and LLMs.

The optimal context varies by the aspect under evaluation, challenging the idea of a universal strategy. The consistent reliability of automatic methods suggests their potential as dependable tools for evaluation. This implies their use in generating supplementary context, eliminating the need for manual determination of context amounts. This streamlines evaluation, enhancing efficiency in context-driven evaluations for TDS. For data lacking topic or preference shifts, heuristics perform effectively. However, LLMs are recommended for shifting conditions, showcasing adaptability not easily discernible with heuristics.

While our primary focus was limited to relevance and usefulness, the proposed experimental design can be extended to other aspects of TDSs evaluation. Moreover, our findings may be task- or dataset-specific, prompting the need for further investigation into their generalizability. As to future work, we aspire to enhance the robustness of our findings by conducting studies on larger-scale datasets. In addition, following previous work by Kazai et al. [116, 117], we would also want to understand the effect of annotator background: experience of interacting with conversational system or prior experience in doing the annotation task on label consistency for TDSs.

3.6 Conclusion

In this chapter, we investigated the impact of varying the dialogue context size and type on crowdsourced evaluation labels. In particular, we crowdsourced evaluation labels for two aspects: *relevance* and *usefulness*. Our findings reveal that the optimal context is dependent on the aspect under evaluation. For relevance, annotators tend to agree more on a label when they have access to the whole dialogue context. However, this does not hold for the usefulness aspect, where we witness high annotator agreement when partial context is available. We show that a simple approach like providing an automatically generated user need through heuristics without revealing the entire dialogue can consistently increase annotator agreement across the two aspects. This implies that we can rely on automatic methods, such as the use of LLMs, to improve the productivity of crowdworkers by reducing the amount of dialogue they have to read before evaluating the current response.

This study contributes towards how LLMs can be integrated into the annotation process to ensure quality labels from the crowdworkers. In this work, we used GPT-4 API, which is not open source. For future work, we will explore the use of open-source LLMs, like Llama-chat [213], to facilitate a more transparent and reproducible experimental framework.

Addressing **RQ2:** *What is the effect of dialogue context on crowdsourced evaluation labels in task-oriented dialogue systems?*, we find that context presentation significantly impacts annotation quality. Our experiments demonstrate that LLM-generated summaries can effectively replace full dialogue history, maintaining evaluation reliability (agreement scores comparable to full context) while reducing cognitive load. This provides a practical solution for scaling dialogue evaluation.

While this chapter focused on the preceding context, another crucial aspect of evaluation is how annotators interpret system responses based on user reactions. Beyond prior context, users provide implicit feedback through their follow-up utterances, signaling satisfaction, confusion, or dissatisfaction. In the next chapter, we examine how assessors, both human and LLM-based, incorporate user feedback into their judgments.

Chapter Appendices

In this section, we provide supplementary materials used to support our main paper. These materials include: experimental conditions elaborated in Section 3.A, quality control measures undertaken to ensure high quality crowdsourced labels and generated supplementary context in Section 3.B and the prompts used to generate the supplementary context in Section 3.C. In Section 3.D we include the annotation instructions and screen dumps of our annotation task. Section 3.E shows sample supplementary context generated by GPT-4.

3.A Experimental Conditions

We list the experimental conditions used for our crowdsource experiments in Table 3.A.1.

Table 3.A.1: Descriptions of the experimental setups used for the crowdsourcing experiments with corresponding relevance and usefulness labels. Unlike relevance, usefulness includes the user's next utterance as feedback. A "turn" denotes a user-system exchange.

Variations	Description
C_0	Current turn with no previous dialogue context
C_3	Current turn with three system-user utterances as previous context
C_7	Current turn with 7 user-system utterances as previous context
C_0-llm	Current turn with an LLM-generated user information need as dialogue context
C_0-heu	Current turn with a heuristically generated user information need as dialogue context
C_0-sum	Current turn with a dialogue summary as dialogue context

3.B Data Quality Control

Generated user information need and summary. To address the potential hallucination of LLMs [43], we implemented a quality control process for the generated user information needs and summaries, ensuring their coherence and factual accuracy. We automatically cross-reference the movies mentioned in both the input dialogues and the summaries. A summary must contain at least two-thirds of the movies mentioned in the input dialogue to be considered valid. If this criterion is not met, the summary is discarded, and a new one is generated following the specified prompt requirements. In total, we discarded and regenerated 15 dialogue summaries. To further ensure coherence, we randomly sampled 30% of the generated summaries and information needs. The authors reviewed them to confirm their coherence and alignment with the information presented in the input dialogue. This process enhanced the quality and reliability of the generated content.

Crowdsourced labels. To ensure a high quality of the collected data, we incorporated attention-checking questions into the HIT. Annotators were required to specify the number of utterances in the dialogues they were evaluating and to identify the last movie mentioned in the system response being evaluated. 10% of the HITs were rejected and returned back to collect new labels. In total, we gathered 1440 data samples from the crowdsourcing task, spanning six variations for relevance and usefulness. We employed majority voting to establish the final relevance and usefulness dialogue label.

3.C Prompts

In Table 3.C.1 we show the final prompts used to generate the user information and dialogue summary with GPT-4.

Table 3.C.1: Prompts used to generate the supplementary context; user information need and dialogue summary with GPT-4.

Dialogue summary prompt
Below you are provided with dialogues between a user and the system about movie recommendations. Generate a complete short and informative summary extractively which is half the length of the dialogue.
User information need prompt
Given the following user and system dialogue in a movie recommendation conversation, generate a concise user's goal in a natural manner. State only the goal without extra text. Start the sentence with "the user wants."

3.D Annotation Instructions and Screen Dumps

Table 3.D.1 details the annotation instructions for the relevance and usefulness evaluations. In Figure 3.D.1 and 3.D.2 we show the annotation interface used for Phase 1 and Phase 2, respectively.

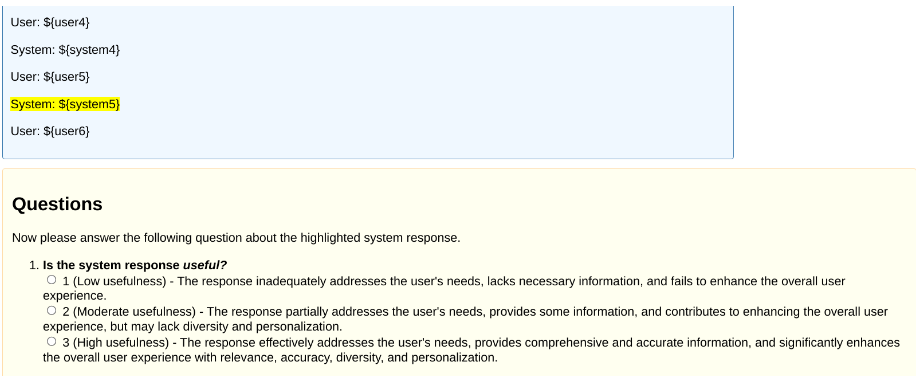


Figure 3.D.1: Annotation interface for phase 1 when evaluating response usefulness for C_3

Table 3.D.1: Annotation instructions provided to the annotators for relevance evaluation. The instructions are the same for usefulness apart from the aspect being evaluated.

Introduction
Thank you for helping us out! Below we explain everything in full detail. Please make sure to read the instructions carefully.
Purpose
This survey aims to evaluate the quality of a system's response. We want to evaluate the dialogue system's performance and gather insights for improvements. We will ask you to evaluate the system response on one metric, that we will discuss in more detail below.
Scenario Outline
Imagine you are evaluating a dialogue system that generates a response to user queries. Your task is to assess the response based on relevance. We will provide examples and detailed explanations of this criteria below.
Task
In each HIT, you will be presented with a dialogue chunk. Your task is to evaluate the last system response based on the given criteria. Please review the explanations and examples for the criteria to ensure your understanding before proceeding with the evaluation. Keeping the scenario that was outlined above in mind, we would like to ask you to judge the system response on relevance.

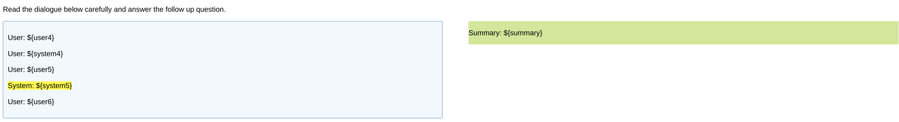


Figure 3.D.2: Annotation interface for phase 2 when evaluating response usefulness with supplementary context

3.E Generated Context

In Table 3.E.1 we show sample user information needs and a summary generated by GPT-4.

Table 3.E.1: Sample dialogue summaries as supplementary context generated by GPT-4.

Dialogue 1
User inquires about a good family movie recommendation similar to ”Real Steel (2011)” or ”The Lego Movie (2014)”. System recommends ”Super (2010)”, an action-comedy about a regular guy who becomes a self-made superhero, describing it as hilarious and entertaining. The user shows interest in this recommendation.
Dialogue 2
The user asked for coming-of-age movie recommendations and mentioned they enjoyed ”My Girl (1991)” and ”Lucas (1986)”. The system suggested watching ”The Spectacular Now (2013)”, a film where Shailene Woodley stars as a character who forms a bond with a troubled classmate.
Dialogue 3
User seeks a dramatic love story to watch. System recommends ”The Notebook (2004)”, but the user has watched it, as well as ”Titanic (1997)”. Both films are favored by the user; they desire to watch something new.
Dialogue 4
The user requests animated movie recommendations following their enjoyment of ”The Incredibles (2004)”. The system suggests other movies, including ”Monsters, Inc. (2001)” and its sequel ”Monsters University (2013)”, which the user approves. The conversation pivots to the topic of successful sequels, citing ”Toy Story 3 (2010)” as an example despite the user’s disagreement, favoring the original movie, ”Toy Story (1995)”.
Dialogue 5
The user wants to find a thrilling crime movie like ”Thor: Ragnarok (2017)” for their weekend. The system suggested they watch ”The Snowman (2017)” but the user declined. However, the system then gave another recommendation, ”First Kill (2001)”.

4

Effect of User Feedback on Humans and LLMs

The evaluation of conversational systems typically involves annotators assessing isolated dialogue turns. However, in natural conversations, users provide valuable feedback about system performance through their responses, such as indicating satisfaction, confusion, or frustration. In this chapter, we continue investigating the reliability of evaluation labels from Chapter 4 by considering another form of context – the user's feedback through follow-up utterances. We address the following research question:

RQ3: How does incorporating user feedback through follow-up utterances affect evaluation judgments by humans and LLMs, and what does this reveal about their respective strengths as annotators?

We examine how access to user feedback through follow-up utterances affects evaluation quality, comparing how human annotators and LLMs assess system responses with and without this additional signal. Through an in-depth analysis, we identify when and how this feedback helps evaluators make more informed and consistent judgments about system performance and show that humans excel at interpreting implicit feedback when assessing usefulness, while LLMs show consistency in evaluating factual aspects like relevance.

4.1 Introduction

Evaluation of systems has been an integral part of the information retrieval (IR) research agenda for decades [e.g., 53]. Traditionally, IR evaluation has relied highly on user actions, including implicit feedback such as click-through rates. However, in a conversational setting such signals are not usually available due to the nature of the interactions. As a result, the evaluation of dialogue systems increasingly relies on human evaluation, leading to a growing interest in user-centric evaluation methods [245]. However, asking for explicit user feedback from a user can be intrusive and may negatively impact user experience [248]. Therefore, in recent years, the assessment of

This chapter was published as C. Siro, M. Aliannejadi, and M. de Rijke. Rethinking the evaluation of dialogue systems: Effects of user feedback on crowdworkers and LLMs. *In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, pages 1952–1962.* ACM, 2024.

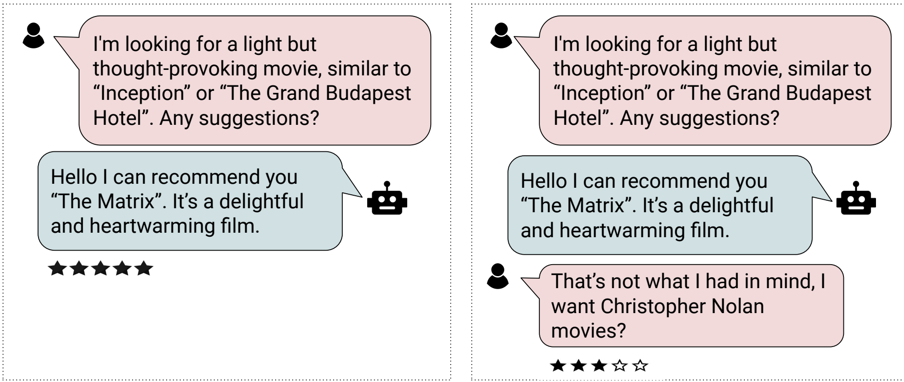


Figure 4.1: A dialogue showing an example of a complex user request with (right) and without (left) the user feedback. The star ratings show the assessment of external assessors judging the usefulness of the system utterance. As can be seen, based on the follow-up utterance the assessors lower their usefulness rating aligning with the user feedback.

conversational systems has relied on crowdsourced evaluation, leveraging the collective wisdom of human annotators.

Turn-level assessments. When gathering evaluation feedback on individual turns in a conversational interaction, various design methods have been considered in the past. These include deciding on the type of judgment scale, as well as formulating annotation guidelines and methods for presenting dialogues under assessment at the turn level [185, 203]. Recent strategies for presenting turn-level utterances to annotators involve two main approaches: one displays both the user's initial request and the system's response for evaluation [201, 207], and the other shows only the system's response [157]. The choice between these approaches often depends on the specific evaluation metric in question. The first method, similar to the query–document pair setting in ad-hoc retrieval systems, operates under the premise that the user's initial request offers sufficient context for annotators to make well-informed evaluations.

Follow-up utterances. Users often do not articulate all of their intentions in a single request. Rather, they engage in an iterative dialogue, clarifying and refining their intentions through successive exchanges [7]. Their queries can be multifaceted, ambiguous, or overly generic, which complicates the process of evaluating individual turns. E.g., in Figure 4.1, the user poses a multifaceted query, leaving substantial room for interpretation. First, the user is looking for a movie suitable for an evening watch, which typically suggests something not overly long or intense. Second, they desire a film that is “light but thought-provoking,” implying a blend of easy-to-digest content with depth in storytelling. Last, by referencing specific movies like “Inception” and “The Grand Budapest Hotel,” the user indicates a preference for a certain style or genre — perhaps celebratory narratives with unique storytelling or visually engaging films. Annotators tasked with evaluating the system's response to this query must consider these multiple layers.

When annotating turns in a conversational interaction, the complexity for annotators lies in assessing the system's response not just for its relevance to an overt request, e.g., for a movie recommendation, but also for its alignment with nuanced, implicit preferences indicated by the user. Systems may not always successfully address all aspects of such requests. When the system's response only partially meets the query's criteria, it becomes challenging for annotators to gauge which aspect of the request was most critical to the user. We believe that this is where a user's follow-up utterance be particularly informative. A user's subsequent response may provide valuable insights into what they valued most in their original request, giving annotators a clearer indication of the user's priorities. Thus, follow-up utterances may serve as crucial cues, helping annotators make a more informed assessment of the system's performance, particularly in how well it navigates and prioritizes the multifaceted aspects of a user's complex request.

Is a user's follow-up utterance crucial in ensuring evaluations align with actual user needs, especially since annotators, as external evaluators, may not fully grasp the user's perspective or context? We hypothesize that annotators who have access to a follow-up utterance produce more accurate and user-centric evaluations, improving the quality of evaluation labels in the process.

Research goals. We investigate the effect of a user's follow-up utterance on the annotation of turns in a task-oriented dialogue system (TDS). We conduct experiments with two types of annotators: human and large language model (LLM)-based. Both types of annotators are asked to provide annotations of turn-level system responses along four dimensions: *relevance*, *usefulness*, *interestingness*, and *explanation quality*, on a 100 level scale, following [185]. We consider two contrastive setups for annotators to provide these annotations: (**Setup 1**) does not consider the user's follow-up utterance, and (**Setup 2**) does consider the user's follow-up utterance. In addition, we collect data on what sources of information human annotators rely on to arrive at their judgments. With this data, we aim to examine the bias introduced by these sources to the crowdsourced labels.

We use a subset of the recommendation dialogue (ReDial) [133] dataset to address the following chapter-level research questions:

RQ3.1 How does a user's implicit feedback from the follow-up utterance influence the evaluation labels collected from both human annotators and LLMs?

RQ3.2 When is implicit user feedback significant in the evaluation of TDSs?

RQ3.3 What are the annotators' perceptions in terms of the sources of information they rely on to make their assessments, and what are the potential biases might that have on their performance?

Findings. Our findings indicate that both the crowdworkers and the LLM exhibit sensitivity to user cues from follow-up utterances. There is a significant difference in the mean ratings from both annotators except for relevance when follow-up utterance is included, indicating user feedback does influence system evaluation. Workers are more susceptible to user feedback in usefulness and interestingness, compared to LLMs in interestingness and relevance. Specifically, there is a clear distinction in relevance and usefulness ratings of crowdworkers **Setup 2** ratings unlike in **Setup 1** where these aspects are often conflated. This indicates that crowdworkers not only evaluate response usefulness based on topical relevance but also align with user needs and preferences expressed in follow-up utterances. This suggests that follow-up utterances enable a more personalized assessment of usefulness, aligning closely with the user's explicit feedback. In **Setup 2**, we observe an increase in annotator agreement. This is particularly evident in scenarios characterized by uncertainty in user requests. Complex user requests with multiple criteria or preferences posed challenges during evaluation, but follow-up utterances helped to clarify the user intent. Similarly, generic user requests, initially broad and challenging to address, became more focused on follow-up utterances, allowing human-/LLM-based annotators to tailor their assessment effectively.

These findings not only show the significance of user feedback in system evaluation but also provide a foundation for integrating user feedback in the automatic evaluation of conversational systems.

4.2 Related Work

Recent studies emphasize evaluating TDSs through a user experience lens [60]. Traditionally, TDSs were assessed primarily for task completion. While task completion remains a fundamental criterion, there is a growing recognition that solely measuring task success may not provide a comprehensive understanding of system performance [201]. Consequently, there is a shift towards incorporating user experience metrics into the evaluation of TDSs [200]. Here, we provide a brief overview of the studies on TDS evaluation from the perspective of user feedback, evaluation bias, and LLMs.

4.2.1 User feedback

In web search, implicit user signals including click-through rates and dwell time on search results are available in vast amounts and these signals are leveraged to evaluate a system and continually improve the search results [122, 123, 141, 142, 248]. However, such signals are not accessible for conversational systems due to their interactive nature. Consequently, automatic evaluation of conversational systems would primarily rely on explicit user feedback in the form of ratings [28, 48], which could be intrusive and lead to poor user experience [27]. In this work, we propose using the user's next utterance as a proxy for both explicit and implicit user feedback. For instance, when a user expresses satisfaction or dissatisfaction in their next utterance following a system recommendation, it represents explicit feedback that should not be disregarded when assessing system performance. Our study investigates how user feedback from the next utterance influences the evaluation labels provided by both human- and LLM-based annotators.

4.2.2 Bias in crowdsourcing evaluation labels

The use of crowdsourcing for evaluating TDSs in IR research, while offering scalability and diversity, brings inherent biases. Research, such as [19, 69, 190], highlights cognitive biases and load as significant factors influencing crowdworker judgments, which can skew the evaluations. For example, workers' preconceived notions or mental fatigue might lead to inconsistent results. Further, Hube et al. [101] and Han et al. [89] emphasize the impact of workers' relevance strategies and personal biases on their assessment of IR systems. To mitigate these biases, strategies such as task design adjustments and worker training are suggested [76]. For instance, presenting tasks neutrally and providing clear, unbiased instructions can help reduce bias, as discussed in [101]. Additionally, the choice and implementation of judgment scales, as explored in [24], play a crucial role in assessor bias. Different from previous studies, in this work we focus on assessing how the sources of information relied on by workers to make their judgment bias their evaluation labels.

4.2.3 LLM-as-judge

Recently, there has been a notable surge in the use of LLMs as annotators in various tasks [43]. These models show good performance comparable to human annotators and in some cases outperform them [82]. Additionally, they have proven to reduce the time and cost for annotation, making them a preferred choice compared to human annotators [222]. However, most research efforts have primarily focused on assessing how well LLMs' labels correlate to human labels [63, 94]. There has been a relative lack of investigation into whether LLMs are susceptible to the same influencing factors as crowdworkers.

Several studies have delved into understanding the factors that impact crowdworkers, encompassing aspects like task design, judgment scales, and protocols designed to enhance the quality of annotations [87, 117, 186]. These studies have contributed valuable insights into optimizing crowdworker performance. However, a similar examination of the impact of these factors on LLMs is notably absent from the research landscape. We contribute towards understanding how task design influences evaluation labels assigned by LLMs. We investigate the influence of user feedback from the user's follow-up utterance on the evaluation of TDSs by both crowdworkers and LLMs.

4.3 The Annotation Task

Our objective is to understand the influence of user feedback from the user's follow-up utterance on the evaluation of TDSs. We conduct our study as an annotation effort with crowdworkers from amazon mechanical turk (AMT) [15]. Additionally, due to the increased use of LLMs as annotators [63, 94] we seek to understand how LLMs are affected by feedback from the user's next utterance. User feedback in this case can be either implicit or explicit. Explicit feedback refers to straightforward, direct responses from users, like specific comments on a certain dialogue aspect (e.g., "I don't like this movie"). Implicit feedback is more subtle, encompassing aspects like tone or contextual hints within the user's follow-up utterance (e.g., "Thanks for the suggestion. How about some action movies?"). We gather annotations for four fine-grained dialogue qualities: relevance, usefulness, interestingness, and explanation quality, across two experimental conditions.

4.3.1 Dialogue qualities

We experiment with four dialogue qualities in the domain of TDSs [21, 201] that have been investigated extensively, elaborated below.

Relevance. The relevance [14, 107, 149, 201] of a dialogue response is a crucial factor in assessing the effectiveness of a TDS. To evaluate relevance, we ask the workers to determine how well the system's response addresses the user's request. This aspect gauges the system's ability to understand and appropriately respond to user input.

Usefulness. Usefulness [151, 203, 216] of a dialogue response pertains to its practical value from the user's perspective. Apart from just being relevant, workers assess whether the system's response gives additional information to the user on the recommended item. In **Setup 2** the user's perspective is captured by asking workers to rely on the user's follow-up utterance to gauge the usefulness of the system response. E.g., if a user says, "I have already watched that," it suggests that the recommendation is not new or helpful to the user, even though it might be relevant. Usefulness helps to measure the system's overall utility in real-world scenarios.

Explanation quality. Understanding how well a TDS communicates its reasoning is important for user trust and comprehension. Explainability in IR systems has witnessed a notable surge in recent times [21, 85, 249, 250]. Following Guo et al. [85] we instruct workers to assess explanation quality, by evaluating the clarity and informativeness of the system's justifications or explanations accompanying its responses. This aspect provides insights into the system's transparency and user-friendly communication.

Interestingness. Beyond system functionality, the interestingness [200] of a system response adds a subjective layer to the evaluation. Workers are asked to evaluate whether the system's responses are engaging, captivating, or exhibit qualities that make the interaction more enjoyable for the user. This encapsulates the language used to make recommendations by the system. This aspect contributes to a holistic assessment of user experience.

4.3.2 Data

We use the ReDial dataset [133], a well-known collection of over 11,000 dialogues specifically focused on movie recommendations. We sample the dialogue turns for annotation by focusing on selecting user utterances that explicitly request movie recommendations or express movie preferences. Phrases like "I prefer," "recommend me," and "my favorite" were key indicators in this selection process. This approach ensured that our sampled data contained user utterances that were explicit and straightforward, facilitating a more accurate assessment of the dialogue system's responses.

Similar to Guo et al. [85], we observe a lack of in-depth explanations in the ReDial dialogues. Our initial analysis of the dataset indicates that longer system utterances often include attempts to explain movie recommendations, whereas shorter ones do not. As a result, we selected system utterances with more than 14 words (average length of system responses in the dataset) to better focus on responses that are more likely to include explanations. In total, we sampled 100 unique dialogue turns from the dataset, each representing a different conversation.

4.3.3 Annotation scale

Following the approach outlined in [185], our study adopts the S100 scale for evaluation purposes. This scale is employed through a sliding window mechanism, allowing annotators to provide detailed feedback on the dialogue systems. The sliding scale's interactive nature enables a more precise and flexible assessment compared to traditional binary or categorical scales. To enhance its usability and ensure intuitive responses, the default value on this slider is set to 0. This design choice is based on the rationale that a neutral starting point encourages annotators to consciously adjust the slider based on their judgment of the dialogue turn's effectiveness, rather than being biased by any preset values. To ensure consistency and accuracy in evaluations, we provide the annotators with several examples demonstrating how to effectively use the S100 scale. We adopt the same scale for annotation with LLM.

4.3.4 Preliminary experiments

Our research included preliminary experiments to refine the design and methodology. These experiments assessed the practicality of our setups, refined annotation guidelines, and identified data collection challenges. Two setups were tested:

Exp 1 *Single worker, two conditions*: Workers evaluated a dialogue turn under two conditions within a single human intelligence task (HIT). The only difference was the presence or absence of the user's follow-up utterance, which served as user feedback.

Exp 2 *Random assignment of conditions*: Workers were randomly assigned to one of the two conditions to incorporate diverse perspectives and reduce potential biases, gathering a range of rationales behind annotator evaluations.

Preliminary results. We used 13 dialogue turns, primarily focusing on comparing **Exp 1** and **Exp 2** to determine the most effective approach. The mean ratings obtained from both setups indicated a high degree of consistency in annotator assessments for relevance and usefulness, suggesting that both methods performed similarly in these aspects. However, an interesting observation emerged concerning annotation time and the diversity of justifications. In **Exp 1**, resulted in shorter annotation times but exhibited limited diversity in justifications. In contrast, **Exp 2**, yielded a more diverse set of justifications. Considering these findings, we decided to proceed with the **Exp 2** setup for our main experiments. This choice was motivated by the goal of obtaining a broader and more diverse range of annotations and justifications, a critical requirement for the comprehensive evaluation of dialogue systems in our study.

4.3.5 Experimental conditions

Following **Exp 2**, we designed two distinct experimental conditions to evaluate the effect of user feedback on the evaluation of TDSs with human annotators, as well as LLMs.

Setup 1 Following the conventional annotation method, this condition provides only the initial user query and the system's response to the annotators and LLMs, omitting the user's follow-up utterance. This setup focuses on evaluating the TDS based on a single interaction, reflecting the traditional approach in dialogue annotation.

Setup 2 This condition incorporates the user's follow-up utterance along with the initial query and the system's response. The aim is to allow annotators and LLMs to evaluate the TDS within the full context of the conversation, assessing the impact of subsequent user feedback on annotations.

4.3.6 Human annotators

For this task, we recruited master workers from AMT. We employed multiple HIT templates to conduct our study, aiming to investigate the impact of the user's next utterance on annotator ratings for various aspects. We collected annotation labels for relevance, usefulness, interestingness, and explanation quality in the two experimental conditions. Each aspect was annotated in a separate HIT. Importantly, we did not disclose the research angle to the annotators, framing it as an annotation effort.

During each HIT, we provided the annotators with instructions, definitions of the aspect to be assessed, and examples. We maintained consistent instructions across all aspects and setups, with variations limited to definitions and examples. In each HIT, annotators rated the aspect and provided justifications for their ratings, a practice known to reduce randomness and enhance assessment quality [153]. Additionally, we sought to understand the sources of information relied on by annotators to make their assessments, asking them to select sources they considered when making their assessments, including personal knowledge, external sources such as web searches, educated guesses, user's request, system response, user's feedback, and an "other" option for sources not covered in the provided options.

84 unique workers participated in the study (46 female and 38 male), with an average age of 30–45. Each worker received a reward of $0.4 per HIT, which was determined based on the minimum wage.

4.3.7 LLMs as annotator

Since LLMs have shown a notable performance as annotators, we investigate whether LLMs are influenced to a comparable degree as human annotators with user feedback from the user's follow-up utterance. This investigation seeks to shed light on the potential similarities and differences in how LLMs and human annotators respond to such contextual input, ultimately contributing to an understanding of LLM behavior in annotation tasks.

We used ChatGPT (gpt-3.5-turbo API[1]) a subseries of GPT models [34] for the annotation task. Typically, crowdworkers are provided with annotation examples to improve the quality of the labels they assign, similarly, we provide the same to ChatGPT, thus using it in the few-shot setting. To ensure consistency, we replicate the experimental conditions used for human annotators across the four aspects. In our experiments, we employ two distinct experimental conditions, varying the prompts and data presentation to align with these setups.[2] For each aspect, we provided ChatGPT with the corresponding human annotation instructions. This approach allows us to comprehensively assess the performance of ChatGPT in generating evaluation labels in a manner that mirrors human annotation practices.

[1] Temperature = 1, Top p = 1

[2] The prompts are available in Appendix 4.A.

Table 4.1: ICC and pairwise Cohen's Kappa for all aspects across both setups.

	ICC		Kappa	
Aspects	Setup 1	Setup 2	Setup 1	Setup 2
Relevance	0.8358	0.7427	0.6978	0.5701
Usefulness	0.7553	0.7419	0.5892	0.5631
Interestingness	0.5456	0.4685	0.2825	0.2231
Explanation quality	0.5136	0.5351	0.2380	0.2812

4.4 Crowdsourced Judgments

Before addressing our research questions (in Sections 4.5–4.7), we examine the judgments collected through crowdsourcing.

Internal agreement. To assess the quality of labels collected from the crowdworkers, we compute pairwise Cohen's Kappa and report the results in Tab. 4.1. The Kappa scores indicate varying levels of agreement in different evaluation setups and aspects:

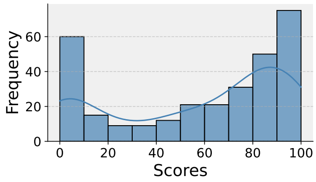
Relevance shows a substantial agreement in **Setup 1**, compared to **Setup 2** (Kappa: 0.69 vs. 0.57). This indicates that the inclusion of the user's follow-up utterance during evaluation introduces complexity, impacting crowdworkers' judgments. Both setups in *usefulness* exhibit high agreement, indicating that the presence or absence of user follow-up utterances has minimal influence on crowdworkers' perceptions of the system's response utility. This suggests that, overall, crowdworkers are consistent in the evaluation of usefulness. **Setup 1** shows a moderate agreement (Kappa: 0.28) in evaluating *interestingness* while **Setup 2** has a slightly lower agreement (Kappa: 0.22), reflecting the added complexity introduced by the user's follow-up utterance. In *explanation quality*, **Setup 2** exhibits higher agreement (Kappa: 0.28 vs. 0.23), possibly because it allows crowdworkers to assess explanations within a broader context that includes both the initial response and the user reaction in the follow-up.

We make similar observations with the intraclass correlation coefficient (ICC) values in the two setups. In general, workers exhibit a substantial to moderate agreement for all aspects. The low agreement in interestingness and explanation quality can be attributed to the nature of their subjectivity and the large scale of evaluation.

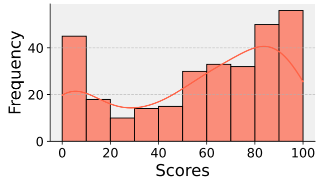
Crowdworker judgments. Figure 4.2 shows the distributions of scores provided by crowdworkers for the four dialogue qualities over the two setups.

Relevance scores in both setups display a long-tailed distribution toward the extremes, as shown in Figure 4.2a and 4.2b. The median rating in **Setup 1** is 71 with 65 for **Setup 2**, suggesting that including the user's follow-up utterance leads to lower relevance scores from the workers. The scores for both setups range between 0 to 100.

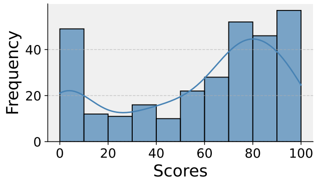
In **Setup 1**, there is a noticeable drop in *usefulness scores* within the 20–40 range, as seen in Figure 4.2c. This setup also indicates similarity in the distribution of relevance and usefulness scores, suggesting that, without user feedback, workers tend to rate usefulness similarly to relevance due to limited information. In **Setup 2**, which includes user follow-up, there is a decrease in responses rated as not useful (0–20) and an increase in the 30–70 range. This indicates that some responses are considered useful even when not directly relevant to the user's initial query, possibly because users



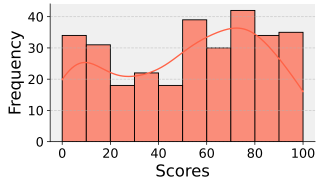
(a) Relevance



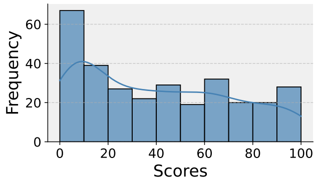
(b) Relevance



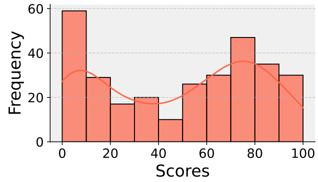
(c) Usefulness



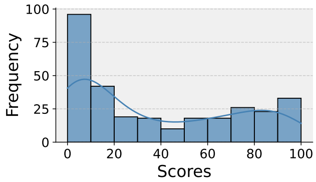
(d) Usefulness



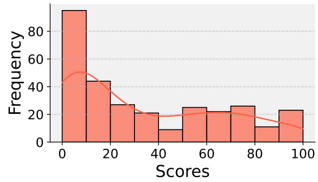
(e) Interestingness



(f) Interestingness



(g) Explanation quality



(h) Explanation quality

Figure 4.2: A comparison of individual worker scores distributions for **Setup 1** (left column) and **Setup 2** (right column).

found the recommended movie intriguing or had not considered it before. The median scores are 70 and 56 respectively in Setups 1 and 2.

The *interestingness* aspect is highly subjective, therefore making it more prone to individual worker bias as observed with the moderate agreement between workers in Table 4.1. In **Setup 1** the scores are skewed towards the left, indicating most workers found the system responses less interesting (see Figure 4.2f) with a median score of 37 compared to 50 for **Setup 2**.

Scores for *explanation quality* are skewed towards the left, with a median score of 25 (**Setup 1**) and 22 (**Setup 2**), as shown in Figure 4.2g and 4.2h. This shows that even though most workers find that system recommendations relevant there is a lack of explanation on why the recommendations are made. These findings are in line with recent work conducted by Guo et al. [85], showing that most conversational recommender system (CRS) dialogues lack explanation in their recommendations.

4.5 Effect of User Feedback

In this section, we answer **RQ3.1**: *How does user feedback from the follow-up utterance influence the evaluation labels collected from both crowdworkers and LLMs?*

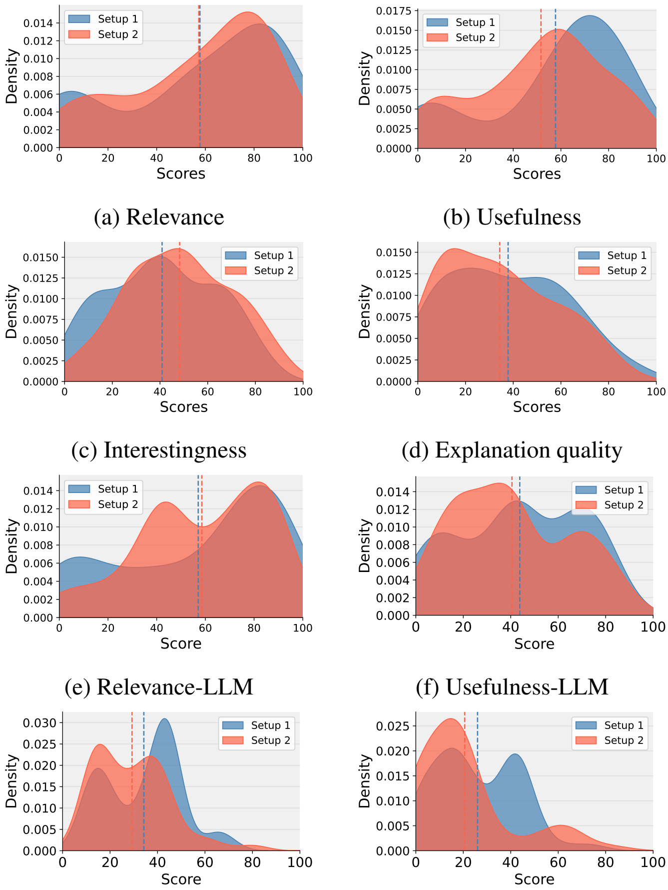
Distributions. For the *crowdsourced labels*, each turn is annotated by three workers; their ratings are averaged to get the overall score per turn. Figure 4.3 (a)–(d) show the density distributions of the scores: *Relevance and usefulness* are skewed towards the right (Figure 4.3a and 4.3b) for both setups, showing that more turns are found to be more relevant and useful by the crowdworkers. For *usefulness*, the peak for **Setup 2** is towards the center compared to **Setup 1**, indicating a decrease in the number of turns that are highly useful as workers have access to the user's follow-up utterance, adding more context during the assessment. Cases where the system makes a relevant recommendation, are rated highly useful in **Setup 1**, but the user's feedback in **Setup 2** changes the worker's rating to lower values in certain cases. E.g., in cases where the user has already watched the movie or even though the movie satisfies their requirements (e.g., genre and actor), they do not like other aspects. *Interestingness* has a more central peak with a wide range showing there was a lot of variability in the assessment (Figure 4.3c). More turns are assessed as interesting in **Setup 2** compared to **Setup 1**, with fewer turns being scored as highly interesting in both setups (80–100). Similar observations pertain to *explanation quality* (Figure 4.3d).

We also plot the distribution of scores from the LLM in Figure 4.3 (e)–(f). The *relevance* kernel density estimation (KDE) plot exhibits a dual-peak distribution with a minor peak at lower values and a more pronounced peak at higher values, notably around 80 and above (Figure 4.3e). In **Setup 2**, the KDE plot shows a distribution peaking in the mid to higher range of the score scale.

In contrast, *usefulness* (Figure 4.3f) shows three peaks in **Setup 1** (not significant from each other), with **Setup 2** having a distinctive peak between scores of 10–40. The slight peak towards the high scores compared to **Setup 1** indicates that with the user's follow-up utterance, the LLM finds most turns not to be highly useful. We observe a different pattern for *interestingness* and *explanation quality* with scores skewed towards the left showing that the LLM rates most turns low on interestingness and explanation similar to observations made from the crowdworker scores.

Humans vs. LLMs. The different peaks between the two setups across the four aspects indicate a significant divergence in how crowdworkers and LLMs perceive and rate the aspects in different setups. **Setup 1** is characterized by high peaks in high-range scores, compared to **Setup 2** which exhibits peaks in the moderate range except for relevance, which has both moderate and high peaks. This contrast suggests that user feedback from the follow-up utterance has a notable impact on both the crowdworkers and LLM assessments.

External agreement. Next, we compute the overall mean score for each setup for both the crowdworkers and the LLM with confidence intervals; see Figure 4.4. There is no significant difference in *relevance scores* between the two annotator groups. This indicates that the presence of the user's follow-up utterance does not significantly affect the relevance assessment. Relevance primarily relies on the system's ability to provide a topically relevant recommendation to the user's request, and having only the user's



(a) Relevance (b) Usefulness

(c) Interestingness (d) Explanation quality

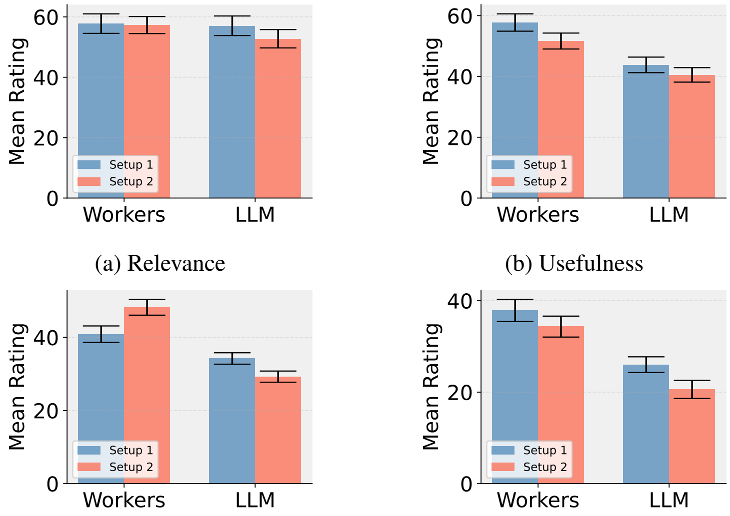
(e) Relevance-LLM (f) Usefulness-LLM

(g) Interestingness-LLM (h) Explanation qual.-LLM

Figure 4.3: Kernel density estimation plots comparing aggregated crowdworker and LLM scores for both setups. The dotted lines represent the overall mean for each setup.

initial request appears sufficient for this assessment. However, some differences in relevance scoring emerge when the follow-up utterance is available to certain workers. In many instances, workers influenced by the follow-up utterances tend to assign high scores to non-relevant system responses if the user accepts the recommendation, even if it deviates from their initial request. Conversely, they may assign low scores to relevant system responses if the user dislikes the recommendation based on specific attributes, despite its topical relevance. Similar patterns are observed in the scores assigned by LLM, with the LLM having a low mean score for **Setup 2**.

LLM-based annotations are consistently lower in terms of *usefulness scores* compared to crowdworkers', as indicated by lower mean scores in both setups (Figure 4.4b). The mean scores are statistically significant for crowdworkers but not for the LLM, highlighting the substantial influence of the follow-up utterance on useful-



(a) Relevance

(b) Usefulness

(c) Interestingness

(d) Explanation quality

Figure 4.4: Mean rating for each aspect across the two setups, for both the crowdworkers and LLM.

ness assessments. In **Setup 1**, crowdworkers assign high usefulness scores to system responses, closely aligning with relevance scores. This suggests that annotators assess usefulness like relevance in the absence of the follow-up utterance.

Conversely, in **Setup 2**, there is a significant drop in the mean usefulness score. This reflects workers transitioning from assessing relevance to considering how well the system response addresses various facets of the user's needs, often revealed in the follow-up utterance. E.g., a user may request an action movie initially, but specific preferences may emerge in the subsequent utterance, such as actor or director preferences as users typically reveal their complete information needs through a back-and-forth exchange [7].

In contrast to other aspects, *interestingness* presents an intriguing observation. Workers assign lower scores in **Setup 1** compared to **Setup 2** (Figure 4.4c). Both groups of annotators assign lower scores for interestingness in both setups, and these differences in mean scores are statistically significant. Examination of annotators' justifications reveals that they hold strict criteria for rating system responses as interesting and had relatively high expectations for a response to be deemed interesting. This is reflected in the score distributions depicted in Figure 4.3c, where a smaller proportion of turns receive a score of 100 for interestingness. The disparity between the setups is particularly notable in **Setup 1**, where only 2% of the turns received a score of 90 or higher, as opposed to 7% in **Setup 2**.

The mean score for explanation quality is not statistically significant for crowdworkers, although there is a noticeable drop of over 3 points from **Setup 1** to **Setup 2**. However, it is statistically significant for LLM annotators. It is worth noting that this aspect consistently yields low mean scores compared to the other aspects, ranging from 20 to 37, as indicated by the mean bars in Figure 4.4d. Fewer turns receive a perfect

Table 4.2: Spearman's r correlation coefficient between the aspects and expert user satisfaction ratings for both the crowdworkers and LLM. * indicate non-significant values ($p \leq 0.05$).

	Crowdworkers		LLM	
Aspects	**Setup 1**	**Setup 2**	**Setup 1**	**Setup 2**
Relevance	0.56	0.51	0.63	0.52
Usefulness	0.55	0.66	0.45	0.41
Interestingness	0.31*	0.39	0.27*	0.21*
Explanation quality	0.44	0.42	0. 54	0.50

score of 100 in the aggregated scores for **Setup 2**; see Figure 4.3d. Both annotator groups agree in assigning lower scores for this aspect, highlighting the lack of recommendation explanations in the dataset. An analysis of annotator justifications reveals varying expectations regarding system explainability. While some workers expect the system to provide explanations related to aspects like the movie's cast or director, others focus on different facets. This subjectivity in worker bias and expectations regarding system explanations contributes to the variation in scores for explanation quality.

Humans vs. LLMs. Overall, both groups agree on assessing relevance, but differences emerge when considering follow-up utterances, influencing relevance scores for LLM. LLMs consistently assign lower usefulness scores than human annotators, indicating the challenge of defining usefulness when follow-up utterances reveal complex user needs. Unlike humans, LLMs do not personalize the system's usefulness to the user. This highlights the importance of including follow-up utterances for more accurate evaluation labels that reflect the user's perspective. Both groups agree on the lack of explanations from the system.

Agreement with expert ratings. Here, we examine how well human and LLM labels align with expert ratings. We collect expert ratings on the user satisfaction aspect to investigate the correlation of the fine-grained aspects to overall user satisfaction following Siro et al. [201]. Using the same setup, we collect the expert ratings from two experts. Since our initial annotation was on an S100 scale we transformed the labels to S3 scale (1–3) for all aspects [88] and then calculated the Spearman's r between each aspect and the expert rating. We report our results in Tab. 4.2.

In **Setup 1**, relevance displays a moderate positive correlation of 0.56 (crowdworkers) and 0.63 (LLM) with expert ratings, indicating a similar alignment with expert satisfaction in the absence of follow-up utterances. Usefulness shows stronger correlations, with 0.55 for crowdworkers and 0.45 for LLM, suggesting that usefulness judgments are significantly influenced by the absence of follow-up utterances in this setup. Interestingness exhibits weaker correlations in both groups, suggesting potential challenges or subjectivity in assessing this aspect. Explanation quality demonstrates moderate correlations (0.47 for crowdworkers and 0.54 for LLM), indicating moderate alignment with expert satisfaction ratings.

In **Setup 2**, relevance maintains positive correlations, 0.51 (crowdworkers) and 0.52 (LLM). Usefulness shows notably stronger correlations, 0.66 for crowdworkers and 0.41 for LLM, indicating that crowdworkers assign scores closely aligned with the user feedback. Interestingness continues to exhibit weaker correlations (0.39 for crowdworkers and 0.21 for LLM). Explanation quality, while still aligned with expert ratings, has slightly weaker correlations (0.44 for crowdworkers and 0.50 for LLM) compared to **Setup 1**.

Humans vs. LLMs. Correlations with expert ratings highlight that relevance and usefulness assessments generally have a stronger alignment with expert user satisfaction ratings across setups and annotator types. However, interestingness shows weaker correlations, indicating potential challenges in assessing this aspect consistently and objectively [200]. In general, we note that humans perform well in assessing user experience measures such as usefulness and interestingness while LLMs performs well in assessing utility measures such as relevance and explanation quality.

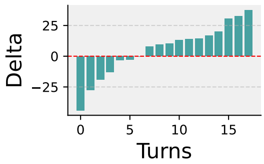
4.6 Significance of User Feedback

In this section, we examine the impact of user feedback from follow-up utterances on reducing annotator variability in their evaluation labels as part of our analysis to answer **RQ3.2**. To assess agreement, we calculate the standard deviation of workers' scores for each turn and categorize the data into two groups: **Group 1** (scores below the median standard deviation, indicating high agreement) and **Group 2** (scores above the median, indicating low agreement). We find that *interestingness* and *explanation quality* consistently exhibit higher agreement among annotators when the system response is uninteresting or lacks explanation. However, there is no clear agreement pattern among workers for *relevance* and *usefulness*.

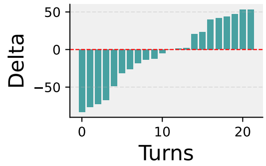
We compare turns in Group 2 from **Setup 1** with Group 1 from **Setup 2**, where Group 2 initially exhibits high variability in evaluation labels, but shows increased agreement in **Setup 2** due to the presence of the user's follow-up utterance. Specifically, there are 18 turns for relevance, 22 for usefulness, 25 for interestingness, and 19 for explanation quality in this analysis (Group 2 initially consisted of 48 to 50 turns). Overall, at least 30% of the turns demonstrate improved worker agreement in **Setup 2**. To quantify score differences between Group 1 and Group 2, we calculate their delta and present the results in Figure 4.5. We observe significant score differences for the same instances under different conditions. Relevance and interestingness have more turns rated highly in Group 1 (positive delta scores) (Figure 4.5a and 4.5c) while usefulness and explanation quality have turns rated low in Group 1 (Figure 4.5b and 4.5d).

Manual analysis. Our manual analysis primarily focused on the usefulness aspect due to its substantial impact, with the highest mean delta difference (35) compared to other aspects (interestingness: 26, explanation quality: 22, relevance: 15). We analyze 22 turns to identify instances where the user's follow-up utterances notably enhance worker agreement, shedding light on cases where the presence of the user's next utterance significantly improves consensus.

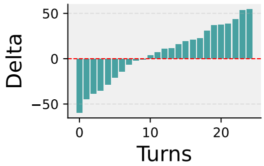
This analysis identifies specific scenarios where user feedback plays a pivotal role, such as addressing ambiguous requests by providing clarity, making generic requests more specific and actionable, simplifying complex requests, and compensating for annotators' lack of domain knowledge. In these scenarios, user feedback consistently improves the overall quality and consistency of the annotation process, highlighting its



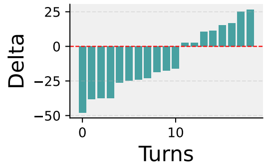
(a) Relevance



(b) Usefulness



(c) Interestingness



(d) Explanation quality

Figure 4.5: Difference in scores assigned to dialogues turns for four aspects in Group 1 with low variability vs. dialogues in Group 2 with high variability between the worker scores from the mean rating.

significance in enhancing system evaluations.[3]

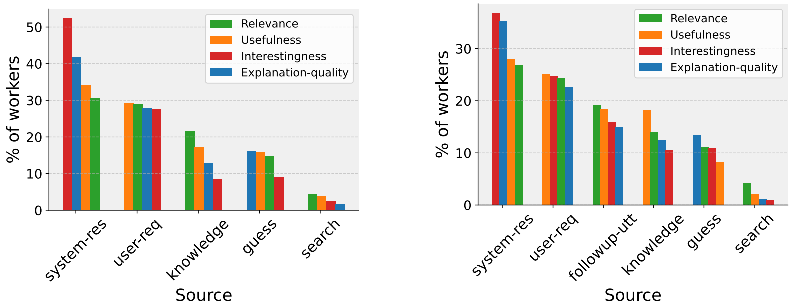
Apart from resolving uncertainty in user requests, follow-up utterances are crucial when annotators encounter unfamiliar topics. An analysis of annotators' justifications reveals that when annotators lack prior knowledge, the user's knowledge about the recommended item, coupled with explicit feedback, bridges the knowledge gap, resulting in more precise evaluations of the system's performance, even when annotators lack subject matter expertise.

4.7 Sources and Bias

In this section, we answer **RQ3.3**. To understand the basis of workers' assessments and their choices when assigning evaluation labels, we conducted a survey where we asked workers to indicate the sources of information they relied on when making judgments. Both setups offered the same options for information sources, except for follow-up utterance, which was available only in **Setup 2**. The available sources included *personal knowledge, searched online, guessed, user request,* and *system response*. Additionally, we performed a manual analysis of the workers' justifications to evaluate any potential biases introduced by their chosen information sources.

Sources. Figure 4.6 shows the percentage of workers who relied on each information source during evaluation. The x-axis represents the information sources, while the y-axis indicates the percentage of workers relying on each source. Across both setups, it is evident that workers predominantly depend on information within the dialogue itself, specifically the user's request and the system's response, for their assessments. Interestingly, the system's response primarily influenced the evaluation of interestingness and explanation quality. While explanation quality considered both the user request and system response, we observe a notably higher reliance on the system's response

[3]A detailed example of a complex user request can be found in Appendix 4.B.



(a) **Setup 1** (b) **Setup 2**

Figure 4.6: Distribution of sources workers relied on to make their judgments for the four aspects in the two setups.

than the user request. On the contrary, for relevance and usefulness assessment, we observe that workers mostly rely on the user request to ensure the system meets the user's need. However, there is a marginal difference between system response and user request for these two aspects, showing that the two are equally important during assessment. Without the user's follow-up utterance, some workers make an educated guess on the relevance and usefulness of the system response.

In **Setup 2** (Figure 4.6b), we note a drop in the percentage of workers relying on the system response during evaluation, showing that the follow-up utterance introduces another dynamic to be considered by the workers during the assessment. Usefulness, which measures how well the system response meets the user's needs, has a high percentage of workers relying on the user's follow-up utterance. Unexpectedly, we note a high number of workers relied on personal knowledge to gauge the usefulness of the system response, showing that they introduced personal bias in assessing this aspect.

A few workers utilize online sources, primarily when assessing relevance and usefulness, implying that workers without domain knowledge leverage online information for more accurate assessments. Interestingly, there is a decrease in the percentage of workers using online sources in **Setup 2**, specifically for evaluating usefulness. This suggests that the introduction of the user's follow-up utterance in **Setup 2** acts as an additional information source, assisting workers lacking domain knowledge in assessing the system response's usefulness.

Biases. Several studies highlight the influence of biases on crowdworkers' judgments e.g., [69, 101, 190]. In our work, we specifically explore how the sources of information outlined in Section 4.7 introduce biases into crowdsourced labels. Figure 4.6 illustrates a reliance on online sources for assessment, which, while potentially augmenting workers' domain knowledge, can introduce specific biases, such as popularity bias [45]. From the analysis of workers' justifications, we observe instances in assessing usefulness where some workers forego the user's feedback on the system's response and rely on online movie reviews. Justifications like "The movie seems to be liked by many so it is useful," and "The movie is not rated highly" are observed, indicating that these external sources could bias workers, leading to ratings that may not accurately reflect the user feedback.

A notable percentage of workers rely on the user's request when assessing interestingness. However, interestingness assessments should primarily be based on the system's response. Therefore, we note several workers get biased by the relevance of the recommended item to the user request during their assessments. To mitigate this bias, it may be prudent to restrict access to the user request when evaluating aspects like interestingness, where the focus is on assessing the system independently of the user's input.

In comparison to other aspects, we notice that workers rely on their knowledge to evaluate the usefulness of the system response (see Figure 4.6), introducing personal preference bias despite the explicit tie to the user's needs. Also, workers display a bias towards rating longer system responses highly for explanation quality.

4.8 Discussion

In this study, we addressed the question of how the inclusion of implicit and explicit user feedback from the user's follow-up utterance influences the evaluation labels from crowdworkers and LLMs. Our analysis revealed intriguing patterns across various aspects, providing valuable insights into the impact of user feedback on the quality of assessments.

Relevance. We considered two experimental setups, one without the user's follow-up utterance and one with. In both setups, crowdworkers and LLM-based annotators largely concur when evaluating relevance. However, subtle differences emerge with the inclusion of follow-up utterances, particularly for LLMs which tend to assign lower scores than humans. Though there is no significant mean difference in **Setup 1** for both annotator groups, we note that LLM scores show a higher correlation to expert user satisfaction ratings than humans. Crowdworkers relied more on the system response, user request, and their prior knowledge to gauge the relevance of the recommended item. With lots of candidate movies to recommend, crowdworkers may lack knowledge of some of these movies to assess their relevance, which results in making an educated guess. However, compared to humans LLMs are rich in internal knowledge on these movies, thus improving their relevance assessment.

Usefulness. The usefulness ratings show a distinctive contrast between the two annotator groups. Human annotators displayed strong correlations with user satisfaction in **Setup 2**, suggesting their ability to personalize the system's usefulness to individual users. In contrast, LLMs consistently assigned lower usefulness scores in both setups, highlighting the challenge of assessing usefulness when follow-up utterances reveal conflicting user needs from their initial request. This shows that when user feedback conflicts with a model's internal knowledge it leads to inconsistency in the ratings.

Interestingness. The aspect of interestingness presented unique challenges. Both crowdworkers and LLMs exhibited lower correlations with user satisfaction ratings, indicating that both annotator groups struggled to capture the user's subjective perception of interestingness. The presence of user feedback had a limited impact on improving assessments in this aspect. This is also observed with the low Kappa and ICC scores in Tab. 4.1. Nonetheless, utterances such as "that's interesting," "sounds good," and "you are funny" lead to an increase in annotator agreement and correlation with user satisfaction ratings (**Setup 2**), emphasizing the significance of user feedback in improving system evaluation.

Explanation quality. Both annotator groups concur on the absence of explanations provided by the system. This shared observation underscores a significant limitation in current CRS, as both human evaluators and LLMs noted the lack of explanatory content in system responses. The LLM shows less sensitivity to user feedback with high correlating scores to overall user satisfaction compared to humans. Humans' ratings are affected by their personal expectations of the system's explainability, which is not evident in the LLM scores. This shows that LLMs can maintain objectivity when assessing system performance.

4.9 Conclusion

In this chapter, we addressed **RQ3:** *How does incorporating user feedback through follow-up utterances affect evaluation judgments by humans and LLMs, and what does this reveal about their respective strengths as annotators?* We carry out experiments in two setups: with and without user's next utterance across two annotator pools – humans and LLMs. Both annotators assess four dialogue aspects in the two setups: *relevance, interestingness, usefulness* and *explanation quality.*

In general, there is a distinct difference in ratings assigned by both annotator groups in the two setups, indicating that user feedback does influence system evaluation. Humans are more susceptible to user feedback in usefulness and interestingness compared to LLMs in interestingness. User feedback leads to personalized usefulness assessment by workers and improves worker agreement when uncertainty arises in the user's request. The lack of adaptability to user feedback by the LLM in assessing usefulness, suggests that LLMs may require additional mechanisms such as prompt engineering to enhance user-centric evaluations by LLMs. Therefore, it is important to assign annotation tasks to LLMs based on the nature of the task, leveraging their strengths in objective assessments like relevance annotation while complementing with human assessors for tasks demanding subjective evaluation or sensitivity to user preferences and feedback. Combining human annotators and LLMs can lead to better system evaluations by leveraging the unique strengths of each type of annotator for specific evaluation tasks.

However, user feedback can sometimes lead to assessments that do not align with overall user satisfaction, resulting in lower correlation scores as observed for the relevance aspect in **Setup 2**. It is important to note that this study employed a single LLM for annotation, and results may vary with different LLMs. Additionally, potential biases in the crowdworker pool and the LLM's training data could influence the findings. For future work, we will conduct further research to validate our findings across diverse conversational systems.

Our first theme in this thesis, *Understanding the Evaluation of Task-based conversational search (CS) Systems* (Chapters 2–4), established frameworks for reliable evaluation of conversational interactions. Through studies on user satisfaction patterns, dialogue context effects, and the role of user feedback, we investigated different aspects of evaluation methodology. Building on these insights about evaluation methodology, our second theme, *Advancing Clarification in CS* (Chapters 5–6), investigates how to improve clarification in CS systems. We examine two key approaches: using LLMs to automate clarifying question generation and evaluation and enhancing these questions with visual elements to improve user understanding.

4.A Prompts

In this section, we provide the prompts used during the annotation task. Below we show prompts used in the annotation of the relevance aspect in **Setup 1** and **Setup 2** respectively. All prompts for the other aspects follow the same structure except for aspect definitions and examples.

```
Task Description

As a Language Model, you are tasked with annotating a series
of interactions between an intelligent system similar to Siri
or Alexa, and a user who is seeking movie recommendations.
Each turn in the conversation consists of two utterances: one
from the user and one from the system. Your task is to examine
the second utterance in each turn, which is the system's
response.

What you need to do:
- Review each turn, particularly focusing on the second
utterance by the system.
- Rate the system's response on the following criteria:
  - Relevance: How well the suggested item aligns with the
  user's request in the current conversation context. (Scale:
  0-100)

Examples:
- Highly Relevant:
  - User: "What's a good sci-fi movie?"
  - System: "I'd suggest watching 'Blade Runner 2049'. It's a
  visually stunning sci-fi film with a captivating storyline!"
  - relevance_score: 90
  - justification: The system's response directly addresses the
  user's request, providing a specific movie title and a brief
  description.

- Moderately Relevant:
  - User: "Recommend a romantic movie with Keanu Reeves."
  - System: "What about 'Constantine'? It's a supernatural
  thriller starring Keanu Reeves."
  - relevance_score: 50
  - justification: The response partially fulfills the user's
  request by suggesting a movie with Keanu Reeves, but it's not
  in the requested genre (romantic).

- Not Relevant:
  - User: "Can you recommend a romantic movie?"
  - System: "Have you watched 'Sherlock Holmes'? It's an
  engaging mystery film."
```

```
  - relevance_score: 0
  - justification: The system's response doesn't align with the
  user's request for a romantic movie.

Please rate the system response on relevance on a scale from 0
to 100, where 0 is 'Not Relevant' and 100 is 'Highly
Relevant', providing a brief justification for your rating.

Output: relevance_score:score, justification:justification.
```

```
Task Description:
As a Language Model, you are tasked with annotating a series
of interactions between an intelligent system similar to Siri
or Alexa, and a user who is seeking movie recommendations.
Each turn in the conversation consists of three utterances:
two from the user and one from the system. Your task is to
examine the second utterance in each turn, which is the
system's response.

What you need to do:
- Review each turn, particularly focusing on the second
utterance by the system.
- Rate the system's response on the following criteria:
  - Relevance: How well the suggested item aligns with the
  user's request in the current conversation context. (Scale:
  1-3)

Examples:
- Highly Relevant:
  - User: "What's a good sci-fi movie?"
  - System: ``I'd suggest watching 'Blade Runner 2049'. It's a
  visually stunning sci-fi film with a captivating storyline!''
  - User: "Thank you! 'Blade Runner 2049' sounds exactly like
  what I was looking for. I'll definitely watch it."
  - relevance_score: 90
  - justification: The system's response is highly relevant to
  the user's request for a sci-fi movie recommendation. It
  provides a specific movie title along with a brief
  description of why it's a good choice, which directly
  addresses the user's query.

- Moderately Relevant:
  - User: "Recommend a romantic movie with Keanu Reeves."
  - System: "What about 'Constantine'? It's a supernatural
  thriller starring Keanu Reeves."
  - User: "Thanks for the recommendation, but I was actually
  looking for a romantic movie. Do you have any other
  suggestions?"
  - relevance_score: 50
  - justification: The response provides a movie starring Keanu
```

```
  Reeves, which partially fulfills the user’s request. However,
  it suggests a supernatural thriller, which is not in the
  requested genre (romantic).

- Not Relevant:
  - User: "Can you recommend a romantic movie?"
  - System: "Have you watched ’Sherlock Holmes’? It’s an
  engaging mystery film."
  - User: "Actually, I’m specifically looking for a romantic
  movie. Can you recommend one?"
  - relevance_score: 0
  - justification: The system’s response is not relevant to the
  user’s request for a romantic movie recommendation. It talks
  about the system’s personal hobby and doesn’t address the
  user’s query.

Please rate the system response on relevance on a scale from 0
to 100, where 0 is ’Not Relevant’ and 100 is ’Highly
Relevant’, providing a brief justification for your rating.

Output: relevance_score:score, justification:justification.
```

4.B Further Analysis

4.B.1 Qualitative analysis

Our manual analysis primarily focused on the usefulness aspect due to its substantial impact, with the highest mean delta difference (35) compared to other aspects (interestingness: 26, explanation quality: 22, relevance: 15). We analyze 22 turns to identify instances where the user’s follow-up utterances notably enhance worker agreement, shedding light on cases where the presence of the user’s next utterance significantly improves consensus.

This analysis identifies specific scenarios where user feedback plays a pivotal role, such as addressing ambiguous requests by providing clarity, making generic requests more specific and actionable, simplifying complex requests, and compensating for annotators’ lack of domain knowledge. In these scenarios, user feedback consistently improves the overall quality and consistency of the annotation process, highlighting its significance in enhancing system evaluations.

For instance, consider a user who requested a movie recommendation, noting their fondness for *‘A Nightmare on Elm Street (1984)’*. This request could imply a preference for several aspects: the classic horror genre, the directorial style of Wes Craven, or the specific era of the 1980s. Initially, the system recommends a popular horror-comedy from 2017, which, while relevant in terms of genre, might not align precisely with the nuanced preferences implied in the user’s request. The initial ratings from the crowdworkers reflect this broad interpretation, with Worker A giving a score of 70/100 based on the assumption of a preference for any modern horror film, Worker B scoring 55/100 considering the genre match but not the era, and Worker C giving

60/100, focusing on the horror aspect but overlooking the director and the time period. However, the user's follow-up utterance, stating a specific enjoyment for Wes Craven's horror style and a preference for 1980s movies, brings new clarity. With this additional information, the crowdworkers in **Setup 2** assess the same recommendation against these more specific criteria. Worker D now scores 40/100, recognizing the mismatch in both director and era, Worker E assigns 30/100, acknowledging the recommendation's lack of alignment with the specific director and the time period, and Worker F scores 35/100, understanding that while the genre is somewhat aligned, the specific preferences for Wes Craven's style and the 1980s era are not met.

This significant change in ratings, from relatively high to low in **Setup 2**, illustrates that without user feedback, crowdworkers may overestimate the usefulness of a system's response based on a broader, less nuanced interpretation of the request. However, when provided with more detailed preferences from the user, their ratings become more aligned and tend to be lower, reflecting a more accurate assessment of how well the system's response meets the user's refined criteria. This highlights the importance of the user's next utterance in accurately gauging the usefulness of system responses, especially in the context of complex requests.

5 Generating and Evaluating Clarifying Questions with LLMs

Clarifying questions, which we study in theme 2, play a crucial role in conversational search systems, helping systems resolve ambiguity when user requests are unclear or incomplete. While their effectiveness has been demonstrated in improving search performance, generating and evaluating these questions at scale remains a significant challenge. In this chapter, we investigate how LLMs can address these scalability challenges through the following research question:

RQ4: How effectively can large language models generate and evaluate clarifying questions for conversational search systems?

We propose AGENT-CQ, an end-to-end LLM-based framework that addresses the generation and evaluation of clarifying questions. We also propose CrowdLLM, a method that emulates crowdsourced judgments by combining outputs from multiple LLM instances.

5.1 Introduction

Conversational search (CS) systems have gained significant attention in recent years, offering users a more natural and interactive way to find information than single-shot search interactions [7, 177, 242]. To resolve the ambiguity inherent in user queries, these systems may ask users clarifying questions [241]. Generating diverse and effective clarifying questions is crucial for improving query understanding and retrieval performance, which remains a challenge.

Existing methods for generating clarifying questions in CS systems rely on manual curation by experts and template-based approaches [8, 239]: human experts craft clarifying questions, relying on their ability to understand complex user intents and contextual nuances intuitively. While this method ensures high relevance and accuracy, it poses challenges for scalability in large-scale applications [60]. Moreover, human curators may not have deep knowledge about all conversation topics. In contrast,

This chapter was published as: C. Siro, Y. Yuan, M. Aliannejadi, and M. de Rijke. AGENT-CQ: Automatic Generation and Evaluation of Clarifying Questions for Conversational Search with LLMs. *Under submission*.

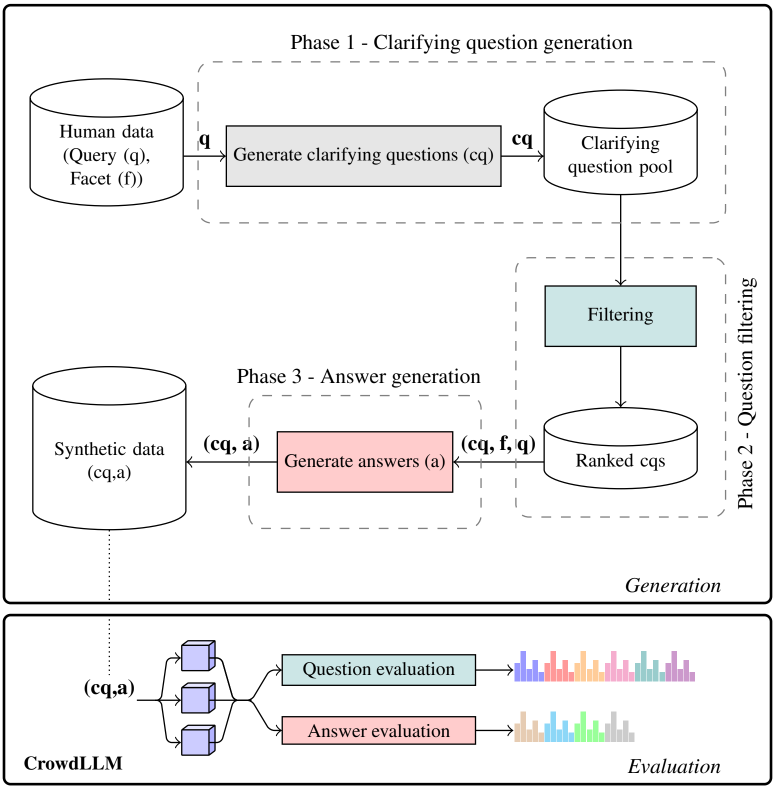


Figure 5.1: The AGENT-CQ framework for generating clarifying questions and simulating user answers (top) and evaluating generated questions and answers (bottom).

template-based methods employ pre-defined templates to automate the generation of clarifying questions, significantly enhancing scalability and efficiency. However, these methods often lack flexibility, leading to generic or less diverse questions that could hurt the overall user interaction experience [234].

Recently, large language models (LLMs) have shown capability in generating quality synthetic data for NLP and IR tasks, e.g., dialogue generation [64], document generation [17], and query generation [104], with little work to explore the generation of clarifying questions. Therefore, we ask: *Can we leverage LLMs to effectively understand user intents and generate high-quality clarifying questions in information-seeking dialogues, thereby enhancing downstream document retrieval performance?*

We propose **AGENT-CQ**, an end-to-end LLM-based framework for generating and evaluating clarifying questions. AGENT-CQ has two stages: *generation* (top) and *evaluation* (bottom); see Figure 5.1. The generation stage has three main phases: question generation (Phase 1), filtering (Phase 2), and answer generation (Phase 3). In Phase 1, we employ LLMs to generate and compare clarifying questions. In Phase 2, we filter out generated questions that do not meet certain quality criteria. In Phase 3, we generate answers to the final set of questions by simulating system-user interactions.

CrowdLLM is the second stage of AGENT-CQ; it is designed to assess our generation stage's effectiveness automatically. It simulates a crowd of workers by employing three instances of an LLM to evaluate the generated questions and simulated answers, mimicking diversified human judgments. CrowdLLM conducts a multidimensional evaluation, assessing the quality of questions on seven metrics and answers on four metrics. Given the challenges of evaluating clarifying questions – where human judgments can be time-consuming, costly, and subjective – CrowdLLM offers a scalable and consistent alternative. We conduct an extensive human evaluation to ensure the quality and robustness of CrowLLM evaluation labels.

Our experiments on the ClariQ dataset [10] show that (i) CrowdLLM is highly effective at evaluating clarifying questions and answers, exhibiting strong inter-rater agreement across most dimensions with a high correlation with human expert ratings; (ii) GPT–Temp, our temperature-variation method, consistently outperforms other approaches, in terms of clarity, relevance, usefulness, and overall quality. Facet-based approaches (e.g., GPT-Facet) demonstrate high specificity but increased question complexity. (iii) LLM-generated clarifying questions, particularly those from GPT-Temp, enhance retrieval effectiveness across both BM25 and BERT models, consistently achieving higher NDCG scores than human-generated questions. This indicates that LLMs are capable of generating effective clarifying questions that help enhance retrieval performance.

Our **main contributions** in this chapter are:

(C1) AGENT-CQ: A scalable methodology for generating and evaluating clarifying questions with LLMs.

(C2) As part of AGENT-CQ, we also share a reliable evaluation framework (CrowdLLM) that balances scalability and accuracy in question assessment.

(C3) A comparative analysis of LLM architectures for generating questions and simulating user responses, and

(C4) LLM-generated clarifying questions, combined with simulated user answers, can improve user intent understanding and enhance retrieval performance.

5.2 Related Work

5.2.1 LLM-based conversational search systems.

CS system is an interactive paradigm where users engage in a dialogue with a search system [177]. In conversational search systems, LLMs enhance the search experience through query understanding, retrieving relevant documents, and generating clear responses [58, 110, 256]. LLMs have been adopted to simulate users and their interactions with the system, reducing the need for human resources [47, 121, 160, 183, 227]. E.g., Abbasiantaeb et al. [1] simulate teacher-student interactions in a conversational setting. Sekulic et al. [197] focus on evaluating query clarification via an LLM-based user simulator.

Our work focuses on query clarification, i.e., the process of refining or elaborating on a user's initial search query or question to better understand their intent [7, 239]. Prior work on the role of clarifying questions in conversational search recognizes their potential to enhance search quality [9, 211, 238] and the user experience [195, 241]. How LLMs can benefit the task remains underexplored. Our work evaluates LLMs on query clarification at the clarifying question and user response levels, assessing their ability to generate effective questions and responses.

5.2.2 Evaluation of generated content.

Traditional automated metrics for evaluating generated content like BLEU [170] and ROUGE [136] often correlate poorly with human judgments for open-ended text generation [139]. Human evaluation is the gold standard but time-consuming and costly to scale [124]. Newer metrics (USR, Mehri and Eskénazi, 2020; BLEURT, Sellam et al., 2020; BERTScore, Zhang et al., 2020) use pre-trained language models to improve correlation with human judgments. Multi-dimensional evaluation frameworks have also emerged [65, 86]. Recent research explores using LLMs as evaluators for natural language tasks [127, 143]. Zheng et al. [252] explore "LLM-as-a-judge" for chat assistants. Lin and Chen [137] introduce LLM-Eval, a multi-dimensional evaluation method for open-domain conversations.

CrowdLLM shares similarities with recent LLM-based evaluation techniques [143, 252] and multi-dimensional frameworks [65]. However, it uniquely employs multiple LLM instances to simulate diverse evaluators, addressing scalability issues of human evaluation. CrowdLLM also incorporates a second tier where humans assess the reliability of LLM-generated evaluations, combining automated efficiency with human judgment accuracy.

5.3 AGENT-CQ Framework

We introduce the two key stages of AGENT-CQ: a framework for generating clarifying questions and a framework for evaluation.

5.3.1 AGENT-CQ: Generation framework

AGENT-CQ's generation framework generates and scores clarifying questions using state-of-the-art LLMs in an end-to-end manner. The framework has three main phases; see Figure 5.1 (top).

Question generation (Phase 1). Let $Q = \{q_1, q_2, \ldots, q_n\}$ be a set of n initial user queries. For each $q_i \in Q$, we aim to generate a set of clarifying questions $C_i = \{c_{i1}, c_{i2}, \ldots, c_{im}\}$, where m is the number of clarifying questions per set. We define a question generation function:

$$G : q_i \to \{C_i^1, C_i^2, \ldots, C_i^k\} \tag{5.1}$$

that generates k sets of clarifying questions for query q_i. We explore two prompt-based approaches (i.e., $p = 2$).

1. **Facet-based approach.** We adopt the approach of diverse query interpretation based on Aliannejadi et al. [10], aiming to generate clarifying questions that address multiple interpretations of a given query. We introduce the facet-based method [196, 239]. Here, an LLM takes a query as input, then generates facets as

a way of exploring the topic of the query, and finally generates a clarifying from a query-facet pair. Algorithm 5.1 (Appendix 5.A.1) details the implementation of this approach.

2. **Temperature-variation-based approach.** This method generates diverse clarifying questions by systematically adjusting an LLM's temperature parameter. Starting from a low temperature and incrementing it over multiple iterations, it produces question sets with progressively increasing diversity. This approach implicitly explores various query facets, potentially uncovering different clarifications without explicit facet modeling. For a detailed description, see Algorithm 5.2 (Appendix 5.A.1).

Question filtering (Phase 2). In this phase, we take two major characteristics of questions to reject low-quality questions. Our preliminary experiments showed that LLMs can sometimes generate questions that are not clarifying questions or on the same topic as the user query. We define a function S to filter out questions based on relevance and clarification potential:

$$S(q_i, C_i^j) = \alpha \cdot R(q_i, C_i^j) + (1 - \alpha) \cdot L(C_i^j), \tag{5.2}$$

where $R(q_i, C_i^j)$ is the relevance score; $L(C_i^j)$ is the clarification score, evaluating the questions' potential to clarify user intent; and α is a weighting parameter. We keep the top 10 ranked questions for each query in the collection for each LLM and experimental setup.

User response simulation (Phase 3). In Phase 3 of AGENT-CQ's generation stage, we simulate user responses to the ranked clarifying questions from Phase 2. Recent work has demonstrated the efficacy of simulated users as cost-effective proxies for real users in conversational systems [1, 236, 251]. Using this insight, we employ an approach that takes LLM as a simulator for generating answers to system-generated clarifying questions. We introduce a *parameterized-user simulation* approach, inspired by Sekulic et al. [197]. This method incorporates user characteristics (U) in the simulation, to generate diverse and realistic answers (for details see Algorithm 5.3 Appendix 5.A.2). Our parameterized function is defined as:

$$a_{ij} = A_p(q_i, u_i, c_{ij}, U), \tag{5.3}$$

where q_i is the original query, u_i is the user information need, c_{ij} is the clarifying question, and U is the set of user characteristics. A_p extends the basic non-parameterized method[1] by incorporating user characteristics U, primarily verbosity, which controls the response length, detail and revealment probability used to determine the likelihood of disclosing the true user information need.

5.3.2 AGENT-CQ: Evaluation framework

Next, we detail CrowdLLM, AGENT-CQ's evaluation framework. CrowdLLM is a multi-LLM and multi-dimensional framework evaluating the generated questions and

[1] Defined as: $a_{ij} = A_N p(q_i, u_i, c_{ij})$

Table 5.1: CrowdLLM ICC and weighted κ (W-κ) agreement scores for different prompting strategies across models including human-generated clarifying questions (H-Gen). Question-C denotes question complexity.

Aspects	**GPT-Baseline**		**GPT-Facet**		**GPT-Temp**		**Llama 3.1**		**H-Gen**	
	ICC	W-κ	ICC	W-κ	ICC	W-κ	ICC	W-κ	ICC	W-κ
Clarification	0.96	0.87	0.95	0.85	0.85	0.72	0.97	0.89	0.97	0.89
Clarity	0.90	0.79	0.80	0.67	0.81	0.77	0.94	0.84	0.95	0.87
On-topic	0.93	0.81	0.87	0.78	0.87	0.82	0.93	0.86	0.96	0.88
Question-C	0.86	0.76	0.93	0.84	0.87	0.80	0.94	0.85	0.78	0.73
Specificity	0.92	0.80	0.83	0.66	0.80	0.66	0.92	0.79	0.96	0.87
Usefulness	0.94	0.84	0.93	0.82	0.92	0.78	0.97	0.89	0.97	0.90
Overall-quality	0.92	0.81	0.88	0.82	0.75	0.68	0.94	0.85	0.95	0.88

simulated responses leveraging the scalability of LLMs to simulate a crowd of evaluators, with validation from human experts.

Multi-LLM evaluation. CrowdLLM employs an LLM-as-a-judge approach [252], in an ensemble of LLM instances with varying temperature settings. We hypothesize that with varying temperatures different LLM instances will bring different angles to the evaluation. This design simulates the setup used with crowdsourced workers, crucial for comprehensive assessment in NLP tasks. Each LLM instance evaluates questions and answers on multiple aspects using a 10-point scale for questions and pairwise comparison between LLM and human answers. To validate CrowdLLM's performance, we incorporate human expert assessment.

Evaluation metrics. Evaluation in CrowdLLM is based on distinct sets of metrics for clarifying questions and simulated answers, drawn from prior work on conversational information seeking and general conversational systems [10, 181, 201, 203, 239]. For clarifying questions, we assess clarification potential, on-topic relevance, specificity, usefulness, clarity, and question complexity. Simulated answers are evaluated on relevance, usefulness, naturalness, and overall quality.[2]

Details about our experimental setup, implementation details and prompts used are included in Appendix 5.A.4 and 5.B.

5.4 Reliability of CrowdLLM

In this section, we study the reliability of our evaluation framework (CrowdLLM) from multiple angles.

[2]Definition of the aspects in Appendix 5.A.3

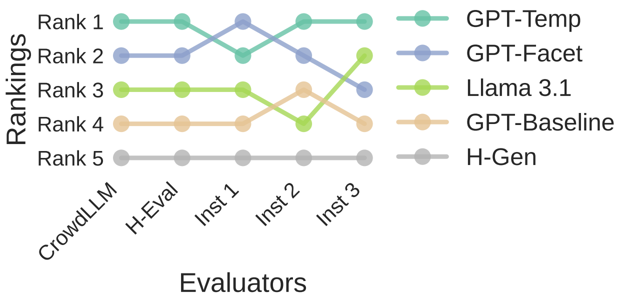


Figure 5.2: Rankings of the question sets from different systems by different evaluators.

5.4.1 Clarifying questions

In Table 5.1, we report the inter-annotator agreement (IAA) among the LLM instances (i.e., GPT-4o) using intraclass correlation coefficient (ICC) and weighted κ [54]. We further assess the agreement of CrowdLLM with human evaluators to inform on the quality of the evaluations. Because of the large number of generated questions by each model and the associated costs, our human evaluators assess a sample of 200 questions from each model and rank them based on preference. For each user information need and initial request, human evaluators see the clarifying questions generated by each system and rank them from most helpful (Rank 1) to least helpful (Rank 5) and give a justification for the least helpful clarifying question. For CrowdLLM evaluations we rank the questions based on overall quality score using Tukey's HSD post-hoc test for pairwise comparisons.

Internal agreement. *CrowdLLM demonstrates moderately consistent performance across most aspects of question quality.* Overall quality shows notable variability ($ICC = 0.75$, $\kappa = 0.68$), lower than other aspects. This suggests that temperature variation in LLM instances leads to diverse perspectives on holistic question evaluation. Specificity ($ICC = 0.80$, $\kappa = 0.66$) and clarity ($ICC = 0.81$, $\kappa = 0.77$) show lower inter-rater agreement. Higher-temperature instances interpret abstract details as specific and rate subtly ambiguous questions as clear, while lower-temperature instances focus on explicit information and apply stricter clarity standards. Despite these variations, high agreement across most aspects indicates CrowdLLM provides a robust method for evaluating clarifying question quality.

External agreement. *CrowdLLM evaluations strongly align with human evaluation results (H-Eval), confirming its effectiveness (Figure 5.2).* Both approaches consistently rank the GPT-Temp question set as most helpful and human-generated questions (H-Gen) as least helpful, with H-Gen receiving an average rank of 4 out of 5 in human evaluation (lower is better). CrowdLLM instances show ranking variations: Instance 1 prioritizes GPT-Facet > GPT-Temp > Llama 3.1; Instance 2 mirrors aggregate rankings with GPT-Baseline > Llama 3.1; Instance 3 ranks Llama 3.1 second, above GPT-Facet and GPT-Baseline. These inconsistencies highlight the value of using multiple instances to produce robust, human-aligned assessments. Despite the mid-rank differences, all evaluation approaches consistently rank GPT-Temp or GPT-Facet first and

Table 5.2: Inter-rater reliability measures for CrowdLLM and Human evaluators in terms of Fleiss' κ, annotator agreement percentage, and Human-CrowdLLM agreement (H-Ag.).

	Fleiss' κ	**% Ag.**	**% H-Ag.**
Naturalness	0.81	89%	53%
Relevance	0.68	86%	73%
Usefulness	0.62	73%	68%
Overall-quality	0.71	79%	75%

H-Gen last.

5.4.2 Simulated answers

We assessed CrowdLLM's reliability in evaluating generated answers, comparing it with human evaluators (H-Eval). Evaluators judged which answer in presented pairs was better for naturalness, relevance, usefulness, and overall quality, with an option to rate them equally. Table 5.2 presents inter-rater reliability results for this comparative evaluation.

Internal agreement. *CrowdLLM shows high internal consistency.* Naturalness has highest agreement ($\kappa = 0.81$, 89% agreement), indicating near-perfect consensus. Relevance ($\kappa = 0.68$, 86% Ag.) and overall quality ($\kappa = 0.71$, 79% Ag.) show substantial agreement. Usefulness, while substantial, has the lowest scores ($\kappa = 0.62$, 73% Ag.).

External agreement. *CrowdLLM and human evaluators align strongly in overall quality (75%) and relevance (73%).* Usefulness agreement was moderate (68%). In Chapter 4, [202] we showed that humans excel at usefulness assessment compared to LLMs, supporting our findings in this chapter. Naturalness had the lowest agreement (53%), despite high internal consistency in both methods. This may be due to LLMs' bias towards their generated answers, while humans better discern between human and LLM-generated responses [189].

5.4.3 Effectiveness of evaluation aspects

Which aspects most influence perceived quality? Table 5.3 shows correlations between question aspects and overall quality. Usefulness has the strongest association ($\tau = 0.80$, $\rho = 0.90$), followed by clarification ($\tau = 0.76$, $\rho = 0.87$), clarity ($\tau = 0.75$, $\rho = 0.85$), and on-topic relevance ($\tau = 0.71$, $\rho = 0.81$). Specificity shows moderate to strong correlation ($\tau = 0.63$, $\rho = 0.73$); query complexity has negligible impact ($\tau = 0.07$, $\rho = 0.08$). Thus, usefulness, clarification, clarity, and topical relevance are key determinants of perceived question quality in CrowdLLM evaluation.

Figure 5.3 shows Spearman's ρ correlation coefficients between answer aspects and overall quality. Usefulness ($\rho = 0.76$) and relevance ($\rho = 0.72$) correlate strongest with overall quality, indicating their critical role in perceived answer quality. Naturalness shows a moderate correlation ($\rho = 0.50$), suggesting less impact. Strong correlations between relevance and usefulness ($\rho = 0.70$) highlight their interconnect-

Table 5.3: Kendall's τ and Spearman's ρ correlation coefficients of CrowdLLM question evaluation aspects with overall quality. Question-C denotes question complexity

Aspect	**Kendall's τ**	**Spearman's ρ**
Clarification	0.76	0.87
Clarity	0.75	0.85
On-Topic	0.71	0.81
Question-C	0.07	0.08
Specificity	0.63	0.73
Usefulness	0.80	0.90

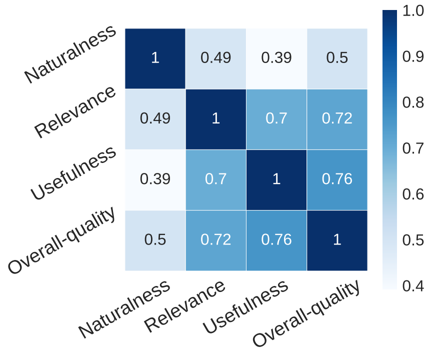


Figure 5.3: Spearman's ρ correlation coefficients between answer evaluation aspects.

edness in high-quality answers. Naturalness correlates less with relevance ($\rho = 0.49$) and usefulness ($\rho = 0.39$); it captures a distinct dimension of answer quality.

In general, *CrowdLLM evaluations reveal that usefulness and relevance are critical for both questions and answers. Questions benefit significantly from clarity and clarification, while answers balance usefulness, relevance, and naturalness.*

5.5 Evaluation of Generated Clarifying Questions

We perform a comprehensive analysis to assess the quality of data generated by AGENT-CQ.

5.5.1 Clarifying question evaluation

We conduct an analysis to explore the characteristics of the generated clarifying questions. We focus on three areas: identifying recurring question patterns [31], categorizing questions based on their intent [31, 239], and classifying the expected response types. We employ a hierarchical matching system to analyze linguistic patterns and key phrases. Categorization uses LLMs to capture subtle distinctions between types such as disambiguation and information seeking as shown in Table 5.D.1 (Appendix 5.D.1).

Table 5.4: Results showing the percentage distribution of question patterns generated by various models including human questions (H-Gen).

Pattern	**Llama3.1 (%)**	**GPT-Baseline (%)**	**GPT-Facet (%)**	**GPT-Temp (%)**	**H-Gen (%)**
Other	29.00	15.90	1.40	11.60	29.44
are you X	22.60	37.63	75.80	50.80	29.64
what specific	20.60	2.62	20.40	5.00	0.00
do you need/want/have	6.80	20.32	0.00	22.80	17.74
would you like	4.60	18.11	0.00	4.80	21.17
how X	2.00	0.60	1.60	0.00	0.20
are you looking for X	1.40	0.00	0.00	0.20	0.00
which specific	1.20	0.00	0.80	0.00	0.00
is there	0.40	2.82	0.00	2.80	0.81

Table 5.5: Results showing the percentage distribution of response types elicited by the generated questions from various models including human questions (H-Gen).

Response type	**Llama3.1 (%)**	**GPT-Baseline (%)**	**GPT-Facet (%)**	**GPT-Temp (%)**	**H-Gen (%)**
Multiple Choice	41.40	22.33	73.20	73.00	10.46
Open-ended	37.20	6.84	8.80	3.40	4.02
Yes/No	16.00	70.22	17.80	22.60	80.68
Factual	5.40	0.60	0.20	1.00	4.83

We classify response types into Yes/No, Multiple Choice, Open-ended, and Factual using a rule-based approach. Detailed analysis techniques in Appendix 5.D.1.

Question length and readability analysis. Table 5.7 shows that human questions are concise (9.71 words) and simple (5th-grade level). LLM outputs vary: GPT-Facet generates complex, lengthy questions (college-level, 23.53 words), Llama 3.1 generates variable-length high school-level questions, and GPT-baseline closely matches human question length but with higher complexity. There is a consistent gap in LLMs' ability to replicate the brevity and simplicity of human-written questions.

Question categories. Table 5.6 shows that all models except Llama 3.1 favor *preference identification* questions, with GPT-Facet leading at 74.00%. Llama 3.1 has a more balanced distribution between *preference identification* (47.20%) and *information gathering* (41.00%), and the highest disambiguation rate (10.20%). Human-generated questions have the highest confirmation rate (17.91%). Comparison questions are consistently low (<1.61%) across all approaches.

Question patterns and response types. Tables 5.4 and 5.5 reveal distinct question patterns and response types across human and LLM outputs. Humans show greater pattern diversity, favoring "Would you like" (21.17%) and "Do you need/want/have" (17.74%), with a strong preference for Yes/No responses (80.68%). In contrast, "Are you X" dominates GPT-Facet (75.80%) and GPT-Temp (50.80%), aligning with their preference for Multiple Choice responses (≈73%). Llama 3.1 most closely mirrors human diversity. GPT-Baseline shows unique tendencies, preferring "Do you need/want/have" (20.32%) and Yes/No responses (70.22%). Some patterns (e.g., "What specific") are almost exclusively LLM-generated, highlighting significant differences in question formulation between humans and LLMs.

Generally, *LLMs exhibit model-specific tendencies in generating clarifying questions, often diverging from human patterns w.r.t. question structure, expected response*

Table 5.6: Percentage of question categories for different models. Columns: Pref: Preference identification, Info: Information seeking, Disamb: Disambiguation, Conf: Confirmation, Comp: Comparison.

Model	Pref.	Info.	Disamb.	Conf.	Comp.
Llama 3.1	47.20	41.00	10.20	0.60	1.00
GPT-Baseline	73.64	15.29	2.41	7.04	1.61
GPT-Facet	74.00	18.00	6.20	0.60	1.20
GPT-Temp	66.80	14.40	16.40	1.60	0.80
Human	64.39	10.66	6.84	17.91	0.20

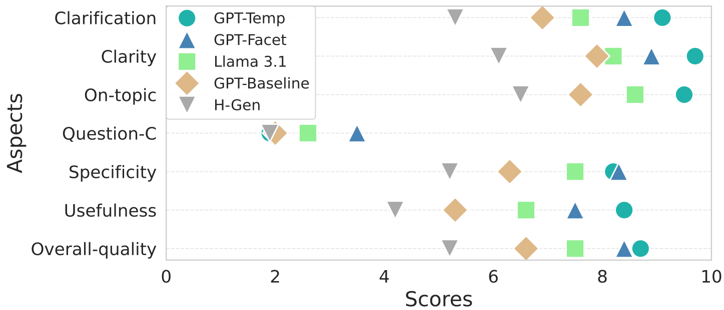


Figure 5.4: Mean question quality scores evaluated by CrowdLLM across all aspects for different models. H-Gen: human-generated questions.

Table 5.7: Question length statistics and readability scores. Length differences are statistically significant, $p \leq 0.05$; Kin: Flesch-Kincaid grade level in brackets.

Model	Mean	Std.	Flesch-Reading (Kin.)
Llama3.1	19.02	9.26	53.77 (9th)
GPT-Baseline	10.72	2.15	61.10 (7th)
GPT-Facet	23.53	4.98	35.58 (14th)
GPT-Temp	15.68	3.33	52.74 (9th)
Human	9.71	2.48	75.99 (5th)

type, category focus, length, and complexity, highlighting the challenges in replicating natural human question-asking behavior.

5.5.2 Quality analysis of clarifying questions using CrowdLLM

Next, we assess the quality of the clarifying questions on seven aspects using CrowdLLM: *clarification*, *on-topic*, *specificity*, *usefulness*, *clarity*, *query complexity*, and *overall quality*. Figure 5.4 shows mean scores across all aspects per model. We use one-way ANOVA with post-hoc Tukey's HSD for statistical significance analysis ($p \leq 0.05$).

GPT-Temp consistently outperforms other approaches across most aspects (Figure 5.4), surpassing GPT-Baseline in usefulness (mean difference $= 3.781$, $p \leq 0.001$). Facet-based models (GPT-Facet and Llama 3.1) show improvements over the baseline, with GPT-Facet often ranking second. GPT-Facet excels in specificity, significantly outperforming GPT-Baseline, as it generates specific facets before producing targeted clarifying questions. Thus, facet-based approaches enhance specificity but lead to more complex questions (GPT-Facet: 3.5) than GPT-Temp and human-generated questions (both 1.9), aligning with our Fleisch readability and Kincaid analysis.

Human-generated questions score lowest across most aspects, except for complexity. GPT-Temp significantly outperforms human questions in usefulness (8.4 vs. 4.2, $p \leq 0.001$), challenging assumptions about human expertise in question formulation. LLMs' superior performance can be attributed to their knowledge and consistent optimization for specific criteria. Recent research shows LLMs may favor their own content [144]; our human evaluators corroborate CrowdLLM's results by also ranking human questions as least helpful. This alignment between human and LLM evaluations validates our conclusions and suggests that CrowdLLM is not biased in this case.

In summary, *LLM-generated clarifying questions, particularly from GPT-Temp, outperform human-generated ones across most quality aspects. GPT-Temp's strong performance and low complexity make it ideal for general-purpose clarification tasks. Facet-based approaches enhance specificity but increase complexity; they best fit specialized domains requiring detailed clarifications.*

5.5.3 Evaluation of simulated answers

We conducted pairwise comparisons of 200 answer pairs across four aspects: relevance, usefulness, naturalness, and overall quality. Each pair was evaluated by three human workers and three CrowdLLM instances. We define a win for a model when at least two out of three (human or LLM) evaluators agree that the model's answer is superior; if the majority rates the answers as equal, we have a tie.

LLM responses are longer (mean 13.21 vs. 8.19 words) and more variable (std dev. 8.06 vs. 4.36) than human-generated ones. Table 5.8 shows that LLM answers perform comparably to human answers in relevance and usefulness, demonstrating our approach's success in generating contextually appropriate and valuable responses. Naturalness assessments yield intriguing results: human evaluators slightly favor LLM answers (34% vs. 32%), while CrowdLLM shows a stronger preference (55.16% vs. 37.74%), suggesting potential bias in automated evaluation systems. Overall quality marginally favors LLM answers with statistical significance, contrasting with previous non-parametric simulations where human evaluators consistently preferred human answers [197].

Overall, *our parametric approach generates LLM-simulated answers that closely match or slightly outperform human answers across key aspects, contrasting with previous non-parametric simulations and demonstrating successful capture of real user response diversity.*

5.6 Retrieval Performance Comparison

Following Aliannejadi et al. [10], we evaluate the impact of clarifying questions on document retrieval performance. Our methodology simulates a typical conversational search scenario: a user initiates a search, the system poses a clarifying question, and

Table 5.8: Percentage of pairwise comparisons won by each model and ties, as evaluated by CrowdLLM and human evaluators (H-Eval). * indicates statistical significance (trinomial test, P-value < 0.05.

Aspects	**CrowdLLM**			**H-Eval**		
	Human	LLM-answer	Tie	Human	LLM-answer	Tie
Relevance	37.34	37.15	19.71	32.7	36.6	30.7
Usefulness	38.26	41.86	14.47	36.6	38.6	24.8
Naturalness	37.74	55.16*	6.82	32.0	34.0	34.0
Overall-quality	45.52	53.17*	1.82	39.9	41.8*	18.3

Table 5.9: BM25 and BERT retrieval and ranking performance with different clarifying questions. In brackets, answer source: (H) - human answers and (L) - LLM answers.

	Questions	**nDCG@1**	**nDCG@5**	**nDCG@10**
BM25	GPT-Baseline (L)	0.180	0.187	0.209
	GPT-Temp (L)	**0.225**	0.199	0.214
	Human (H)	0.201	**0.221**	**0.246**
	Human (L)	0.173	0.193	0.215
BERT	GPT-Baseline (L)	0.283	0.294	**0.303**
	GPT-Temp (L)	**0.312**	**0.296**	0.301
	Human (H)	0.307	0.288	0.301
	Human (L)	0.267	0.259	0.277

the user provides an answer. The retrieval system then uses this additional information to retrieve an updated set of documents. We hypothesize that higher-quality clarifying questions and their corresponding answers lead to improved post-QA retrieval performance.

Table 5.9 presents the retrieval performance using BM25 and BERT models with clarifying questions and their corresponding answers. GPT-Temp-generated clarifying questions significantly enhance retrieval effectiveness for both the BM25 and BERT models. For BM25, GPT-Temp questions achieve the highest nDCG@1 (0.225), demonstrating superior performance in top-rank retrieval. In BERT-based retrieval, GPT-Temp leads with nDCG@1 of 0.312 and nDCG@5 of 0.296. This performance aligns with our earlier findings, where both evaluators ranked GPT-Temp questions as the most helpful; its ability to generate precise, contextually relevant questions directly translates to improved retrieval outcomes.

Human-generated questions with human answers perform better in BM25 retrieval at nDCG@5 (0.221) and nDCG@10 (0.246). This effectiveness likely stems from two factors: humans' tendency to use terms overlapping with the original query, enhancing lexical matching, and the overall quality of human answers, contributing to improved ranking quality in term-based retrieval. However, when human questions are paired with LLM-generated answers, performance declines across both retrieval models. This contrasts with our finding that LLM-simulated answers are often indistinguishable from human answers in quality assessments. Suggesting that while our parametric approach successfully mimics human-like responses, it may not fully capture the nuanced interactions between clarifying questions and answers crucial for retrieval tasks.

Generally, *GPT-Temp, excels in generating high-quality clarifying questions, enhancing top-ranked retrieval across models. However, BM25 benefits more from precise human-generated content, while BERT leverages contextually rich LLM questions effectively. This highlights the need for LLMs to better integrate query-relevant terms and context for optimal cross-model retrieval performance.*

5.7 Conclusion

We introduce AGENT-CQ, a framework for generating and evaluating clarifying questions and answers in CS systems. Our study reveals that GPT-Temp, our temperature-variation prompting method, consistently outperforms other methods in generating high-quality clarifying questions. Surprisingly, human-generated questions, despite lower quality ratings, excelled in term-based retrieval at nDCG@5 and 10. LLM-simulated answers, while matching human answers in quality assessments, underperformed in retrieval tasks when paired with human questions. CrowdLLM, our evaluation framework, shows general alignment with expert assessments but demonstrates potential biases towards LLM-generated content. Future work should explore enhancing LLMs' ability to generate retrieval-effective questions and answers and improving the integration of LLM-generated content with existing retrieval models.

Addressing **RQ4:** *How effectively can large language models generate and evaluate clarifying questions for conversational search systems?*, AGENT-CQ demonstrates that LLMs can effectively scale both generation and evaluation tasks. Our experiments show that LLM-generated questions significantly improve retrieval effectiveness for both BM25 and cross-encoder models compared to the baseline and human-generated questions. Additionally, CrowdLLM's multi-instance evaluation approach achieves a high correlation with human judgments while enabling assessment at scale.

We continue exploring clarifying questions in the following chapter by investigating how multimodal elements, particularly images, can enhance the effectiveness of clarifying questions and improve user performance.

Chapter Appendices

5.A Additional Methodology Details

In this section, we give additional details on the implementation of AGENT-CQ.

5.A.1 Clarifying question generation algorithms

The two alternative methods used for clarifying question generation in Phase 1 of AGENT-CQ are detailed in Algorithm 5.1 and 5.2.

Algorithm 5.1 Facet-based clarifying question generation

Require: Query q_i
Ensure: Set of clarifying questions C_i

```
F_i ← φ(q_i)                                  ▷ Generate facets
C_i ← {}
for each facet f_ij ∈ F_i do
    c_ij ← ψ(q_i, f_ij)                       ▷ Generate questions
    C_i ← C_i ∪ {c_ij}
end for
return C_i
```

Algorithm 5.2 Temperature-variation-based clarifying question generation.

Require: Query q_i, Temperature variations k
Ensure: Set of clarifying questions C_i

```
C_i ← {}
for j ← 1 to k do
    τ ← min(0.9, 0.5 + (j − 1) ∗ 0.1)
                                              ▷ Update LLM's temperature τ
    c_ij ← ψ(q_i, τ)                          ▷ Generate questions
    C_i ← C_i ∪ {c_i^j}
end for
return C_i
```

5.A.2 User response simulation algorithm

Algorithm 5.3 describes the parameterized answer simulation approach by LLM.

The `ConstructParameterizedPrompt` function generates a structured prompt by incorporating the original query q_i, user information needs u_i, and clarifying question c_{ij}, along with verbosity level and reveal probability randomly selected from the user characteristics set U. This approach generates a wider range of responses, better reflecting real-world user behavior diversity. It also provides a richer dataset for training and evaluating conversational search systems, enabling a systematic study of user characteristics' impact on system performance.

Algorithm 5.3 Parameterized User Response Generation

Require: Query q_i, Facet f_i, Set of clarifying question C_i, User characteristics U
Ensure: Set of parameterized responses A_i

```
A_i ← ∅
for c_ij ∈ C_i do
    prompt ← ConstructParameterizedPrompt(q_i, f_i, c_ij, U)
    a_ij ← ψ(q_i, f_i, c_ij, U)                ▷ Generate response
    A_i ← A_i ∪ {a_ij}
end for
return A_i
```

5.A.3 CrowdLLM question and answer evaluation metrics

Question quality metrics. CrowdLLM evaluated the quality of the generated clarifying questions from a multidimensional perspective, capturing the following quality aspects:

(A1) Clarification: Assesses how well the question seeks to understand the original query without introducing unrelated topics.
(A2) On-topic: Measures the question's direct relation to the subject matter of the original query.
(A3) Specificity: Evaluates the question's focus on particular aspects of the query rather than being general.
(A4) Usefulness: Gauges how much answering the question would improve the response to the original query.
(A5) Clarity: This measure evaluates how easily understood and unambiguous the clarifying question is from the user's perspective.
(A6) Question complexity: This aspect examines whether the clarifying question introduces technical terms, or specialized concepts, or requires domain-specific knowledge not present in the original query.
(A7) Overall quality: Assesses the overall quality of the question based on the above metrics

Answer quality metrics. Answers were also evaluated from a multidimensional perspective on the following three metrics and overall quality.

(M1) Relevance: How directly the user's answer addresses the system's clarifying question.
(M2) Usefulness: The value of the user's answer in clarifying their original information need.
(M3) Naturalness: The human-like quality and conversational tone of the user's response.
(M4) Overall quality: Holistic assessment of the answer's effectiveness in aiding system understanding.

5.A.4 Implementation details

Dataset. We use an existing question clarification dataset, named ClariQ [10]. ClariQ is one of the most widely used question clarification dataset and aligns well with our setting. Each data sample in ClariQ includes a topic originated from the TREC Web Track 2009–2012 [50] that represents an initial user query. These topics can be further divided into multiple facets that capture the user's true intent. For each facet, a set of manually collected clarifying questions are provided which helps the system better understand the underlying user intention. Subsequently, user responses are collected for each clarifying question, providing insights to the corresponding facet. Notably, the ground-truth retrieved documents are also attached given each topic-facet pair. Specifically, in our setting, we reuse the queries from ClariQ but prompt LLMs to generate diverse clarifying questions and simulate user responses. This diversification allows us to better simulate real-world scenarios where users may have different perspectives or require more specific information. Overall, it consists 198 topics with 891 different facets and over 8k questions, with 9.49 terms on average per question.

Models. For our experiments, we use GPT [34] and Llama models [212] in our framework. For question generation, we primarily employ GPT-based models. In the facet-based method, we use a hybrid approach: GPT-3.5 generates query facets, which are then fed to Llama for question generation, as Llama alone was ineffective in facet generation. The same generation model is used in the filtering stage to evaluate and select the most appropriate questions. We used the 8B variant of Llama-3.1. For simulating user responses, we rely on GPT-3.5 due to its versatility in generating diverse and contextually appropriate answers. Our CrowdLLM evaluation framework uses GPT-4o [165] as the base model, leveraging its advanced capabilities for assessing question and answer quality.

Hyperparameters. Our framework employs various hyperparameters, carefully chosen to balance performance and diversity:

1. **Question generation:**
   - Temperature variation: We use temperatures ranging from 0.5 to 0.9, incrementing by 0.1. We set $n\_sets = 3$.
   - Facet-based approach: Temperature is set to 0.7, top_p = 0.95, for Llama: top_k = 50 and max_length = 1024.
   - Baseline: A fixed temperature of 0.7 is used to generate 10 questions for each query.
2. **Question filtering:**
   - We set $\alpha = 0.4$ in the filtering stage to balance relevance and clarification potential of the selected questions.
   - Temperature is set to 0.7.
3. **User simulation:**
   - Verbosity: 10–60 tokens

- Cooperativeness: reveal probabilities 0.0–0.9
- Answer generation: temperature = 0.7, top_p = 0.98, frequency_penalty = 0.5, presence_penalty = 0.2

These parameters simulate diverse user behaviors while maintaining coherent responses.

4. **CrowdLLM evaluation:** We use three GPT-4 instances to simulate diverse human judgments:
   - Conservative judge (temperature 0.2): Produces predictable, focused judgments, simulating a strict evaluator.
   - Balanced judge (temperature 0.5): Provides a mix of creativity and focus, representing a typical evaluator.
   - Creative judge (temperature 0.7): Generates more exploratory judgments, simulating a lenient evaluator.

The selection of these hyperparameters was based on: (i) Extensive experimentation with various setups to optimize performance. (ii) Analysis of output quality and diversity across different parameter combinations. (iii) Alignment with observed patterns in human evaluation behaviors from prior crowdsourcing studies.

5.B Prompts

In this section, we list the prompts used in different prompting strategies and stages of AGENT-CQ.

5.B.1 Facet-based prompt

```
For the user query: '{query}'
Generate a list of 40 diverse facets that this query might be
addressing.
This query represents multiple user information needs.
Generate diverse facets to capture these varied needs.
Ensure each facet is unique and explores different aspects or
interpretations of the query. Avoid repetition and strive for
a wide range of perspectives in your facets.
```

```
For the user query: '{query}'
And considering this specific facet: '{facet}'
Generate a clarifying question that addresses this facet and
helps to better understand the user's specific information
need.
Use diverse language and question structure to formulate the
questions.
```

5.B.2 Temperature-variation prompt

```
for i in range(n_sets):
   For the user query: '{query}'

   Generate a set of 10 clarifying questions. The goal is to
   better understand the user's specific information need.

   This query represents multiple user information needs.
   Generate diverse clarifying questions to capture these
   varied needs.
   Ensure each question is unique and explores different
   aspects or interpretations of the query. Avoid repetition
   and strive for a wide range of perspectives in your
   questions.

   IMPORTANT GUIDELINES:
   1. Each question should aim to clarify a different aspect
   of the user's intent or information need.
   2. Ensure all questions are unique. Do not repeat questions.
   3. Focus on questions that will help narrow down or specify
   the user's request.
   4. Consider potential ambiguities or multiple
   interpretations of the query.
```

5.B.3 Scoring and filtering prompt

```
Evaluate the following question for the user query: '{query}'
Question: "{question}"
Consider these aspects:
   1. Clarification: How well does this question help to
   better understand the user's original query?
   2. On Topic: To what degree does this question directly
   relate to the subject matter of the user's original query?

Provide a score (0-10) for each aspect and a brief explanation.
```

5.B.4 User response simulation prompt

```
You are a user who initially made this request: '{query}'.
Your actual information need is: '{facet}'.
Respond to the clarifying question based on this information
need.

Your verbosity level is {verbosity_level}.
Your reveal probability is {reveal_probability:.2f}.
Keep your response short, ideally under
{verbosity["max_tokens"]} tokens.
```

```
Remember: Your answer should not include any additional
information that is not part of your actual information need
('{facet}').
```

5.B.5 CrowdLLM prompt

Below is an example of CrowdLLM prompt for question complexity. Other metrics follow the same prompt except for the definition of the metric. Each metric is evaluated independently to avoid bias from previous metric rating. Overall quality followed a slightly different approach, apart from having access to the query and system clarifying question, it also included the ratings from the other six metrics in order to ground the overall quality on these metrics.

```
As a user, you are evaluating the complexity of the system's
clarifying question in relation to your original query.

Definition:
   - Question Complexity: The degree to which the clarifying
   question introduces technical terms, specialized concepts,
   or requires domain-specific knowledge not present in the
   original query.

Scale:
   1-10, where 1 is very simple (uses only general terms and
   concepts) and 10 is highly complex (introduces specialized
   terminology or concepts).

Your original query: "{original_query}"
System's clarifying question: "{system_question}"

Evaluate the complexity of the system's question compared to
your original query. Consider:
   1. Does it introduce technical terms or jargon not present
   in the original query?
   2. Does it require specialized knowledge that might not be
   evident from the original query?
```

```
As a user, you are providing an overall evaluation of the
system's clarifying question, taking into account your ratings
from other aspects.

Definition:
   - Overall Quality: Your comprehensive assessment of how
   well the system's clarifying question helps you get a
   better response to your original query, considering
   clarity, relevance, specificity, and usefulness.

Scale:
```

```
   1-10, where 1 is the lowest quality and 10 is the highest
   quality.

Your original query: "{original_query}"
System's clarifying question: "{system_question}"

Your rating from the other metrics: {other_ratings}

Consider these ratings and provide an overall evaluation of
the system's clarifying question quality. Explain your
reasoning, referencing your other metric ratings.
```

5.C Human Evaluation

To quantify the effectiveness of our evaluation framework (CrowdLLM) we conducted human evaluation to assess both the question and answer quality. We employed the so-called Master crowdworkers from Amazon Mechanical Turk from the US with an approval rate of more than 95% in over 10,000 HITs. Each HIT was done by 3 workers and they were paid $8.5 per hour. In total, the question assessment was done by 30 workers, 18 male and 12 female while the answer assessment was done by 18 workers, 7 male and 11 female.

Different from CrowdLLM where we evaluated each question on six dimensions before assessing overall quality, humans assessed the questions based on preference. This is due to the large number of questions and associated costs. In each HIT, a worker was shown the initial user request and five generated clarifying questions for the query from each system: Llama 3.1, GPT-Facet, GPT-Temp, GPT-Baseline, and Human question. Using a drag-and-drop option, they were tasked to rank the questions from the most helpful (Rank 1) to the least helpful (Rank 5). To avoid position bias, where a system's question is always placed at the top, we uniformly randomized the order of the questions at each HIT so that each system question was placed at the top in 20% of the HITs. A total of 1000 questions were assessed, 200 from each system.

Similar to CrowdLLM, humans assessed a pair of answers on three dimensions: relevance, usefulness, naturalness, and overall answer quality. The comparison was between human answers and LLM-simulated answers to human clarifying questions. We used human clarifying questions because they already had human-generated answers. The dataset is well known and has been used in several studies, allowing us to compare with our LLM-generated answers, thus avoiding collecting new human answers. At each HIT the workers were presented with a user information need which we call facet in our case, the initial user query, the human clarifying question, and two answers; one from a human and one from an LLM. Their task was to assess this pair of answers on the four dimensions and choose which of the answers was more relevant, useful, natural, and overall of high quality. If both answers were of the same quality then an option of "Equal" could be selected. Similarly, the order of the answers was randomly swapped. The workers assessed 100 answer pairs and in total 200 answers were assessed.

5.D Supplementary Results and Analyses

5.D.1 Analyses

Question categories. We developed a classification framework for clarifying questions based on Zamani et al. [239] and Braslavski et al. [31]. Table 5.D.1 presents each question type with descriptions and examples, including user questions (UQ) and corresponding clarifying questions (CQ) to illustrate their application in real conversations. This taxonomy provides a robust framework for comparing clarification strategies across various LLMs in conversational information seeking.

Initial rule-based classification attempts proved inadequate for capturing nuances. For instance, "Did you mean the book or the movie?" could be categorized as disambiguation or information gathering, depending on context. To address this, we employed GPT-3.5 for categorization, leveraging its context awareness to select the most appropriate category. This approach enabled more accurate classification, especially for questions requiring nuanced interpretation or potentially fitting multiple categories.

Question patterns. We developed a systematic approach to identify and classify question patterns using a hierarchical matching system. This process analyzes the linguistic structure and key phrases, starting with primary question words (e.g., "what," "how," "are you") and then examining subsequent words for more specific patterns. For example, "what specific" and "what kind of" are categorized differently from general "what" questions. "how" questions are differentiated based on inquiries about methods, duration, or extent. We also consider compound structures like "are you looking for" or "do you need", which are common in clarifying questions. The implementation uses a combination of regular expressions and string-matching algorithms, balancing flexibility in pattern recognition with consistency in categorization. This approach enables nuanced analysis of how different prompting strategies formulate questions.

Elicited response types. We classified expected response types of clarifying questions into four categories: yes/no, Multiple Choice, Open-ended, and Factual. Yes/No questions are identified by auxiliary verb initiation (e.g., "are", "is", "do"). Multiple Choice questions contain explicit options or suggest selections from a limited set. Factual questions use specific question words (e.g., "when", "where", "who") seeking concise information. Open-ended questions, not fitting other categories, typically invite elaboration. We implemented this classification using regular expressions and conditional logic.

To ensure accuracy, particularly for edge cases, we followed the automated classification with a manual review process. This combined approach allowed us to systematically analyze large volumes of clarifying questions while maintaining high classification accuracy. By examining patterns, categories, and response types, we gained insights into how different models and prompting strategies influence the structure and intent of clarifying questions generated by language models in conversational information-seeking contexts.

Table 5.D.1: Clarifying question categories and examples. UQ stands for User Query, which represents an example of a typical user question. CQ stands for Clarifying Question, which shows how a system might respond to the UQ by asking for more specific or relevant information.

Category	**Description**	**Example**
Disambiguation [239]	Addresses queries that are ambiguous and could refer to different concepts or entities.	UQ: I'm looking for information on Java CQ: Are you referring to Java the programming language, Java the island, or Java coffee?
Preference Identification [239]	Clarifies the user's specific preferences, including personal, spatial, temporal, or purpose-related information.	UQ: I want to buy a new laptop CQ: What will be the primary use of this laptop? Gaming, work, or general use?
Information Gathering [31, 239]	Seeks additional details, verifications, or narrows down broad topics.	UQ: Tell me about artificial intelligence CQ: Which aspect of artificial intelligence are you most interested in learning about: machine learning, neural networks, or natural language processing?
Comparison [31, 239]	Involves comparing entities or options to aid decision-making.	UQ: I'm researching electric cars CQ: Would you like to compare the range, performance, or price of different electric car models?
Confirmation [31, 239]	Questions that seek to verify or confirm previously provided information or assumptions.	UQ: I need a new phone CQ: Are you specifically looking for a smartphone, or would you consider other types of mobile phones?
General [31]	Broad questions that prompt for additional details or elaboration on a topic.	UQ: I want to start a business CQ: Can you provide more details about your business idea and what stage of planning you're in?

5.D.2 Question categories

Table 5.D.1 shows the categories of the questions used to categorize the generated clarifying questions.

Table 5.E.1: Number of queries used in the experiments and total questions generated and filtered for each model.

Model	Queries	Generated	Filtered
Llama3.1	50	2000	500
GPT-Baseline	50	500	500
GPT-Facet	50	2000	500
GPT-Temp	50	1500	500
H-Gen	50	1200	500
Total	50	7200	2500

5.E Data Statistics and Sample Generated Questions

In Table 5.E.1 we report the number of questions generated by different systems and the overall filtered questions. Table 5.E.2 shows sample clarifying questions generated by various models for each user query. Table 5.E.3 shows a sample of user responses to the human-generated clarifying questions by humans and an LLM.

Given that GPT-Temp demonstrated superior performance in question quality, we utilize it to generate clarifying questions for all queries in the ClariQ collection. To maintain consistency with the original dataset, we match the number of clarifying questions per facet as present in the original data. This approach ensures a fair retrieval comparison while leveraging the enhanced question generation capabilities of GPT-Temp. For comparative analysis, we also generate a dataset using a GPT-Baseline model. This allows us to assess the relative performance gains achieved by GPT-Temp and provides insights into the effectiveness of different prompting strategies in generating clarifying questions for CS systems.

Table 5.E.2: Clarifying questions generated by different models for various queries

Query	**Llama3.1**	**GPT-Baseline**	**GPT-Facet**	**GPT-Temp**	**H-Gen**
Tell me about atypical squamous cells	What specific information about atypical squamous cells are you looking for?	Is there a specific type of testing or screening you're curious about?	What specific cytopathological techniques are you interested in regarding the diagnosis of atypical squamous cells?	Are you interested in the treatment options available if atypical squamous cells are detected?	Would you like to know if there is a cure?
Tell me information about computer programming.	What specific aspect of the history of computer programming are you interested in?	Do you want to know about programming tools and software?	What specific problem-solving strategies or techniques in programming are you interested in learning more about?	Are you looking for beginner resources, advanced topics, or something in between?	Are you interested in a coding bootcamp
Tell me about defender	Are you interested in learning more about the features and capabilities of defender as a cybersecurity tool?	Are you referring to a specific type of defender, like in sports?	Are you looking for an explanation of defensive coding practices in general, or are you interested in specific techniques or examples of how they are implemented in programming?	Are you referring to a specific type of defender, such as a legal defender, a sports defender, or a cybersecurity defender?	Are you interested in the television series defender
How to cure angular cheilitis	What specific over-the-counter treatments are you looking for to cure angular cheilitis?	What symptoms are you experiencing with angular cheilitis?	Are you looking for home remedies or professional medical treatments to manage angular cheilitis in children?	Would you prefer natural remedies, over-the-counter options, or prescription medications?	Are you looking for the definition of angular cheilitis

Table 5.E.3: Comparison of Human and LLM answers to Human-Generated clarifying questions

Query	User-information-need	Clarifying-question	Human-answer	LLM-answer
Tell me about atypical squamous cells	What do atypical squamous cells mean on a pap smear test?	Would you like to know if there is a cure?	no i want to know what it means when atypical squamous cells appear on a pap smear test	Not specifically. Just curious about their meaning.
Tell me information about computer programming.	What type of careers are there for computer programmers?	Are you interested in a coding bootcamp	no i want to know what career options programmers have	I'm curious about the various job roles within computer programming.
Tell me about defender	Find information on the Land Rover Defender sport-utility vehicle.	Are you interested in the television series defender 3	no i am looking for information about the land rover defender	Not exactly, I'm interested in a specific vehicle model.
How to cure angular cheilitis	What home remedies are there for angular cheilitis?	are you looking for the definition of angular cheilitis	no home remedies	I'm more interested in finding home remedies for angular cheilitis.

6 Effect of Images on Clarifying Questions

In the previous research chapters, we focused on understanding and evaluating text-based interactions in conversational systems. However, modern conversational systems increasingly incorporate multiple modalities to enhance user interaction. This evolution towards multimodality is driven by both technological advances and natural human communication patterns. In daily life, humans frequently combine visual and verbal elements to clarify meaning and convey information more effectively. Clarifying questions, which we explored in Chapter 5 through AGENT-CQ, present an interesting case for multimodal integration: while images might help clarify ambiguous queries, their effectiveness likely varies with task type and user expertise. In this chapter, we investigate:

RQ5: How do images in clarifying questions affect user performance and preferences in conversational search across different tasks and user expertise levels?

Through a user study, we examine the impact of images on two search tasks: answering clarifying questions and query reformulation. This investigation returns to the human-centered evaluation principles established in Chapter 2, focusing directly on user interaction patterns and performance.

6.1 Introduction

Understanding a user's query and intent is one of the main challenges in information retrieval (IR). Users' queries are often very short and can be interpreted in various ways. Search result diversification [49] is a traditional solution to this problem; it aims to present results that cover various aspects or interpretations of the same query. In conversational search (CS), however, due to limited bandwidth, search result diversification is not an effective approach. Other interaction modes and mixed-initiative strategies are employed to enhance the system's understanding of user intent [177], such as preference elicitation, asking clarifying questions, feedback, and query reformulation. Asking clarifying questions is typically used to find out the intent behind the user's query [7].

This chapter was published as C. Siro, Z. Abbasiantaeb, Y. Yuan, M. Aliannejadi, and M. de Rijke. Do images clarify? A study on the effect of images on clarifying questions in conversational search. *In CHIIR'25: ACM SIGIR Conference on Human Information Interaction and Retrieval, Melbourne, Australia, page 273–291.* ACM, 2025.

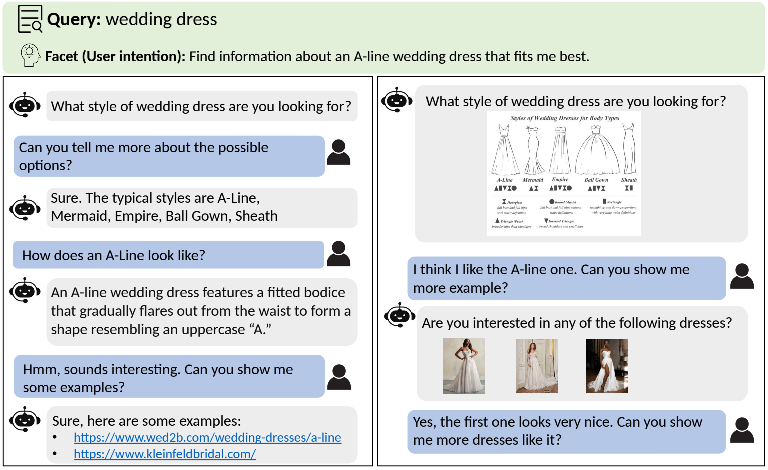


Figure 6.1: Two example conversations with text-only clarifying questions, representing the existing systems (on the left), as well as multimodal clarifying questions, representing our proposed system setup (on the right).

Much work has been done on understanding the impact of clarifying questions both on system and user performance [129, 241, 258]. Studies on text-only clarifying questions show that asking just one clarifying question can lead to considerable improvements in retrieval performance [7, 226]. Clarifying questions are an important element of the user experience in both conversational and ad-hoc retrieval, as determined through controlled user studies [120, 258] and large-scale log analyses [241]. User studies on text-only questions [120, 258] lead to multifold findings, where the usefulness of clarifying questions highly depends on their quality, as well as the user's prior knowledge about the task, greatly impacting user satisfaction.

While clarifying questions enhance CS effectiveness, text-only approaches face limitations in handling queries with visual attributes. Users often struggle to interpret and respond to questions about visual concepts, spatial relationships, or physical attributes through text alone. For instance, describing the specific style of dress you are seeking (Figure 6.1) or explaining the symptoms of a skin condition can be challenging without visual reference points – common elements in domains such as medical diagnosis, product search, and architectural design [61, 97, 134].

Multimodal search research has demonstrated that visual elements can significantly enhance the search process by reducing cognitive load, providing immediate context, and enabling faster recognition of relevant information compared to text-only descriptions [57, 73]. Building on these insights, recent work has explored multimodal clarifying questions in CS, where systems augment textual questions with relevant images to provide additional context and facilitate user understanding [238]. As illustrated in Figure 6.1, such visual enhancements can provide crucial contextual information while potentially influencing users' perception of system understanding.

While previous research has examined the system-side benefits of multimodal clarifying questions [238], understanding user interaction behavior and experience remains crucial for developing effective multimodal CS systems. We address this gap by investigating how the visual enhancement of clarifying questions affects both user experience and their performance. Specifically, we examine how users perceive and use image-enhanced clarifying questions across two fundamental search tasks: answering clarifying questions and query reformulation. Through a within-subject controlled study, we present participants with search scenarios comprising an initial query (e.g., "wedding dress") and its corresponding information need (e.g., "Find information about an A-line wedding dress that fits me best"). Participants interact with clarifying questions (e.g., "What style of wedding dress are you looking for?") under both with-mage and without-image conditions, providing responses aligned with the given information need (e.g., "I want to know more about A-line styled wedding dresses"). This experimental design allows us to systematically investigate how visual enhancement influences user behavior, satisfaction, and performance across different search tasks and user expertise levels.

In this chapter, we address the following chapter-level research questions:

RQ5.1 How do images influence users' answers to clarifying questions in CS?

RQ5.2 What effect do images have on query reformulation in CS?

RQ5.3 When are images useful in CS?

We examine these questions in different types of search task and levels of user expertise. Our findings reveal several important patterns in how visual elements influence search interaction. First, the impact of images varies significantly between search tasks: while users strongly prefer multimodal clarifying questions for direct question answering, their preferences are more nuanced during query reformulation. Second, we find that visual elements play a crucial role in bridging expertise gaps: images help maintain engagement across different knowledge levels in answering clarifying questions, whereas text-only questions show declining effectiveness as user expertise increases. Third, our analysis reveals an interesting disconnect between user preference and their performance: while users generally prefer image-enhanced interactions, their performance (measured by the retrieval effectiveness of their answers) varies by task type. In query reformulation, images help users generate more precise queries, but in question answering, text-only responses often lead to better retrieval outcomes as users provide more comprehensive textual information.

Our study and findings provide useful insights into the design and use of images in multimodal conversational systems. Overall, users find the images useful and helpful in the clarification process; however, their performance reveals mixed results where images are more useful for certain types of search tasks, and clarifying questions. This suggests that when deciding to add an image to a clarifying question, the system should take into account the nature of the user's query, as well as the question type, as they greatly impact the usefulness of images.

6.2 Related Work

6.2.1 User intent clarification

Asking clarifying questions enables IR systems to collect users' explicit feedback, making it an effective interaction mode for various applications, such as product search

[257], voice queries [120], question answering [31, 181, 232], and information-seeking systems [7, 91, 239]. In mixed-initiative IR systems, where either system or the user can take the initiative [11, 95], text-only user intent clarification has been studied extensively [167, 195, 241]. It is known that, in mixed-initiative systems, clarifying questions have the potential to improve search quality and user experience [9, 158, 217].

Research highly depends on user-system interaction data. Aliannejadi et al. [7] present a set of clarifying questions and their answers on TREC Web track [55] queries, which has later been extended to include more topics and clarification need labels, as part of the ConvAI3 competition [10]. Zamani et al. [240] propose a template-based question generation framework, and release a dataset based on the questions generated by their algorithm after being deployed at Bing, called MIMICS. Sekulic et al. [195] propose a model trained on the MIMICS dataset, to predict user engagement for a given clarifying question, and use it as a proxy of quality. Knowing when to ask clarifying questions is an important problem [8], as it can negatively affect the system performance [129] and user experience [258]. Aliannejadi et al. [9] simulate user-system interactions in a conversational information-seeking scenario and study different strategies where they find that depending on the strategy and user's interaction preference, different numbers of clarifying questions can lead to a better cost-gain trade-off. Wang and Ai [226] argue that in various cases, abstaining from asking clarifying questions leads to better retrieval performance (on top of less user effort) and propose a reinforcement-learning-based algorithm to ask a question (or not) based on the expected information gain.

While these studies have established the importance of clarifying questions in CS, they predominantly focus on text-based interactions. The potential of visual elements in the clarification process remains largely unexplored, despite the recognized benefits of multimodal interaction in other search contexts [214]. Our work extends this body of research by examining how visual elements influence both the clarification process and user responses, contributing new insights into multimodal CS.

6.2.2 Multimodal information retrieval

Multimodal IR aims at improving the user experience and system performance by incorporating multimodal information in the interface and retrieval process [223], applied in various IR scenarios such as query reformulation [237], question answering [42, 209], cross-modal retrieval [176]. In mixed-initiative systems, Murrugarra-Llerena and Kovashka [159] propose an image retrieval system that dynamically decides whether the system or user's initiative would be more beneficial to the system's performance. Ma et al. [148] combine mixed-initiative and mixed-modal interactions to improve user experience while interacting with conversational recommender systems. Yuan et al. [238] propose integrating images in the clarification process and releases a multimodal query clarification dataset on the task, based on ClariQ. They show that incorporating images in the clarification process leads to significant retrieval improvements.

Our work differs from these works as they focus on system performance or other IR applications. Our work, on the other hand, focuses on the effect of images on user experience and performance in mixed-initiative conversational systems.

6.2.3 User aspects

A lot of research has focused on system and utility aspects of asking text-only and multimodal clarifying questions. As a relatively recent means of user-system interaction, various user-related aspects of asking clarifying questions are yet to be studied. Kiesel et al. [120] conduct one of the first user studies on the impact of asking clarifying questions on user performance and experience in systems dealing with voice queries. They find that users find such an interaction helpful even in cases where it does not provide a helpful interaction experience. Azzopardi et al. [20] propose a theoretical framework based on an economics model of IR, accounting for various interaction modes involving clarifying questions. In an attempt to discover the effect of clarifying questions of different qualities on the user experience, Zou et al. [258] run a large-scale controlled study where they simulate the clarification system deployed by Bing [239] and present the participants with questions of different qualities. They find that low-quality questions can lead to worse user experience and performance. Presenting multiple questions in the same search session can reduce the risk. In a follow-up study, they examine the effect of clarifying questions in multi-question sessions. Although it is less risky to ask multiple clarifying questions in the same session, they find that users start to lose their trust in the system if the system starts the session with low quality questions [258].

While inspiring our study, these studies do not focus on image-enhanced clarifying questions. Also, in most cases, the information-seeking task mimics a web search scenario, whereas we focus on multimodal conversational information-seeking.

6.3 Study Design

To investigate the role images play during search clarification, we conducted a controlled user study. We examined the impact of images on CS interactions, focusing on two key tasks: answering clarifying questions and query reformulation. We selected these tasks as they represent the primary actions users take during CS clarification – either directly answering system questions or modifying their queries based on the interaction [9]. The tasks represent two distinct yet interconnected aspects of the search process. When users respond to clarifying questions, they must interpret the system's request and provide relevant information, a process that may be altered by the presence of visual cues. Similarly, query reformulation represents users' evolved understanding of their information need, potentially influenced by the clarifying question and its accompanying image(s). Based on prior work in CS systems [7] and mixed-initiative search [18], we propose the following hypotheses:

H1 Multimodal clarifying questions will lead to higher user engagement and satisfaction compared to text-only questions.

Rationale: Visual elements could provide context that complements textual information in clarifying questions. A multimodal approach would enhance user understanding and engagement by offering multiple channels for processing information. The complementary nature of images and text could help maintain efficient task completion times. The integration of visual and textual elements aligns with cognitive theories suggesting that multiple representational formats can enhance information processing and comprehension [152].

H2 Users' background knowledge on a search topic will influence their perception and dependence on images in clarifying questions.

Rationale: Based on research showing that domain knowledge significantly affects how users process and integrate information during search tasks [228], we hypothesize that experts and novices may differ in their ability to extract and utilize visual information effectively during the clarification process.

H3 The utility of images in clarifying questions will vary based on query type, with higher perceived usefulness for queries with inherent visual attributes (e.g., descriptive or visual information needs) than abstract or conceptual queries.

Rationale: Visual content naturally excels at conveying physical and spatial characteristics that text alone struggles to describe efficiently [26, 214]. When answering clarifying questions about visual attributes (like product appearances or spatial arrangements), users can reference images directly rather than interpreting textual descriptions [174]. Similarly, during query reformulation, images provide concrete visual anchors that help users articulate visual concepts more precisely in their refined queries [99].

6.3.1 Topic selection and pre-study analysis

To investigate the role of images in clarifying questions, we need to carefully curate a set of search tasks where images can have a meaningful impact. Although the ClariQ dataset [10] provides a foundation of CS topics, it has not been designed with visual elements in mind. MELON [238], an extension of ClariQ incorporating images, has inconsistencies in image quality and relevance due to its crowdsourced image collection process (as revealed by our initial inspection). Previous studies show that crowdworkers may not consistently select optimal images for search tasks [162]. To control for these variables, we curate our own image collection and conduct a systematic pre-study to identify suitable topics for this study.

In our pre-study task, we sampled 100 topics from ClariQ and employed two appen[1] assessors to judge the potential benefit of image augmentation for clarifying questions. We asked the assessors to provide detailed justifications as to why the topics would benefit from visual elements (or not). Analysis of workers' justifications revealed three key characteristics that made topics amenable to visual augmentation: (a) Physical structures or objects (e.g., "hip roof construction," "solar water fountains"); (b) Medical conditions with visual symptoms (e.g., "carpal tunnel syndrome"); and (c) Natural elements requiring visual identification (e.g., "norway spruce characteristics").

Based on these insights, we selected 24 topics (Table 6.1) and formulated 6 additional ones, making it a total of 30 topics for the study. We modified the information needs to align with identified visual enhancement opportunities while maintaining the original search context from ClariQ. For example, "hip roof" is extended to address visual aspects of roof structure and design elements. Images are sourced from Google image search by querying the topic, aiming to complement the clarifying questions.

[1] `https://www.appen.com/`

Table 6.1: Topics and facets selected from ClariQ for the study with the reference to ClariQ facet ids in brackets. Note: Topics without a reference are those that we reformulated ourselves.

Topics	Topics
T1: dangers of asbestos (F0075)	T16: volvo (F0459)
T2: norway spruce (F0736)	T17: land surveyor (F0455)
T3: home theater systems (F0468)	T18: american military university (F0146)
T4: grilling (F0243)	T19: ct jobs (F0889)
T5: dinosaurs (F0162)	T20: cass county missouri (F0032)
T6: kids earth day activities (F0491)	T21: electoral college 2008 results (F0507)
T7: teddy bears (F0607)	T22: rick warren (F0497)
T8: hip roof (F0616)	T23: angular cheilitis (F0206)
T9: solar water fountains (F0493)	T24: barrett's esophagus (F0600)
T10: carotid cavernous fistula treatment (F0716)	T25: moths (F0921)
T11: ham radio (F0543)	T26: patron saint of mental illness (F0609)
T12: carpal tunnel syndrome (F0477)	T27: altitude sickness (F0757)
T13: cloud types	T28: car dashboard symbol
T14: bike repair	T29: office chair
T15: coffee table	T30: pipe fittings

6.3.2 Study setup

We employed a within-subjects design. Both studies used the same set of 30 topics, with each topic presented in two setups: with- and without-image. This design allowed for a direct comparison of how the presence or absence of an image affected participants' responses to the same questions.

Task structure. For each task, participants were presented with:

1. *Information need:* A detailed description of the user's context and search goal, derived from the ClariQ facets.
2. *Initial query:* The first search query a user entered, based on their information need.
3. *Clarifying question:* A question from the system, designed to better understand the user's intent or narrow down the search focus.
4. *Image:* An image added to a clarifying question, intended to provide additional context or information.

Tasks. We explain the two tasks below:

1. **Task 1: Answering clarifying questions.** We asked the participants to imagine themselves as the actual users seeking information and to answer the clarifying question as if they had the given information need.
2. **Task 2: Query reformulation.** We asked the participants to act as users with the information need who had submitted their initial query. Their task was to reformulate their initial query after being exposed to the clarifying questions.

Questionnaire design. Following established practices in interactive IR evaluation [118], we designed three questionnaires to measure aspects of the user experience with multimodal clarifying questions:

Pre-task demographics questionnaire. Before beginning the tasks, participants completed a questionnaire capturing:

- Demographic information (age, gender, education level),
- Experience with conversational systems (5-point scale: no experience to very experienced), and
- Experience with the task.

The demographic questions were selected based on factors known to influence search behavior [228] and to ensure our sample represented diverse user characteristics.

Post-clarification questionnaire. After each clarifying question, participants completed a brief questionnaire designed to capture immediate feedback. The choice of questions was motivated by aspects of CS interaction [10]:

1. Information sources relied upon when answering the clarifying question (e.g., image, initial query, clarifying question, personal knowledge).
2. Background knowledge level on the topic (3-point scale: expert, familiar, new topic).
3. Clarity of the provided clarifying question (5-point scale: very unclear to very clear).
4. Usefulness of the provided clarifying question (5-point scale: not useful at all to very useful).
5. Overall satisfaction (5-point scale: very poor to excellent).

Exit questionnaire. We designed the exit questionnaire to capture overall patterns across the interaction. Questions covered:

1. Overall experience with multimodal clarifications (5-point scale: significantly difficult to significantly easy).
2. Overall experience with text-only clarifications (5-point scale: significantly difficult to significantly easy).
3. Impact of images on clarification efficiency (5-point scale: greatly slowed to greatly sped up).
4. Frequency of cases where images provided helpful context (5-point scale: never to always).
5. Preference for setup (multimodal, text-only, no preference).
6. Justification for their preference (open-ended).
7. Situations where images were most helpful in understanding clarifying questions (e.g., physical objects, processes[2]).

[2]These categories were derived from our pre-study analysis of topic characteristics and aligned with the types of queries where visual elements were predicted to be most beneficial.

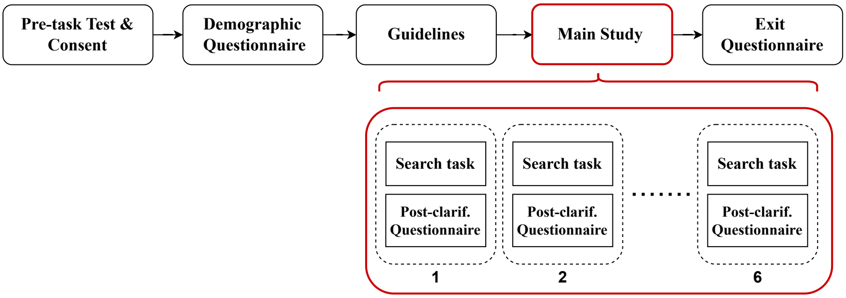


Figure 6.2: Our user study procedure. "Main Study" refers to either query reformulation or answering the clarifying question tasks.

6.3.3 Procedure

The study was conducted in two separate experiments (i.e., answering clarifying questions and query reformulation), each with its own set of participants. Both experiments included the following steps before starting the study of main search tasks as shown in Figure 6.1: (i) pre-task test and consent to the collection of their demographic information; (ii) filling in the demographic information; and (iii) reading instructions and examples on completing the task.

Search task distribution. In total, we had 60 search tasks for each experiment. We randomly grouped them into 10 batches where each batch included 6 search topics. We ensured that each participant encountered an equal number of multimodal and text-only clarifying questions and that each clarifying question was evaluated in both setups across the participant pool. Each batch was assigned to at least 3 users. To mitigate learning effects, we limited each participant's involvement to a maximum of two batches for a single experiment. While participants were exposed to both with- and without-image clarifying questions, we structured the batches to ensure that no participant encountered the same clarifying question in both setups within a single batch. The users were given the tasks one by one. After successful completion and submission of each task, they were enabled to move to the next task. The entire procedure for each experiment was designed to be completed within approximately 10–15 minutes. The detailed procedure for each experiment is as follows:

Task 1: Answering clarifying questions. Each participant engaged with six search topics. The order of the questions and the presence/absence of images were randomized and counterbalanced to mitigate learning effects. After each question, participants completed the post-clarification questionnaire. Upon responding to all six questions, participants answered the exit questionnaire.

Task 2: Query reformulation and exploration. A different set of participants responded to six search topics, each involving query reformulation. As in Task 1, the order of questions and the presence/absence of images was randomized and counterbalanced. Participants completed the post-clarification questionnaire after each question and the exit questionnaire after completing all six questions.

Quality assurance. To ensure the quality of our user study we implemented the following:

– We started the study with a pre-task test including two attention check questions. Participants were allowed to only do this test once. In case of failure, participants were not allowed to continue with the study. In case of successful completion of the pre-task test, participants were redirected to the main task.
– We provided a comprehensive guideline for the participants to complete the tasks. We included detailed descriptions and examples.
– After completing the study by participants, we manually checked the responses provided by them and discarded the low-quality annotations. We discarded low-quality responses from 6 participants in Task 1, and 7 participants in Task 2.
– We ran 3 pilot studies to improve our guidelines and questions and in each pilot, we used 4 participants.
– We maintained the integrity of the experiment by recruiting separate groups of participants for each task. Participants in Task 1 were not eligible to participate in Task 2, and vice versa. This approach prevented potential bias and cross-contamination between the two tasks.

6.3.4 Data collection

We collected the following data: Text responses (answers to clarifying questions, reformulated queries); Self-reported information reliance; Ratings on various aspects from the questionnaires (pre-, post-clarification, and exit questionnaires; and Open-ended responses from the exit questionnaire (justification on image preference). Overall, we collected 360 data points; 180 for each task. Per task, we collected 90 samples per condition (with- vs. without-image). All collected data is stored in a local password-protected computer to ensure data privacy.

6.3.5 Participants

We recruited participants through Prolific.[3] To ensure high-quality responses, we applied strict filters, requiring participants to have a 95% or higher approval rate and to have completed more than 3000 tasks on Prolific. Eligible participants were 18 years or older, had English as their native language, and were regular users of search engines and digital assistants.

Task 1, which focused on answering clarifying questions, involved 36 unique participants ($N = 36$), each participant completed one batch. The age distribution: 38–47 years ($n = 10$), 28–37 and 48–57 years ($n = 7$ each), 68+ years ($n = 6$), 58–67 years ($n = 5$), and 18–27 years ($n = 1$). Male participants ($n = 25$) outnumbered female ($n = 11$) participants. Most participants held bachelor's degrees ($n = 25$), followed by master's degrees ($n = 6$), PhDs ($n = 2$), and other ($n = 3$). The majority ($n = 27$) had no prior experience in answering clarifying questions, while some ($n = 9$) reported previous experience. Participants reported having moderate knowledge levels on the topics involved in the task (2.59/3).

[3] https://www.prolific.com/

Task 2 involved 37 unique participants ($N = 37$), each completing one batch. The age distribution was: 38–47 years ($n = 10$), 48–57 and 58–67 years ($n = 8$ each), 28–37 years ($n = 6$), 18–27 years ($n = 4$), 68+ years ($n = 1$). There were 21 males, 13 females and 3 preferred not to say. Bachelor's degree holders formed the largest group ($n = 26$), followed by master's degree holders ($n = 6$), while high school graduates, PhD holders, and those with other qualifications each comprised $n = 2$ participants. Regarding task experience in reformulating clarifying questions, $n = 24$ participants reported no prior experience, while $n = 13$ indicated previous experience. For query reformulation, participants reported having an average knowledge level of 2.6 on the search topics.

6.3.6 Data analysis

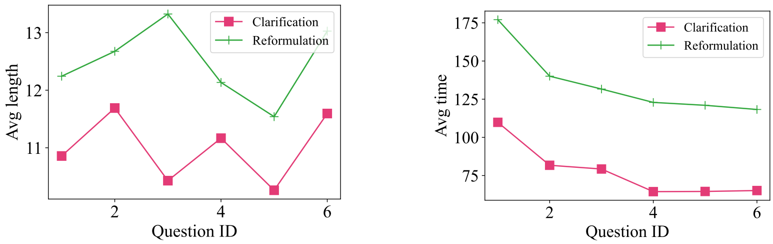
Our study employed a mixed-methods approach to analyze the collected data, combining quantitative statistical analyses with qualitative content analysis.

Quantitative analysis. The quantitative analysis examined responses from both post-clarification and exit questionnaires using multiple statistical approaches. To account for the hierarchical nature of our data, where responses are nested within participants, we employed linear mixed-effects models with setup (with- vs. without-image) as a fixed effect and participant as a random effect. This approach allowed us to control for individual participant variations while examining the effect of image presence. To complement the mixed-effect analysis, we conducted independent t-tests comparing responses between with- and without-image setups, with Bonferroni correction ($\alpha = 0.05/6 = 0.0083$) to control for multiple comparisons. Effect sizes were calculated using Cohen's d. We conducted one-way ANOVA tests with post-hoc Tukey's HSD to examine the influence of participants' background knowledge levels (novice, familiar, expert) on their interactions. All statistical analyses were performed with a significance level of $\alpha = 0.05$, with appropriate corrections for multiple comparisons where applicable.

Qualitative analysis. The qualitative analysis examined three types of open-ended responses: answers to clarifying questions, reformulated queries, and justifications for setup preferences. Two independent coders analyzed these responses using thematic analysis. The coders first independently identified recurring patterns in how participants used images across different response types. They then developed a coding scheme through discussion and iteration, focusing on patterns that emerged in both with- and without-image conditions. The final coding scheme was applied to all responses.

6.4 Results

In this section, we present findings from our user study examining both tasks: answering clarifying questions (Task 1) and query reformulation (Task 2). We analyze results across two conditions (with and without image).



(a) Average answer length per question

(b) Average completion time per question

Figure 6.3: Average time taken by participants to complete each question (a) and the average length of the clarifying question and reformulated queries (b).

6.4.1 Descriptive statistics

Participants spend on average 77.52 seconds answering each clarifying question in Task 1; Task 2 requires substantially more time with an average of 135.15 seconds per query reformulation. This difference highlights the increased cognitive demand of query reformulation compared to direct question answering. Analysis of completion times reveals a consistent decrease as participants progressed through the questions (Figure 6.3b). To distinguish between potential fatigue and learning effects, we examine the relationship between completion time and response length. While time decreases across questions, the average length of both answers and reformulated queries showed consistent variation (Figure 6.3a), suggesting that faster completion times result from task familiarity rather than decreased engagement.

The exit questionnaire completion times also reflect the differential cognitive load between tasks: Task 1 takes 84.50 seconds (median: 69.00 seconds) and Task 2, 92.14 seconds (median: 74.00 seconds), with a mean difference of 7.64 seconds, and median difference of 5.00 seconds. This difference suggests that the nature of the preceding task (answering clarifying questions vs. reformulating queries) influences participants' response times and potentially their depth of reflection. We hypothesize that the longer completion times in Task 2 may be due to the complexity of the task, thus leading to more extensive cognitive processing, not only during the task itself but also during subsequent reflection.

6.4.2 Task 1: Answering clarifying questions

Table 6.2 presents the results, combining mean rating, mixed-effects analysis, and independent t-test results.

Question clarity. Our analysis reveals that clarifying questions are generally perceived as clear across both setups, with ratings predominantly in the 4–5 range (Figure 6.4a). Interestingly, questions without images are rated marginally clearer (3.93 vs. 3.85), with our analysis showing a small negative effect of images ($\beta = -0.079$, $d = -0.078$), where questions without images receive slightly more maximum clarity ratings (5/5). This raises an interesting possibility that images might occasionally introduce complexity rather than clarity to the question-answering process. The high

Table 6.2: Analysis results for Task 1: Clarifying Question Answering. Note: Mean rating show M (SD). β represents the fixed effect of condition (image presence); SE = Standard Error of the coefficient estimate. ICC = Intraclass Correlation Coefficient; d = Cohen's d effect size.

	Mean rating		Mixed-Effects		Independent t-test		
Measure	With Image	Without Image	β (SE)	ICC	t(180)	p	d
Topic Knowledge	2.60 (0.51)	2.58 (0.54)	0.024 (0.061)	0.160	0.360	0.719	0.045
CQ-Usefulness	3.78 (1.14)	3.72 (1.13)	0.056 (0.124)	0.243	0.389	0.698	0.049
CQ-Clarity	3.85 (1.03)	3.93 (1.00)	-0.079 (0.098)	0.420	-0.622	0.535	-0.078
Satisfaction	3.67 (0.88)	3.70 (0.82)	-0.032 (0.083)	0.399	-0.297	0.767	-0.037
Answer Length	11.30 (6.50)	10.53 (5.75)	0.770 (0.596)	0.407	0.996	0.320	0.125
Completion Time	79.00 (54.72)	76.05 (44.92)	2.952 (5.078)	0.355	0.468	0.640	0.059

Table 6.3: Analysis Results for Task 2: Query Reformulation. Note: Mean rating show M (SD). β represents the fixed effect of condition (image presence); SE = Standard Error of the coefficient estimate. ICC = Intraclass Correlation Coefficient; d = Cohen's d effect size.

	Mean rating		Mixed-Effects		Independent t-test		
Measure	With Image	Without Image	β (SE)	ICC	t(180)	p	d
Topic Knowledge	2.64 (0.54)	2.56 (0.53)	0.081 (0.065)	0.177	1.129	0.260	0.152
CQ-Usefulness	3.24 (1.17)	3.14 (1.20)	0.108 (0.128)	0.351	0.681	0.496	0.091
CQ-Clarity	3.56 (0.96)	3.55 (1.05)	0.009 (0.114)	0.286	0.067	0.947	0.009
Satisfaction	3.43 (0.85)	3.37 (0.79)	0.063 (0.088)	0.357	0.575	0.566	0.077
Refor. Query l	12.29 (5.26)	12.63 (5.37)	-0.342 (0.546)	0.417	-0.480	0.632	-0.064
Completion Time	135.63 (61.70)	134.68 (72.97)	0.955 (6.393)	0.503	0.105	0.916	0.014

ICC for clarity (0.420) suggests that participants were moderately consistent in their clarity evaluations across questions.

Question usefulness. Participants find the clarifying questions useful overall, with ratings skewed towards the higher end of the scale (Figure 6.4b). Questions with images are rated marginally more useful (3.78 vs. 3.73), showing a small positive effect ($\beta = 0.056$, $d = 0.049$). The substantial consistency in individual ratings ($ICC = 0.399$) indicates that participants maintain stable opinions about question utility across different scenarios, regardless of the setup

Overall satisfaction. Participants report high satisfaction levels across both setups (mean of 3.68 on a 5-point scale). The presence of images has minimal impact on satisfaction ratings (3.66 with images vs. 3.69 without), with our analysis confirming this negligible difference ($\beta = -0.032$, $d = -0.037$). The high consistency in individual ratings ($ICC = 0.399$) indicates that participants maintain stable satisfaction levels across different questions, suggesting that image presence does not substantially alter their overall experience with the clarification process.

User answers. On average, participants provide relatively concise answers regardless of the experimental setup, with longer answers when images are included (11.30 words, $SD = 6.50$) than without images (10.53 words, $SD = 5.75$). With-image questions take slightly longer time to complete (79.00 vs. 76.05), suggesting that users need additional time to process visual information. The moderate ICC (0.355) for completion time indicates that while individual differences exist in response speed, they are not as pronounced as in other measures.

6.4.3 Task 2: Query reformulation

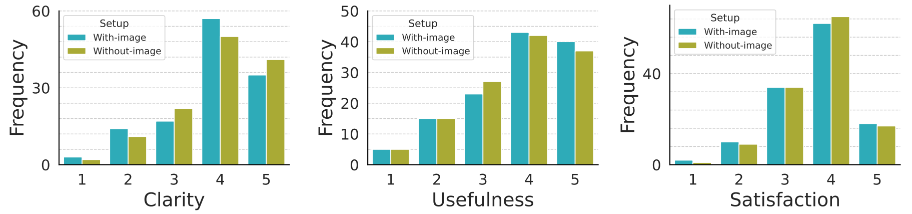
Task 2, focusing on query reformulation, shows distinctive patterns in participant ratings for clarity and usefulness, as shown in Figure 6.4.

Question clarity. While question clarity ratings are positive, Table 6.3 indicates a minimal difference between setups (3.56 vs. 3.55, $\beta = 0.009$). The lower ICC for clarity (0.286) compared to Task 1 suggests more variability in how participants evaluate clarity in the reformulation tasks.

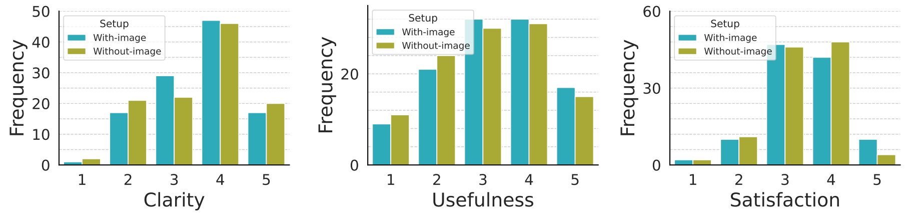
Usefulness. Usefulness ratings are moderate with Figure 6.4e showing a shift towards lower ratings compared to Task 1. While multimodal questions were rated slightly more useful (3.24 vs. 3.14) ($\beta = 0.108$, $d = 0.091$), the higher ICC (0.351) indicates more consistent individual preferences.

Overall satisfaction. Satisfaction levels remain moderate to high, with a small advantage for multimodal conditions (3.43 vs. 3.37, $\beta = 0.063$, $d = 0.077$). The high consistency in individual ratings ($ICC = 0.357$) indicates that participants maintain stable satisfaction levels across different reformulation tasks.

Reformulated queries. For query length, while original queries are quite short (mean of 3.13 words), reformulated queries are significantly longer (mean of 12.46 words). When comparing the two setups, we observe a subtle difference in reformulated query lengths: queries are slightly shorter when images are included (12.29 vs. 12.63 words). This suggests that images may lead to more concise query reformulations, possibly by helping users focus their information needs more precisely.



(a) Clarity (b) Usefulness (c) Satisfaction level



(d) Clarity (e) Usefulness (f) Satisfaction level

Figure 6.4: Rating distributions of main task aspects as rated by participants in Task 1 – row 1 and Task 2 – row 2.

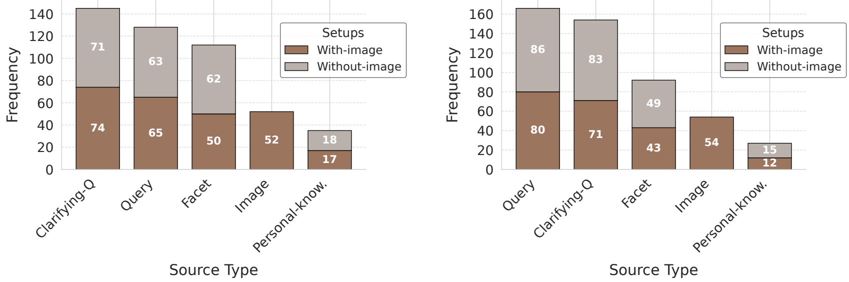
Completion time. The higher completion time ($M = 135.63s$, $SD = 61.70s$ with images) compared to Task 1 reveals the more cognitively demanding nature of query reformulation. The high ICC for completion time suggests that individual differences in reformulation strategies are more stable and pronounced than in answering clarifying questions.

Summary. Our findings do not support **H1**, that visually-enhanced clarifying questions lead to higher user satisfaction and engagement. While images show some positive effects on usefulness ratings and minor benefits for satisfaction, these differences are not substantial. The impact of images varies by task type: in Task 1, they lead to longer answers but slightly reduced clarity, while in Task 2, they result in more concise reformulations with minimal impact on clarity. This suggests that the value of visual enhancement may be more nuanced and task-dependent than initially hypothesized.

6.4.4 Information sources relied on when answering clarifying questions and reformulating queries

Figure 6.5 presents participants' utilization of information sources across two search tasks: answering clarifying questions (Task 1) and query reformulation (Task 2), comparing with-image and without-image setups.

Task-based patterns. The primary information source shifts between tasks, with clarifying questions dominating in Task 1 (145 total references) and initial queries in Task 2 (166 references). This aligns with task requirements: question answering naturally emphasizes the clarifying question, while reformulation centers on modifying the original query. Task 2 generally exhibits higher frequencies of information source usage than Task 1, particularly for clarifying questions and initial queries. On average each



(a) Task 1 (b) Task 2

Figure 6.5: Information sources relied on by participants when a) answering clarifying questions and b) reformulating queries.

participant relies on 3.2 information sources in Task 2 compared to 1.8 in Task 1. This difference likely reflects the more complex cognitive process involved in reformulating a query, which requires synthesizing information from multiple sources to generate a new, improved query.

Impact of visual enhancement. The presence of images influences information source utilization patterns. In Task 1, visual cues enables more balanced use of clarifying questions (74) and queries (65), while its absence leads to increased facet use (62 vs. 50). Task 2 shows stronger compensation patterns in the without-image condition, with higher reliance on both queries (86 vs. 80) and clarifying questions (83 vs. 71).

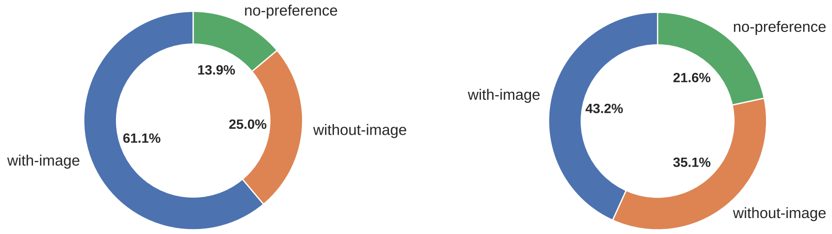
Across both tasks, personal knowledge remains consistently the least referenced source, with minimal variation between conditions, suggesting that participants primarily relied on provided information rather than prior knowledge regardless of visual support.

Summary. These patterns demonstrate that the presence of images alters how users approach information gathering in CS. While images serve as an additional information source, their impact extends beyond direct usage – they appear to streamline the information integration process and reduce reliance on textual sources. This suggests that visual elements not only provide direct information but also help users more efficiently process and combine information from multiple sources during search tasks.

6.4.5 Exit questionnaire aspects

The exit questionnaire survey results, depicted in Figure 6.7, reveal participants' perceptions and preferences regarding multimodal clarifying questions across both tasks.

Task 1. In Task 1, images significantly enhance the question-answering process. 36% of participants find images "Sometimes" helpful, with 33% reporting improved task understanding and completion speed. Users strongly prefer multimodal setups, with 61.1% favoring this setup (Figure 6.6a), reflected in the high percentage rating on their experience with image setup. The ease of use is evident, with 72% rating multimodal setups as easy or significantly easy. These results clearly demonstrate that visual aids provide immediate context and facilitate comprehension when answering clarifying questions. The straightforward nature of this task allows users to directly leverage



(a) Task 1 (b) Task 2

Figure 6.6: Self-reported preference between setups.

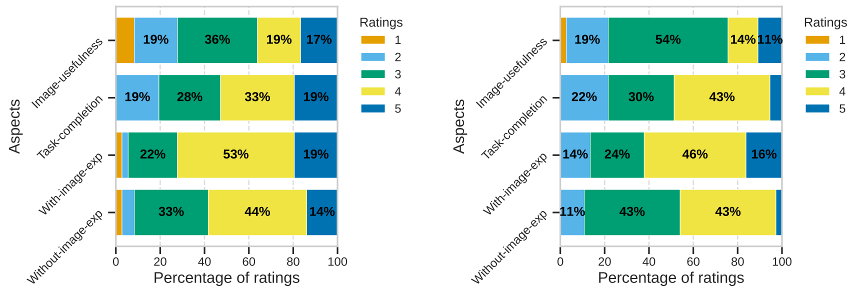
visual information, explaining the consistent, moderately beneficial impact of images.

Task 2. Task 2, images prove more consistently useful, with 68% rating them as "Sometimes" or "Often" helpful (Figure 6.7b). 43% of users report improved task understanding and completion speed with images. However, preferences are more evenly distributed: 43.2% prefer multimodal setups, 35.1% favor text-only setups, and 21.6% have no preference. This balanced distribution aligns with the polarized experience ratings for multimodal setups in Task 2. These findings indicate that while users recognize the potential value of images in providing rich, multi-dimensional context that can spark new ideas or highlight previously unconsidered aspects of a topic, the actual preference for using images in this task is more varied. This variation stems from the complex nature of query reformulation, which requires synthesizing information and generating new search terms. While, images may provide rich, multi-dimensional context that can spark new ideas, the same richness can be distracting for some users or query types.

Summary. These findings provide partial support for **H1** (visually-enhanced clarifying questions leads to higher user satisfaction and engagement). While images clearly enhance satisfaction in the question-answering task, their impact on query reformulation is more variable. This suggests that the effectiveness of visual enhancement depends on the specific task context: straightforward for direct question answering, but more complex and individual-dependent for query reformulation. The higher reported usefulness but lower preference for images in Task 2 indicates that visual aids can provide valuable context while potentially introducing cognitive complexity in more demanding tasks.

6.4.6 Image usefulness in query clarification

Figure 6.8 reveals the distribution of situations where participants find images to be useful; Table 6.4 has examples. In Task 1, physical object scenarios dominate at 50%. This high percentage indicates that when users seek information about tangible items, such as dinosaurs (T5), images are likely to be highly beneficial. Context-related and process-related situations each account for 14%, demonstrating equal importance in scenarios like clarifying hip roof construction (T8) or explaining grilling techniques (T4). Abstract concepts represent 8% of situations, while technical details and data visualization account for 6% and 3%, respectively.



(a) Task 1 (b) Task 2

Figure 6.7: Rating distributions of final task aspects as rated by participants in Task 1 and Task 2.

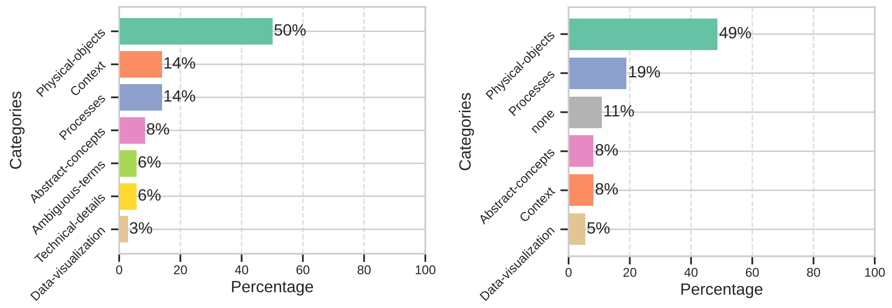
Task 2 maintains the prominence of physical object scenarios at 49%, showing the consistent value of visual representations for tangible items. It reveals a significant increase in process-related situations to 19%. This increase reflects the greater utility of visual demonstrations when users refine queries about specific processes, such as detailed grilling methods (T4). Context-related scenarios decrease to 8%, indicating less need for broad visual overviews during query reformulation. Abstract concepts remain steady at 8%; data visualization shows a slight increase to 5%.

To better understand when images are perceived as useful or not, we analyze participants' justifications for their setup preferences through thematic analysis. Our analysis reveals four primary useful aspects and four not-useful aspects of image utility in search interactions (Table 6.5).

Beneficial aspects. Images prove most valuable for contextual support (28.4% of participants), particularly when users need additional context to understand and reformulate their queries. As one participant notes, "The images helped me to have some more context to rephrase the question." Cognitive facilitation is the second most prominent benefit (24.3%), with participants highlighting how images simplify information processing and reduce cognitive load, especially for complex topics. Creative stimulation (21.6%) is another key benefit, with images inspiring alternative query formulations and new perspectives: "They made me think of alternative angles for the questions that I might not have thought about otherwise." Enhanced engagement and focus (16.2%) round out the beneficial aspects, with images helping maintain task attention and improve understanding.

Limiting aspects. Our analysis also reveals limitations. Cognitive overload from irrelevant visuals (13.5%) is the most frequently cited problem, particularly when images introduce unnecessary complexity. Some participants (10.8%) find images irrelevant to their task goals, noting misalignment between visual content and search objectives. Individual preferences play a role, with 9.5% expressing a clear preference for text-based interaction. Finally, 8.1% view images as redundant, adding no value beyond the textual information.

Summary. Our findings provide strong support for **H3**: image utility varies systematically with query type. The clear dominance of physical objects and process-related



(a) Task 1 (b) Task 2

Figure 6.8: Participants self-reported situations where images were useful.

Table 6.4: Examples of topics and clarifying questions for different image utility situations.

Situation	**Topic (T)**	**Clarifying Question (CQ)**
Physical objects	T5: Dinosaurs	CQ1: Are you interested in the types or their history?
Processes	T4: Grilling	CQ2: What techniques or tips for grilling vegetables would you like to know more about?
Context	T8: Hip roof	CQ3: What specific information do you need regarding the construction or specifications of a hip roof?
Abstract concepts	T10: Carotid cavernous fistula treatment	CQ4: What particular treatment options are you most curious about?
Technical details	T11: Ham radio	CQ5: Are you looking for information on the different designs of ham radio antennas and their specific applications?
Data visualization	T12: Carpal tunnel syndrome	CQ6: Which types of exercises are you considering—those for immediate relief or for long-term prevention?

scenarios in perceived image usefulness, coupled with the lower utility reported for abstract concepts, confirms our hypothesis. Our thematic analysis further reinforces this pattern, showing that positive perceptions of images (contextual enhancement, cognitive efficiency) are predominantly associated with queries having inherent visual attributes, while negative perceptions (cognitive overload, irrelevance) are more common with abstract or conceptual queries.

6.4.7 Effect of background knowledge

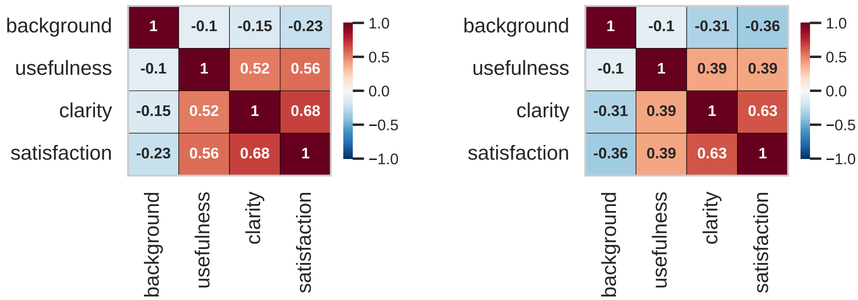
We examine how users' background knowledge influences their perception of clarifying questions with and without images. Figures 6.9 and 6.10 present Spearman's ρ correlation coefficients and mean ratings for usefulness, clarity, and satisfaction across different knowledge levels for both tasks.

In Task 1, while measures show strong positive correlations among themselves (Figure 6.9a, $\rho > 0.6$, $p \leq 0.01$), we observe a consistent negative correlation between background knowledge and all other measures. Similarly, Task 2 exhibits even stronger negative correlations between background knowledge and user ratings (Figure 6.9b, $\rho \leq -0.4$, $p \leq 0.01$). Given these observed negative correlations, we conduct further

Table 6.5: Thematic analysis of Participants setup preference justifications to understand when images are useful and not.

Category	**Characteristics**	**Example quote**	**Freq. (%)**
Useful aspects			
Contextual enhancement	Provides additional context; Enhances query understanding; Clarifies search intent	"Gave me additional information to enable me to reformulate my question"	28.4
Cognitive efficiency	Reduces mental effort; Simplifies information processing; Enables faster comprehension	"A picture very often helps clarify things without a lot of written, and sometimes superfluous information"	24.3
Creative stimulation	Inspires new query angles; Suggests alternative formulations; Promotes exploration	"They made me think of alternative angles for the questions that I might not have thought about otherwise"	21.6
Increased engagement and focus	Increases task engagement; Enhances attention; Improves task focus	"I thought the images added a good initial visual prompt this helped in determining context"	16.2
Not Useful Aspects			
Cognitive overload	Introduces unnecessary complexity; Creates processing burden; Complicates task	"They overloaded me with information and I found it more difficult to construct a straightforward question"	13.5
Irrelevant to user	Misaligns with user expectations; Lacks task relevance; Provides unhelpful information	"Images were not always relevant to what I needed to know"	10.8
User preference for text	Individual preference for textual information; Comfort with text-based search	"I am better with words than images"	9.5
Redundancy	Adds no additional value; Duplicates textual information	"I didn't see the images as providing any more information that I required to reformulate the query"	8.1

Note: Percentages represent the proportion of participants who mentioned each aspect in their responses. Individual responses often contained multiple aspects, hence percentages sum to more than 100%. For example, a single participant might mention both contextual benefits and cognitive facilitation in their response.



(a) Task 1 (b) Task 2

Figure 6.9: Spearman's ρ correlation coefficients for aspects in the main questionnaire.

statistical analysis to understand this relationship. For each task, we perform one-way ANOVA tests separately for with and without-image setups, followed by Tukey's HSD for post-hoc comparisons.

Task 1. The without-image setup (Figure 6.10a) reveals significant effects of background knowledge (F(2,97) = 8.45, $p \leq 0.05$). Post-hoc comparisons show significant differences between medium and high knowledge levels, with usefulness ratings dropping from 4.7 to 3.5 and satisfaction from 4.3 to 3.5. The with-image setup (Figure 6.10b) shows no significant differences across knowledge levels (F(2,97) = 2.13, $p = 0.12$), suggesting visual aids help maintain engagement regardless of expertise.

Task 2. For Task 2, without-image setup (Figure 6.10c), we find significant knowledge effects on both satisfaction (F(2,97) = 9.23, $p \leq 0.05$) and clarity (F(2,97) = 10.11, $p \leq 0.05$), with ratings decreasing substantially for higher expertise levels (clarity: 5.0 to 3.3; satisfaction: 4.0 to 3.1). While the with-image setup (Figure 6.10d) maintained significant effects (F(2,97) = 7.84, $p \leq 0.05$), the decline was less pronounced (clarity: 4.7 to 3.4; satisfaction: 4.3 to 3.2).

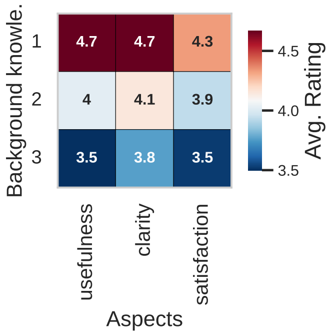
Summary. The consistency of these patterns across tasks, despite their different cognitive demands, strongly supports **H2**. The findings suggest that background knowledge plays a crucial role in how users perceive clarifying questions, with three key implications:

– Expert users generally find text-only clarifying questions less valuable, possibly due to their advanced understanding of the domain;

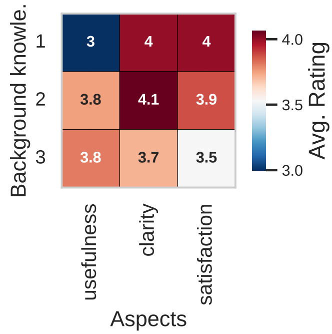
– Images serve as knowledge mediators, providing additional context that remains valuable even at higher expertise levels; and

– The moderating effect of images is particularly important in complex tasks like query reformulation, where the gap between novice and expert perceptions is largest.

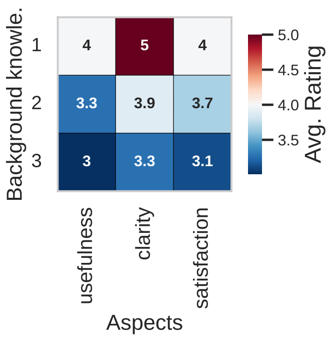
This suggests that multimodal clarifying questions may be crucial for creating universally effective search experiences, as they provide layered context that can be interpreted differently based on user expertise. The stronger moderating effect in Task 2



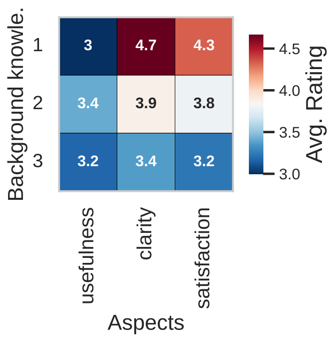
(a) Without-image



(b) With image



(c) Without-image



(d) With-image

Figure 6.10: Correlation of the average rating for usefulness, clarity, and satisfaction against the self-reported background knowledge of the participants. Task 1 row 1 and Task 2 row 2.

also indicates that visual support becomes more valuable as task complexity increases.

6.5 Retrieval Effectiveness

Building on our analysis of user interactions with multimodal clarifying questions, we examine whether visual enhancement impacts retrieval performance. Our evaluation focuses on whether the presence of images in clarifying questions leads to responses and reformulations that improve document retrieval.

We employ BM25 as our retrieval model for several key reasons. First, as a robust, training-free model for text-based ranking, it eliminates potential biases from training data. Second, by focusing on text-based retrieval, we avoid the complexity of aligning text and image features that multimodal models would require, allowing us to directly assess how visual elements influence user-generated text (responses and reformulations). Our experimental setup evaluates three query configurations:

Table 6.6: BM25 retrieval and ranking performance with different clarifying questions and answers.

Query	**Input**	**nDCG@1**	**nDCG@5**	**nDCG@10**
Original	Query only	0.059	0.067	0.074
	Query + QA	0.142	0.130	0.148
Reformulated	Query only	0.127	0.129	0.149
	Query + QA	**0.162**	**0.136**	**0.168**

Table 6.7: BM25 retrieval and ranking performance on with/without image queries with different clarifying questions and answers.

Query	**Input**	**nDCG@1**	**nDCG@5**	**nDCG@10**
Original	Query only	0.059 / 0.059	0.067 / 0.067	0.074 / 0.074
	Query + QA	0.079 / 0.185	0.120 / 0.152	0.138 / 0.172
Reformulated	Query only	**0.204** / 0.091	**0.155** / 0.120	**0.182** / 0.140

– Original query alone,

– Original query augmented with clarifying question-answer pairs, and

– User-provided reformulated queries

For each configuration, we retrieve the top 100 documents per facet and evaluate performance using nDCG@k ($k \in \{1, 5, 10\}$). Tables 6.6 and 6.7 present comparative results across with-image and without-image conditions.

6.5.1 Retrieval performance findings

Analysis of Tables 6.6 and 6.7 reveals four key patterns in retrieval performance:

Clarifying questions. Incorporating clarifying questions significantly enhances retrieval performance compared to using queries alone, consistent with findings from previous studies [7, 195, 238]. This improvement demonstrates the value of clarification in enriching search context.

Query reformulation. Reformulated queries consistently outperform original queries in retrieval effectiveness, aligning with recent research [167, 225]. This suggests that the reformulation process helps users better articulate their information needs.

Image impact on answers. The presence of images significantly influences answer characteristics and subsequent retrieval performance. Without images, users achieve higher retrieval scores (nDCG@1 0.185 vs. 0.079 in Original Query+QA), primarily because they provide more comprehensive textual responses to compensate for the absence of visual context. Conversely, with images, users tend to generate more specific answers that reference visual elements inaccessible to the text-based retrieval system.

Image impact on reformulations. Images substantially improve query reformulation quality, with nDCG@1 increasing from 0.091 to 0.204 in the Query-only condition. This improvement suggests that visual context helps users formulate more precise queries by providing concrete reference points and complementary contextual information for articulating search intent.

6.5.2 Impact of images on answer consistency

To understand how images influence user answer consistency, we analyzed performance distribution across five representative facets, selected to cover diverse query types. The results are shown in Figure 6.11a and 6.11b. The five selected facets are: *Facet 1*: *What specific health risks are associated with exposure to asbestos?*; *Facet 2*: *What home remedies are there for angular cheilitis?*; *Facet 3*: *What salary range does a land surveyor receive?*; *Facet 4*: *What were the results of the electoral college for the 2008 US presidential race?*; (5) *Facet 5*: *Is there a link between Barrett's Esophagus and cancer?*.

In most facets, images help reduce performance variability, leading to more consistent results across different answers. This effect is particularly evident in *Facet 2*, where the performance distribution becomes more centered and narrower with images. For instance, when users can see actual images of angular cheilitis rather than relying on text descriptions alone, they generate more consistent answers focused on remedies specifically suited to the visible symptoms.

However, certain facets show contrasting effects. For *Facet 3* and *Facet 4*, which involve specific numeric information (salary ranges and electoral results), the performance distribution with images shows greater variability. This wider distribution indicates that images can introduce distractions or unnecessary complexity for queries seeking precise, quantitative information, potentially leading to less consistent answers and retrieval outcomes.

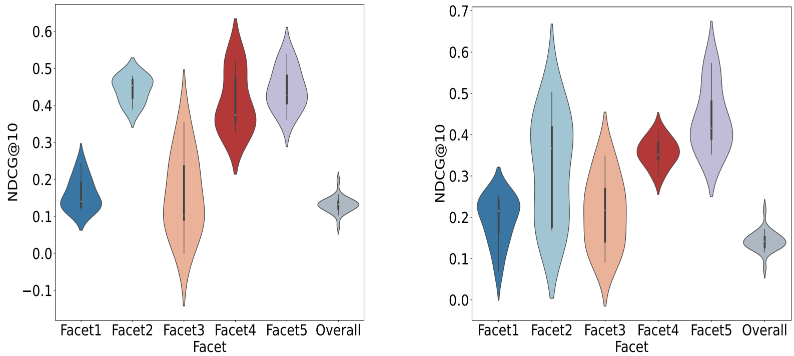
In summary, while images could help reduce response variability and improve retrieval consistency in some facets, their overall impact is mixed. This suggests that the effectiveness of images in retrieval is facet-dependent and should be determined in specific search scenarios.

6.5.3 Effect of reformulated queries on retrieval performance

We present examples of the best- and worst-performing reformulated queries for a specific facet, comparing their performance across both with-image and without-image queries.

Image-aligned reformulation. Visual information appears to guide users toward more effective query reformulation when the images directly relate to key search aspects. For example, in Table 6.8, when presented with images of ham radio antennas (row 3), users formulated queries that specifically referenced the visible antenna types, achieving better performance compared to the more general reformulation without visual context (row 4). This demonstrates how visual cues can help users incorporate relevant technical details into their queries.

Precision in reformulation. Our analysis reveals that precise, focused reformulations consistently outperform broader queries. In the angular cheilitis case (Table 6.9, rows 1–2), a specific query about non-prescription remedies achieved substantially higher performance (nDCG: 0.479) compared to a more general reformulation (nDCG: 0.390). This precision advantage appears particularly pronounced when images help



(a) With-image (b) Without-image

Figure 6.11: Performance distribution across facets with different answers.

users identify specific aspects of their information needs.

Facet-specific targeting. Query reformulations that directly address user needs show superior performance. The Norway spruce example (Table 6.9, row 3) illustrates this effect, where a focused query about planting and cultivation guidance (nDCG: 0.312) outperformed a broader, less targeted reformulation. Visual context appears to help users better align their reformulations with specific information needs rather than generating general-purpose queries.

These patterns suggest that visual enhancement can guide users toward more effective query reformulation strategies, particularly when images help users identify and incorporate specific, relevant details into their queries.

6.6 Analysis of Clarifying Answers and Reformulated Queries

We conduct qualitative analysis to explore the impact of image availability on: responses to clarifying questions (Task 1) and query reformulation (Task 2).

6.6.1 Task 1: Responses to clarifying questions

Our analysis of Task 1 reveals a profound impact of visual cues on users' responses to clarifying questions. The presence of images consistently leads to more specific, detailed, and confident answers across various domains.

A striking finding is the marked increase in response specificity when images are available. For a meteorological topic, e.g., when asked about cloud types, a user with access to an image responds, "I am seeing a stratus shaped cloud." This response not only demonstrates careful observation but also applies specific meteorological terminology. In contrast, without an image, another user answers the same question with a more general description: "They are long, thin clouds." This comparison illustrates how visual cues can trigger the use of domain-specific language and more precise observations.

We observe a notable increase in the comprehensiveness of responses when im-

Table 6.8: Best and worst performed a reformulated query for each facet (with image).

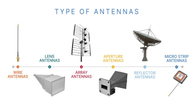
Facet	Ori. Query	Reformulated Query	Clarifying Question	Image	NDCG	
What salary range does a land surveyor receive?	land surveyor	How much would a land surveyor typically earn.	Are you interested in salary comparisons for land surveyors based on experience or geographic location?		0.354	👍
What salary range does a land surveyor receive?	land surveyor	What kind of remuneration does a land surveyor receive?	Are you interested in salary comparisons for land surveyors based on experience or geographic location?		0.00	👎
I need to learn more about ham radio antennas. What types are there, and how are they used in communication?	ham radio	What types of ham radio antennas are there and how does a facilitate communication?	Are you looking for information on the different designs of ham radio antennas and their specific applications?		0.305	👍
I need to learn more about ham radio antennas. What types are there, and how are they used in communication?	ham radio	Where can I learn about amateur radio antennas and their applications?	Are you looking for information on the different designs of ham radio antennas and their specific applications?		0.140	👎

Table 1

Typical Costs for a Sole-proprietor Business

Fixed costs	
Small office or part of the home	$1,500.00
Truck or vehicle(s)	$1,000.00
Equipment (field survey)	$1,700.00
Equipment (office)	$1,000.00
Insurance	$2,100.00
Office computers (3)	$ 750.00
Subtotal	$8,050.00
Variable costs	
Personnel salaries	$33,106.67
Fringe benefits	$ 8,276.67
Taxes, unemployment insurance	$13,242.66
Utilities	$250.00
Supplies	$ 550.00
Vehicle fuel and maintenance	$1,550.00
Legal and other professional fees	$ 650.00
Subtotal	$57,626.00
Total example expenses	**$65,676.00**

Table 6.9: Best and worst performed reformulated query for each facet (without image).

Facet	Ori. Query	Reformulated Query	Clarifying Question	NDCG	
What home remedies are there for angular cheilitis?	angular cheilitis	non-prescription remedies for angular cheilitis	What types of home remedies are you interested in, or do you have specific concerns related to angular cheilitis?	0.479	👍
What home remedies are there for angular cheilitis?	angular cheilitis	What is the best home treatment for Cheilitis?	What types of home remedies are you interested in, or do you have specific concerns related to angular cheilitis?	0.390	👎
I'm interested in gathering information on how to plant and cultivate Norway Spruce trees, including the best growth conditions and care tips.	norway spruce	Can you provide me with a guide to planting and cultivating Norway Spruce trees?	What particular aspects of planting and caring for Norway Spruce trees are you most focused on, such as growth conditions or maintenance tips?	0.312	👍
I'm interested in gathering information on how to plant and cultivate Norway Spruce trees, including the best growth conditions and care tips.	norway spruce	Where can I grow Norwegian spruce trees successfully in the world?	What particular aspects of planting and caring for Norway Spruce trees are you most focused on, such as growth conditions or maintenance tips?	0.178	👎

ages are absent, suggesting a compensatory strategy. For example, when asked about grilling techniques, an image-enhanced response is concise: "How to enhance flavours." Without an image, users tended to provide more elaborate answers, such as "I would like you to present a comprehensive list of vegetable grilling suggestions to enhance the flavour of vegetables." This pattern indicates that in the absence of visual cues, users attempt to fill the information gap with more detailed verbal descriptions.

Furthermore, the presence of visual cues leads to an increased display of confidence in responses. When queried about dinosaur types, a user with image access stated, "I am interested in a list of different kinds of dinosaurs." Without an image, responses are more tentative: "I'm wanting a list of the types/names of dinosaurs including images to accompany." This difference in tone and directness underscores the role of visual information in reinforcing user confidence during the clarification process.

The results from Task 1 reveal that: First, images serve as cognitive anchors, enabling users to access and articulate domain-specific knowledge more precisely, particularly in technical and medical domains. Second, without images, users adopt a compensatory strategy of providing more comprehensive textual descriptions, effectively recreating missing visual context and leading to improved retrieval performance. Third, visual support increases response confidence, suggesting that images help validate user understanding and potentially streamline early-stage information exchange.

6.6.2 Task 2: Query reformulation

The analysis of Task 2 uncovers distinct patterns in how visual cues influence query reformulation. The presence of images leads to more focused, action-oriented, and specific query refinements.

In the domain of automotive information, we observed a clear trend toward increased specificity in image-enhanced reformulations. An initial query about "Car dashboard symbols meaning" was refined to "Could you explain what the airbag indicator light means on a car dashboard?" when an image was present. Without visual cues, the reformulation remained broad: "what are the meaning of car dashboard symbols?" This demonstrates how visual information enables users to pinpoint specific aspects of their information needs during the reformulation process.

Visual cues also prompt a shift towards more action-oriented queries. E.g., an initial query about "grilling" evolved into "Vegetable grilling techniques to enhance flavour and presentation" with image assistance. Without visual cues, users tended towards more comprehensive information gathering: "I want some specific techniques when grilling that will help the taste, without ruining the visual appeal of the dish." This pattern suggests that visual information encourages users to focus on practical applications and specific actions.

The influence of visual cues on query scope is particularly noteworthy. In a query about Norway spruce trees, image-enhanced reformulation led to a focused request: "Can you provide me with a guide to planting and cultivating Norway Spruce trees?" Without images, users sought broader information: "What are the best growth conditions and maintenance tips for the planting and caring of Norway spruce trees?" This divergence in query scope highlights how visual cues can shape the breadth and depth of information seeking behavior.

Task 2 findings highlight four key effects of visual cues on query reformulation strategies. First, images guide users toward more specific and focused queries, potentially enabling more efficient search through precise scope definition. Second, visual information promotes action-oriented reformulations, suggesting that images help users conceptualize practical applications of their information needs. Third, users demonstrate compensatory behavior without images, but unlike Task 1's verbose descriptions, this manifests as broader query scope. Finally, visual support increases directness in query formulation, indicating that images enhance user confidence in articulating specific information needs.

Summary. Our analysis reveals the selective effectiveness of visual enhancement in CS: the impact of image availability varies with query type and domain. Technical and product-related queries showed substantially different patterns of specificity and detail between with- and without-image conditions, highlighting domains where visual support plays a crucial role. However, for general or conceptual questions, these differences were notably less pronounced, suggesting that visual enhancement provides selective rather than universal benefits. This domain and query-dependent pattern has important implications for the design of multimodal search systems, indicating that visual enhancement strategies should be adaptively deployed based on query characteristics rather than uniformly applied across all search scenarios.

6.7 Discussion and Implications

6.7.1 Discussion

In **RQ5.1** we ask *"How do images influence users' answers to clarifying questions in CS?"* Our findings reveal several key insights about the role of visual elements in CS interactions. First, despite the common assumption that visual aids enhance clarity, we found that text-only questions were perceived as marginally clearer. This unexpected finding suggests that integrating visual information may introduce additional complexity to the clarification process. However, this increased complexity does not necessarily detract from the overall user experience. Users consistently demonstrated a strong preference for multimodal questions (61.1%). The minimal increase in task completion time suggests efficient integration of visual information, indicating that any cognitive overhead from processing images was effectively balanced by their contextual benefits. This efficient integration of visual information aligns with dual coding theory [168], which suggests that parallel processing of visual and verbal information can be cognitively efficient.

The varying impact of images across different types of search tasks is particularly notable. The dominance of physical object queries (50%) among the scenarios in which images proved the most useful suggests that visual elements are particularly valuable when clarifying tangible or spatial characteristics. This finding extends previous research on visual information in search tasks [238] by identifying specific contexts where visual enhancement is most beneficial. Intriguingly, our results shed new light on how visual elements influence users with different levels of domain expertise. We observed that user background knowledge had less impact on their performance with multimodal questions, suggesting that image-enhanced questions provide enough context to enable users with different background knowledge equally. This finding contributes to our understanding of expertise-based differences in search behavior [163] and highlights the potential of visual elements to create more adaptive and inclusive search experiences.

While users report positive experiences with image-enhanced interactions in certain contexts, this did not consistently translate to improved performance. Higher user satisfaction with images did not necessarily correlate with better retrieval outcomes of their answers. For clarifying question answers, despite the positive reception of images by users, responses without images achieved better retrieval scores (nDCG@1 0.185 vs. 0.079), probably because users provided more comprehensive textual information when compensating for the lack of visual context, and the fact that our retrieval model did not include visual embeddings into account.

In **RQ5.2** we examine "*What effect do images have on query reformulation in CS?*" We show that images play a distinct and complex role in query reformulation. While all users significantly expanded their queries (from 3.13 to 12.46 words on average), those with access to images produced slightly more concise reformulations. This suggests that visual information helps users focus their information needs more precisely, potentially leading to more efficient search strategies. This finding extends previous work on query reformulation [99] by demonstrating how visual elements can help streamline query modification. However, the small differences in usefulness and satisfaction suggest that images are not always perceived as universally beneficial during query reformulation, indicating that images can either enhance or distract users, depending on the task at hand. For some participants, visual cues may provide the additional context needed to clarify their initial queries, while for others, this additional information may introduce unnecessary complexity or cognitive load.

Interestingly, the low frequency of direct image usage in questions where images are not explicitly required suggests that users tend to rely on visual information only when they perceive it to be necessary. This indicates that users favor visual content when they feel it adds substantial value. Participants were less likely to refer to their explicitly stated information needs when images were present. Images provide an implicit form of context that enables users to internalize and restructure their information needs without needing to explicitly recall them. This phenomenon aligns with Lu et al. [146]'s theory of cognitive offloading, where users rely on external tools, such as images, to simplify the mental processes required for query reformulation.

In **RQ5.3** we investigate *"When are images useful in query clarification?"* Our investigation of image utility in CS, combining situational analysis and thematic coding of user responses, reveals three critical dimensions that determine the effectiveness of visual elements in clarifying questions:

First, query characteristics emerge as a primary determinant, supported by both our situational and thematic analyses. Images demonstrate clear utility for queries with inherent visual attributes, particularly those involving physical objects (approximately 50% of beneficial cases) and procedural information (14–19%). Our thematic analysis reveals how this utility manifests through contextual support (28.4%) and cognitive facilitation (24.3%), especially when visual elements align closely with query intent. However, for abstract queries, images can create cognitive overload (13.5%) or become irrelevant to task goals (10.8%), suggesting that visual enhancement decisions should carefully consider query type (Section 6.4.6). This finding extends previous work on multimodal search clarification [238] by specifically identifying which query types benefit most from visual enhancement in conversational contexts.

Second, the type of search task significantly moderates image effectiveness, influencing how users use visual information. While direct question answering benefits from images through creative stimulation (21.6%) and increased engagement (16.2%), query reformulation shows a more complex relationship with visual elements, sometimes leading to redundancy (8.1%). This task dependency suggests that mixed-initiative systems need to consider not only whether to include images, but when in the conversation they would be most beneficial.

Third, user expertise plays a crucial role in image utility in our thematic findings. Expert users demonstrate particularly effective use of visual information when present, showing an enhanced ability to integrate visual cues into their search strategies. Although few users (9.5%) expressed preference for text-only interactions, our analysis shows that the level of expertise significantly influences how effectively visual information is used.

6.7.2 Implications and limitations

Our findings reveal requirements for designing effective multimodal CS systems. First, systems need to implement adaptive visual enhancement strategies based on task characteristics. Our results demonstrate distinct patterns of image utility between answering clarifying questions and query reformulation tasks, suggesting the need for mechanisms that dynamically assess when to present visual information based on the current search stage and task requirements.

Second, user expertise significantly influences image effectiveness, indicating the need for expertise-aware visual support. Systems should adapt their visual strategy based on user expertise signals, potentially varying both the complexity and quantity of visual information. This could involve developing presentation strategies for novice users who benefit from basic visual context versus expert users who can leverage more complex visual information.

Third, query characteristics should drive image deployment decisions. The clear correlation between query type and image utility suggests implementing automatic assessment of query visual dependency. Systems should prioritize visual enhancement for queries with inherent visual attributes (e.g., physical objects, processes) while limiting visual elements for abstract queries.

However, several limitations should be considered when interpreting these implications. Our study used a curated subset of ClariQ dataset topics, which may not fully represent the diversity of real-world search scenarios. The controlled nature of our user study, while enabling systematic analysis, used pre-defined queries rather than natural CS interactions. Additionally, while we employed systematic image selection, we did not explore how different image characteristics might affect search interaction. Given the complex nature of visual elements, a deeper study of the images exhibiting specific visual characteristics can complement our work. Our work, however, provides important insights into this task and the role of images on user's perceptions that can inform future studies.

6.8 Conclusion

In this chapter, we have investigated the effect of images on CS, through a user study examining their role in answering clarifying questions and query reformulation. Our findings reveal that while images generally enhance the search experience, their impact varies significantly based on search task, user expertise, and question characteristics. Across tasks, images are found to enhance clarity and usefulness in specific contexts, particularly when users seek information about physical objects or processes. However, even though images can help users refine their queries and improve search efficiency, their utility is task-dependent. Visual aids are most effective when clarifying tangible items and complex procedures, but less so for abstract concepts and context-related queries.

Our work contributes to the growing understanding of multimodal interaction in CS by providing empirical evidence of when and how images can improve search outcomes. Future work should explore how to optimize the integration of images into real-world search systems and investigate the long-term effects of multimodal interfaces on user satisfaction and performance.

In this chapter we addressed **RQ5:** *How do images in clarifying questions affect user performance and preferences in conversational search across different tasks and user expertise levels?* Our findings show that integrating images into clarifying questions affects user interaction in CS. While images improve user understanding in tasks requiring visual interpretation, their effectiveness depends on user expertise and query type. Novice users benefit more from visual support, while expert users may find text-based clarifications sufficient. Additionally, image usefulness varies across different query types, indicating the need for adaptive strategies in multimodal conversational systems.

This chapter concludes this thesis, progressing from evaluation methodology to system enhancement, establishing frameworks for assessing and improving conversational interactions. In the next chapter, we conclude this thesis and discuss our main findings, the limitations, and future research directions.

7
Conclusions

This thesis has investigated the evaluation and enhancement of conversational systems, focusing on understanding user satisfaction, improving evaluation methodologies, and examining the role of clarifying questions in both text and multimodal contexts. Our work addresses five research questions organized around two central themes: understanding the evaluation of task-based conversational systems and advancing clarification in conversational search.

7.1 Research Findings and Implications

Theme 1: Understanding the Evaluation of Task-Based Conversational Systems

Our studies on the evaluation task-based conversational systems yielded several key insights across three research questions:

RQ1 Which dialogue aspects influence user satisfaction in a conversational recommender system, and can we effectively predict user satisfaction using these dialogue aspects?

To understand what influences user satisfaction in conversational recommender systems (CRSs), we defined six quantifiable dialogue aspects (i.e., relevance, interestingness, understanding, task completion, interest arousal, and efficiency) to capture different dimensions of conversation quality. We conducted an annotation study of 200 dialogues in which annotators evaluated these aspects and provided satisfaction ratings at both the turn and dialogue levels. Using these annotated dialogues, we analyzed the relationship between dialogue aspects and satisfaction levels, then developed predictive models to leverage multiple aspects to assess both turn-level response quality and overall dialogue satisfaction. Our analysis revealed that users get satisfied differently at the turn and dialogue levels. Although utterance relevance drives immediate turn-level satisfaction, the overall dialogue satisfaction depends more on the system's ability to arouse the user's interest in novel recommendations and complete the user's task. Furthermore, our models demonstrated that combining multiple dialogue aspects yielded stronger predictive performance than using single aspects or turn-level satisfaction alone, confirming the multifaceted nature of user satisfaction in CRSs.

RQ2 What is the effect of dialogue context on crowdsourced evaluation labels in task-oriented dialogue systems?

In Chapter 3, we addressed the research question by conducting a crowdsourcing study to investigate how different amounts of dialogue context affect the quality and consistency of evaluation labels, focusing on relevance and usefulness. The study featured three setups: no previous dialogue context, partial previous dialogue context, and full dialogue context. We show there is no one-size-fits-all solution, as optimal context depends on the aspects being evaluated. Therefore, we propose to use an LLM-as-an-assistant approach to generating the supplementary context in the form of user information needs and dialogue summaries. These automatically generated contexts enhanced annotator agreement in no-context conditions and reduced annotation time compared to providing full dialogue context. These findings demonstrate how task design, especially context presentation, can significantly impact the quality of crowdsourced evaluations. Additionally, they highlight the potential of LLMs to optimize the trade-off between annotation effort and reliability, offering practical solutions to improve evaluation workflows.

RQ3 How does incorporating user feedback through follow-up utterances affect evaluation judgments by humans and LLMs, and what does this reveal about their respective strengths as annotators?

To answer **RQ3,** we investigated how the user's follow-up utterance influences the quality of task-oriented dialogue system (TDS) evaluation labels. Through a comparative study, we examined two evaluation conditions: assessors evaluating system responses (i) with and (ii) without access to the user's next utterance. Both crowdworkers and large language models (LLMs) evaluated responses across four aspects: relevance, usefulness, interestingness, and explanation quality. We show that human annotators relied heavily on user feedback when assessing usefulness and interestingness, conditioning their ratings based on user responses. Thus improving the annotator agreement, particularly for ambiguous user requests. LLMs maintained more stable ratings across both conditions, performing well on relevance assessment but showing limited adaptation to user feedback signals. These findings highlight the distinct capabilities of human and LLM assessors, where humans excel at incorporating user feedback for subjective assessments, while LLMs provide consistent objective evaluations.

Theme 2: Advancing Clarification in Conversational Search

Our work on clarifying questions in conversational search led to several insights across the two research questions:

RQ4 How effectively can large language models generate and evaluate clarifying questions for conversational search systems?

To address **RQ4**, we proposed AGENT-CQ (Automatic GENeration, and evaluaTion of Clarifying Questions), an end-to-end LLM framework for generating and evaluating clarifying questions in conversational search (CS). The framework consists of two main components. First, a generation stage that employs two distinct approaches:

(i) a facet-based method that explicitly models different query interpretations and (ii) a temperature-variation approach (GPT-Temp) that systematically adjusts LLM temperature to produce diverse questions. This stage includes filtering mechanisms and simulates user responses using parameterized LLM generation. Our experiments on the ClariQ dataset reveal that GPT-Temp consistently outperforms baselines in generating high-quality clarifying questions, achieving the highest nDCG@1 scores (0.225 for BM25, 0.312 for BERT). However, human-generated questions performed better in BM25 retrieval at nDCG@5 (0.221) and nDCG@10 (0.246), likely due to better term overlap with original queries. For answer simulation, our parametric approach generates responses that match or slightly outperform human answers in quality assessments, though they show lower retrieval performance when paired with human questions. The evaluation stage, CrowdLLM, demonstrates strong capability in simulating crowdsourced assessment across multiple quality metrics. However, we observe potential biases: while human assessors slightly favor LLM answers (34% vs. 32%), CrowdLLM shows a stronger preference (55.16% vs. 37.74%). While GPT-Temp's low complexity makes it ideal for general purpose clarification tasks, facet-based approaches might better suit specialized domains requiring detailed clarifications. Our results demonstrate that LLMs can effectively automate both the generation and evaluation of clarifying questions with GPT-Temp achieving better retrieval performance and CrowdLLM showing strong alignment with expert judgments.

RQ5 How do images in clarifying questions affect user performance and preferences in conversational search across different tasks and user expertise levels?

Through a crowdsourced user study with 73 participants, we examined the role of visual elements in clarifying questions for conversational search to answer **RQ5**. The study compared multimodal and text-only variants across two fundamental tasks: (i) answering clarifying questions and (ii) reformulating queries. Participants engaged with both formats, allowing us to analyze how images affect user performance and preferences. We show that users strongly prefer multimodal setups when answering clarifying questions, though text-only setups led to more comprehensive responses and better retrieval performance. This difference occurs because users provide more detailed textual information when compensating for the absence of visual context. In the query reformulation experiment, images helped users construct more precise queries, particularly when searching for physical objects or processes, substantially improving retrieval performance. This improvement suggests that visual context provides concrete reference points for articulating search intent. However, effectiveness varied with both user-reported expertise level and query type.

In summary, the findings from theme 1 and theme 2 advance our understanding of the evaluation and enhancement of conversational systems in several key directions:

1. We demonstrated that effective evaluation requires both turn-level and dialogue-level assessment, with different aspects influencing user satisfaction at each interaction level (Chapter 2)
2. We established how dialogue context and user feedback influence evaluation quality, showing that LLMs excel at objective assessments while humans better capture subjective aspects (Chapters 3 and 4)

3. We developed scalable approaches for generating and evaluating clarifying questions using LLMs while identifying task-specific benefits of incorporating visual elements in conversational interfaces (Chapters 5 and 6)

7.2 Future Research Directions

Our research in this thesis reveals several important limitations that shape future research. In evaluating user satisfaction (Chapter 2), we relied on external assessors rather than direct user feedback, limiting our understanding of personal satisfaction factors. Our studies on dialogue context and user feedback (Chapters 3 and 4) showed that while LLMs excel at objective assessments, they struggle with subjective aspects and adapting to user-specific preferences. The AGENT-CQ framework (Chapter 5) demonstrated strong performance but revealed potential biases of LLMs toward machine-generated content. Our multimodal study (Chapter 6) was limited to specific types of search tasks and user groups.

These limitations, combined with the evolving landscape of conversational AI, lead us to three future research directions: the interplay between human and machine evaluation, context-aware conversational systems, and diverse and representative evaluation methods.

Below, we detail these research directions.

7.2.1 Hybrid evaluation frameworks

In Chapters 2, 3, and 4, we relied on human assessors, which, while effective for capturing subjective insights, face scalability issues and potential demographic biases influenced by age, gender, and cultural background [69, 117]. LLMs offer scalable and cost-effective alternatives for evaluation [94, 130, 180, 204], yet they have notable drawbacks. For example, they may overlook the cultural context in subjective evaluations, reinforce existing biases present in their training data, or lack the specialized expertise required for certain fields [44]. Their decision-making is often unclear, hindering interpretability and accountability in high-stakes applications. Furthermore, LLM evaluations are sensitive to prompt wording and formatting, leading to inconsistent and irreproducible results [180].

Our findings in Chapter 4 highlight the complementary strengths of humans and LLMs. Humans excel at subjective evaluations that require user-centric insights, while LLMs provide a consistent and efficient evaluation of objective criteria. This suggests the potential for hybrid frameworks that strategically combine both approaches. For example, using LLMs for scalable assessments such as relevance scoring while reserving subjective evaluations for human assessors. In addition, ensuring human accountability in high-stake domains like health care, where LLM errors can be detrimental, is essential. Several technical challenges need to be addressed in the development of effective hybrid frameworks. These include optimal task allocation (determining which aspects and domain should be evaluated by humans versus LLMs) [132] and disagreement resolution (developing methods to handle conflicting assessments between human and automated assessors).

Furthermore, research questions remain about LLM-based evaluation itself, such as methods for sampling diverse opinions from LLMs, techniques for simulating demographic diversity through persona assignment, and approaches to a robust evaluation of multiturn context-rich interactions. The ongoing LLM4Eval workshop [179] series at SIGIR and WSDM has begun to explore these challenges through discussions of hybrid models, bias mitigation strategies, and scalable metrics, laying the foundation for future research.

7.2.2 Context-aware conversational systems

Our research revealed important limitations in how current conversational systems adapt to user needs and contexts. In Chapter 4, we show that follow-up utterances contain valuable signals about interaction quality that systems could use for real-time adaptation. Chapter 6 demonstrated that the effectiveness of multimodal elements significantly depends on user expertise and query type, challenging the prevalent one-size-fits-all approach to visual presentation. Importantly, we found that while visual elements benefit some users in specific contexts, they can be unnecessary or even detrimental to others. These findings highlight two research directions.

First, we need frameworks that can effectively leverage real-time user feedback for system adaptation. Building on Bruyn et al. [35]'s work on implicit feedback detection, future research should explore how to detect and respond to evolving satisfaction signals while maintaining coherent conversations. A key challenge lies in balancing immediate adaptation with long-term interaction coherence.

Second, our findings on visual elements [205, 238] demonstrate the need for personalized multimodal interaction. Rather than applying visual elements uniformly, systems should dynamically assess when and how to incorporate visual information based on user expertise, task requirements, and interaction context.

7.2.3 Diverse and representative evaluation methods

A major limitation in conversational systems, including the work presented in this thesis, lies in the dominance of benchmarks focused on English and other resource-rich languages, such as GLUE [220] and XTREME [98]. This linguistic and cultural bias restricts the generalizability of models across diverse populations, leaving underrepresented languages and communities without equitable access to conversational technologies. Furthermore, methods like reinforcement learning from human feedback (RLHF) [166] rely on data and user preferences sourced from narrow demographics, often excluding perspectives from low-resource and marginalized groups. Future research must address how benchmarks can better represent linguistic diversity, particularly languages with unique typologies, such as tonal (e.g., Yoruba) or agglutinative (e.g., Swahili) languages. How can benchmarks capture such diversity while ensuring fair evaluation across different linguistic structures?

Similarly, the data collection strategies for RLHF and preference optimization must be revisited to include global user signals or more diverse local signals. Questions arise about how cultural and regional differences shape user preferences and how diverse feedback can be integrated without skewing model optimization toward overrepresented groups. Initiatives like Masakhane[1] have created a path to address some of these challenges through the creation of participatory community datasets [2, 3, 199], African-centric evaluation metrics [221], and models [6, 164].

[1] https://www.masakhane.io/

Bibliography

[1] Z. Abbasiantaeb, Y. Yuan, E. Kanoulas, and M. Aliannejadi. Let the LLMs talk: Simulating human-to-human conversational QA via zero-shot LLM-to-LLM interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024, Merida, Mexico*, pages 8–17. ACM, 2024. URL `https://doi.org/10.1145/3616855.3635856`. (Cited on pages 91 and 93.)

[2] D. I. Adelani, J. Abbott, G. Neubig, D. D'souza, J. Kreutzer, C. Lignos, C. Palen-Michel, H. Buzaaba, S. Rijhwani, S. Ruder, S. Mayhew, I. A. Azime, S. H. Muhammad, C. C. Emezue, J. Nakatumba-Nabende, P. Ogayo, A. Anuoluwapo, C. Gitau, D. Mbaye, J. Alabi, S. M. Yimam, T. R. Gwadabe, I. Ezeani, R. A. Niyongabo, J. Mukiibi, V. Otiende, I. Orife, D. David, S. Ngom, T. Adewumi, P. Rayson, M. Adeyemi, G. Muriuki, E. Anebi, C. Chukwuneke, N. Odu, E. P. Wairagala, S. Oyerinde, C. Siro, T. S. Bateesa, T. Oloyede, Y. Wambui, V. Akinode, D. Nabagereka, M. Katusiime, A. Awokoya, M. MBOUP, D. Gebreyohannes, H. Tilaye, K. Nwaike, D. Wolde, A. Faye, B. Sibanda, O. Ahia, B. F. P. Dossou, K. Ogueji, T. I. DIOP, A. Diallo, A. Akinfaderin, T. Marengereke, and S. Osei. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021. URL `https://aclanthology.org/2021.tacl-1.66/`. (Cited on page 154.)

[3] D. I. Adelani, M. Masiak, I. A. Azime, J. O. Alabi, A. L. Tonja, C. Mwase, O. Ogundepo, B. F. P. Dossou, A. Oladipo, D. Nixdorf, C. C. Emezue, S. S. Al-Azzawi, B. K. Sibanda, D. David, L. Ndolela, J. Mukiibi, T. Ajayi, T. M. Ngoli, B. Odhiambo, A. T. Owodunni, N. C. Obiefuna, M. Mohamed, S. H. Muhammad, T. M. Ababu, S. A. Salahudeen, M. G. Yigezu, T. Gwadabe, I. Abdulmumin, M. T. Bame, O. O. Awoyomi, I. Shode, T. A. Adelani, H. Abdulganiyu, A. Omotayo, A. Adeeko, A. Abeeb, A. Anuoluwapo, S. Olanrewaju, C. Siro, W. Kimotho, O. R. Ogbu, C. E. Mbonu, C. Chukwuneke, S. Fanijo, J. Ojo, O. Awosan, T. K. Guge, S. T. Sari, P. Nyatsine, F. Sidume, O. Yousuf, M. Oduwole, K. P. Tshinu, U. Kimanuka, T. Diko, S. Nxakama, S. G. Nugussie, A. T. Johar, S. A. Mohamed, F. M. Hassan, M. A. Mehamed, E. Ngabire, J. Jules, I. Ssenkungu, and P. Stenetorp. Masakhanews: News topic classification for african languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali*, pages 144–159. Association for Computational Linguistics, 2023. URL `https://doi.org/10.18653/v1/2023.ijcnlp-main.10`. (Cited on page 154.)

[4] A. Al-Maskari and M. Sanderson. A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology*, 61(5):859–868, 2010. URL `https://doi.org/10.1002/asi.21300`. (Cited on pages 1 and 16.)

[5] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between ir effectiveness measures and user satisfaction. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherland*, page 773–774. ACM, 2007. URL `https://doi.org/10.1145/1277741.1277902`. (Cited on pages 11 and 49.)

[6] J. O. Alabi, D. I. Adelani, M. Mosbach, and D. Klakow. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea*, pages 4336–4349. International Committee on Computational Linguistics, 2022. URL `https://aclanthology.org/2022.coling-1.382`. (Cited on page 154.)

[7] M. Aliannejadi, H. Zamani, F. Crestani, and W. B. Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France*, pages 475–484. ACM, 2019. URL `https://doi.org/10.1145/3331184.3331265`. (Cited on pages 66, 77, 89, 91, 115, 116, 118, 119, and 138.)

[8] M. Aliannejadi, J. Kiseleva, A. Chuklin, J. Dalton, and M. Burtsev. ConvAI3: Generating clarifying questions for open-domain dialogue systems (ClariQ). *CoRR*, abs/2009.11352, 2020. URL `https://arxiv.org/abs/2009.11352`. (Cited on pages 5, 89, and 118.)

[9] M. Aliannejadi, L. Azzopardi, H. Zamani, E. Kanoulas, P. Thomas, and N. Craswell. Analysing mixed initiatives and search strategies during conversational search. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia*, pages 16–26. ACM, 2021. URL `https://doi.org/10.1145/3459637.3482231`. (Cited on pages 92, 118, and 119.)

[10] M. Aliannejadi, J. Kiseleva, A. Chuklin, J. Dalton, and M. Burtsev. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP 2021, Online and Punta Cana, Dominican Republic*, pages 4473–4484. Association for Computational Linguistics, 2021. URL `https://aclanthology.org/2021.emnlp-main.367`. (Cited on pages 91, 92, 94, 100, 105, 118, 120, and 122.)

[11] J. Allen, C. I. Guinn, and E. Horvitz. Mixed-initiative interaction. *IEEE Intelligent Systems & Their Applications*, 14, 1999. URL `https://doi.org/10.1109/5254.796083`. (Cited on page 118.)

[12] J. F. Allen, G. Ferguson, and A. Stent. An architecture for more realistic conversational systems. In *Proceedings of the 6th International Conference on Intelligent User Interfaces, IUI 2001, Santa Fe, NM, USA*, pages 1–8. ACM, 2001. URL `https://doi.org/10.1145/359784.359822`. (Cited on page 1.)

[13] O. Alonso and R. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland*, volume 6611 of *Lecture Notes in Computer Science*, pages 153–164. Springer, 2011. URL `https://doi.org/10.1007/978-3-642-20161-5_16`. (Cited on page 18.)

[14] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42 (2):9–15, 2008. URL `https://doi.org/10.1145/1480506.1480508`. (Cited on pages 49, 50, and 70.)

[15] Amazon Mechanical Turk. `https://www.mturk.com`, 2023. (Cited on pages 52 and 69.)

[16] A. Anand, L. Cavedon, M. Hagen, H. Joho, M. Sanderson, and B. Stein. Conversational search - A report from dagstuhl seminar 19461. *CoRR*, abs/2005.08658, 2020. URL `https://arxiv.org/abs/2005.08658`. (Cited on pages 1 and 16.)

[17] A. Askari, M. Aliannejadi, C. Meng, E. Kanoulas, and S. Verberne. Expand, highlight, generate: RL-driven document generation for passage reranking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore*, pages 10087–10099. Association for Computational Linguistics, 2023. URL `https://doi.org/10.18653/v1/2023.emnlp-main.623`. (Cited on page 90.)

[18] S. Avula, B. Choi, and J. Arguello. The effects of system initiative during conversational collaborative search. *Proc. ACM Hum. Comput. Interact.*, 6(CSCW1):66:1–66:30, 2022. URL `https://doi.org/10.1145/3512913`. (Cited on page 119.)

[19] L. Azzopardi. Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia*, pages 27–37. ACM, 2021. URL `https://doi.org/10.1145/3406522.3446023`. (Cited on page 68.)

[20] L. Azzopardi, M. Aliannejadi, and E. Kanoulas. Towards building economic models of conversational search. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 31–38. Springer, 2022. URL `https://doi.org/10.1007/978-3-030-99739-7_4`. (Cited on page 119.)

[21] K. Balog and F. Radlinski. Measuring recommendation explanation quality: The conflicting goals of explanations. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China*, pages 329–338. ACM, 2020. URL `https://doi.org/10.1145/3397271.3401032`. (Cited on pages 69 and 70.)

[22] K. Balog and C. Zhai. User simulation for evaluating information access systems. *Found. Trends Inf. Retr.*, 18(1-2):1–261, 2024. URL `https://doi.org/10.1561/1500000098`. (Cited on page 39.)

[23] T. Baltrusaitis, C. Ahuja, and L. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019. URL `https://doi.org/10.1109/TPAMI.2018.2798607`. (Cited on page 2.)

[24] D. L. Barbera, K. Roitero, G. Demartini, S. Mizzaro, and D. Spina. Crowdsourcing truthfulness: The impact of judgment scale and assessor bias. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, Part II*, volume 12036 of *Lecture Notes in Computer Science*, pages 207–214. Springer, 2020. URL `https://doi.org/10.1007/978-3-030-45442-5_26`. (Cited on page 69.)

[25] J. N. Belkin. Anomalous states of knowledge in information retrieval. *Canadian journal of information science*, 5:133–143, 1980. (Cited on page 1.)

[26] E. Bobek and B. Tversky. Creating visual explanations improves learning. 2014. URL https://escholarship.org/uc/item/3gh2z11f. (Cited on page 120.)

[27] P. K. Bodigutla, L. Polymenakos, and S. Matsoukas. Multi-domain conversation quality evaluation via user satisfaction estimation. *CoRR*, abs/1911.08567, 2019. URL http://arxiv.org/abs/1911.08567. (Cited on pages 17, 30, and 68.)

[28] P. K. Bodigutla, A. Tiwari, S. Matsoukas, J. Valls-Vargas, and L. Polymenakos. Joint turn and dialogue level user satisfaction estimation on multi-domain conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online*, pages 3897–3909. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.findings-emnlp.347. (Cited on pages 16, 31, 37, and 68.)

[29] M. Boguslav and K. B. Cohen. Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing. In *MEDINFO 2017: Precision Healthcare through Informatics - Proceedings of the 16th World Congress on Medical and Health Informatics, Hangzhou, China*, volume 245 of *Studies in Health Technology and Informatics*, pages 298–302. IOS Press, 2017. URL https://doi.org/10.3233/978-1-61499-830-3-298. (Cited on page 54.)

[30] A. Bouyamourn. Why LLMs hallucinate, and how to get (evidential) closure: Perceptual, intensional, and extensional learning for faithful natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore*, pages 3181–3193. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.emnlp-main.192. (Cited on page 51.)

[31] P. Braslavski, D. Savenkov, E. Agichtein, and A. Dubatovka. What do you mean exactly?: Analyzing clarification questions in CQA. In *CHIIR '17: ACM SIGIR Conference on Human Information Interaction and Retrieval, Oslo, Norway*, pages 345–348. ACM, 2017. URL https://doi.org/10.1145/3020165.3022149. (Cited on pages 97, 110, 111, and 118.)

[32] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. URL https://doi.org/10.1023/A:1010933404324. (Cited on page 31.)

[33] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984. ISBN 0-534-98053-8. (Cited on page 31.)

[34] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html. (Cited on pages 72 and 105.)

[35] M. D. Bruyn, E. Lotfi, J. Buhmann, and W. Daelemans. Open-domain dialog evaluation using follow-ups likelihood. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea*, pages 496–504. International Committee on Computational Linguistics, 2022. URL https://aclanthology.org/2022.coling-1.40. (Cited on page 153.)

[36] P. Budzianowski and I. Vulić. Hello, it's GPT-2 - How can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22, Hong Kong, 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-5602. (Cited on page 47.)

[37] A. Burmania, S. Parthasarathy, and C. Busso. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Trans. Affect. Comput.*, 7(4):374–388, 2016. URL https://doi.org/10.1109/TAFFC.2015.2493525. (Cited on page 2.)

[38] W. Cai and L. Chen. Predicting user intents and satisfaction with dialogue-based conversational recommendations. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 33–42. ACM, 2020. URL https://doi.org/10.1145/3340631.3394856. (Cited on pages 16 and 17.)

[39] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996. URL https://aclanthology.org/J96-2004. (Cited on page 54.)

[40] G. Casella and R. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002. URL https://books.google.nl/books?id=0x_

vAAAAMAAJ. (Cited on page 28.)
[41] C. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria*. OpenReview.net, 2024. URL https://openreview.net/forum?id=FQepisCUWu. (Cited on page 2.)
[42] Y. Chang, G. Cao, M. Narang, J. Gao, H. Suzuki, and Y. Bisk. Webqa: Multihop and multimodal QA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA*, pages 16474–16483. IEEE, 2022. URL https://doi.org/10.1109/CVPR52688.2022.01600. (Cited on page 118.)
[43] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie. A survey on evaluation of large language models. *CoRR*, abs/2307.03109, 2023. URL https://doi.org/10.48550/arXiv.2307.03109. (Cited on pages 51, 61, and 69.)
[44] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), Mar. 2024. ISSN 2157-6904. URL https://doi.org/10.1145/3641289. (Cited on page 152.)
[45] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He. Bias and debias in recommender system: A survey and future directions. *ACM Trans. Inf. Syst.*, 41(3):67:1–67:39, 2023. URL https://doi.org/10.1145/3564284. (Cited on page 81.)
[46] L. Chen and P. Pu. Critiquing-based recommenders: Survey and emerging trends. *User Modeling and User-Adapted Interaction*, 22(1):125–150, 2012. (Cited on page 15.)
[47] S. Chen, M. Wu, K. Q. Zhu, K. Lan, Z. Zhang, and L. Cui. LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *CoRR*, abs/2305.13614, 2023. URL https://doi.org/10.48550/arXiv.2305.13614. (Cited on page 91.)
[48] J. I. Choi, A. Ahmadvand, and E. Agichtein. Offline and online satisfaction prediction in open-domain conversational systems. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China*, pages 1281–1290. ACM, 2019. URL https://doi.org/10.1145/3357384.3358047. (Cited on page 68.)
[49] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore*, pages 659–666. ACM, 2008. URL https://doi.org/10.1145/1390334.1390446. (Cited on page 115.)
[50] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 web track. In *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA*, volume 500-278 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2009. URL http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf. (Cited on page 105.)
[51] C. W. Cleverdon. The Cranfield tests on index language devices. In *Aslib Proceedings*, volume 19, pages 173–192, 1967. (Cited on page 11.)
[52] C. W. Cleverdon. The significance of the cranfield tests on index languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, (Special Issue of the SIGIR Forum)*, pages 3–12. ACM, 1991. URL https://doi.org/10.1145/122860.122861. (Cited on page 11.)
[53] C. W. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. In *ASLIB Cranfield research project*. 1966. (Cited on page 65.)
[54] J. Cohen. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. URL https://10.1037/h0026256. (Cited on page 95.)
[55] K. Collins-Thompson, C. Macdonald, P. N. Bennett, F. Diaz, and E. M. Voorhees. TREC 2014 web track overview. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA*, volume 500-308 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2014. URL http://trec.nist.gov/pubs/trec23/papers/overview-web.pdf. (Cited on page 118.)
[56] W. S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37, 1971. URL https://doi.org/10.1016/0020-0271(71)90024-6. (Cited on page 18.)

[57] A. Dagan, I. Guy, and S. Novgorodov. An image is worth a thousand terms? analysis of visual e-commerce search. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada*, pages 102–112. ACM, 2021. URL https://doi.org/10.1145/3404835.3462950. (Cited on page 116.)

[58] Y. Deng, L. Liao, L. Chen, H. Wang, W. Lei, and T. Chua. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore*, pages 10602–10621. Association for Computational Linguistics, 2023. URL https://doi.org/10.18653/v1/2023.findings-emnlp.711. (Cited on page 91.)

[59] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation WMT 2014, Baltimore, Maryland, USA*, pages 376–380. Association for Computational Linguistics, 2014. URL https://aclanthology.org/W14-3348. (Cited on page 12.)

[60] J. Deriu, Á. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak. Survey on evaluation methods for dialogue systems. *Artif. Intell. Rev.*, 54(1):755–810, 2021. URL https://doi.org/10.1007/s10462-020-09866-x. (Cited on pages 2, 15, 16, 19, 47, 49, 68, and 89.)

[61] W. Di, N. Sundaresan, R. Piramuthu, and A. Bhardwaj. Is a picture really worth a thousand words?: – on the role of images in e-commerce. In *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA*, pages 633–642. ACM, 2014. URL https://doi.org/10.1145/2556195.2556226. (Cited on page 116.)

[62] D. E. Difallah, E. Filatova, and P. Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA*, pages 135–143. ACM, 2018. URL https://doi.org/10.1145/3159652.3159661. (Cited on page 49.)

[63] B. Ding, C. Qin, L. Liu, Y. K. Chia, B. Li, S. Joty, and L. Bing. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada*, pages 11173–11195. Association for Computational Linguistics, 2023. URL https://doi.org/10.18653/v1/2023.acl-long.626. (Cited on pages 2 and 69.)

[64] N. Ding, Y. Chen, B. Xu, Y. Qin, S. Hu, Z. Liu, M. Sun, and B. Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore*, pages 3029–3051. Association for Computational Linguistics, 2023. URL https://doi.org/10.18653/v1/2023.emnlp-main.183. (Cited on page 90.)

[65] Y. Dou, M. Forbes, R. Koncel-Kedziorski, N. A. Smith, and Y. Choi. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland*, pages 7250–7274. Association for Computational Linguistics, 2022. URL https://doi.org/10.18653/v1/2022.acl-long.501. (Cited on page 92.)

[66] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA*, pages 155–161. MIT Press, 1996. URL http://papers.nips.cc/paper/1238-support-vector-regression-machines. (Cited on page 31.)

[67] N. Dziri, E. Kamalloo, K. Mathewson, and O. Zaiane. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HLT 2019, Volume 1 (Long and Short Papers), Minneapolis, Minnesota*, pages 3806–3812. Association for Computational Linguistics, 2019. URL https://aclanthology.org/N19-1381. (Cited on page 17.)

[68] J. Edlund, J. Gustafson, M. Heldner, and A. Hjalmarsson. Towards human-like spoken dialogue systems. *Speech Communication*, 50:630–645, 2008. URL https://www.sciencedirect.com/science/article/pii/S016763930800054X. (Cited on page 23.)

[69] C. Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 162–170. ACM, 2018. URL https://doi.org/10.1145/3159652.3159654. (Cited on pages 58, 68, 81, and 152.)

[70] G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, and H. Wachsmuth. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information*

Retrieval, ICTIR 2023, Taipei, Taiwan, pages 39–50. ACM, 2023. URL https://doi.org/10.1145/3578337.3605136. (Cited on pages 39 and 48.)

[71] X. Feng, X. Feng, and B. Qin. A survey on dialogue summarization: Recent advances and new frontiers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria*, pages 5453–5460. ijcai.org, 2022. URL https://doi.org/10.24963/ijcai.2022/764. (Cited on page 51.)

[72] N. Ferro and M. Maistro. Evaluation of IR systems. In *Information Retrieval: Advanced Topics and Techniques*, volume 60 of *ACM Books*, pages 111–191. ACM, 2024. URL https://doi.org/10.1145/3674127.3674132. (Cited on page 1.)

[73] C. Fichter and K. Jonas. Image effects of newspapers: How brand images change consumers' product ratings. *Zeitschrift für Psychologie/Journal of Psychology*, 216(4):226–234, 2008. URL https://doi.org/10.1027/0044-3409.216.4.226. (Cited on page 116.)

[74] S. E. Finch and J. D. Choi. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue SIGDIAL 2020, 1st virtual meeting*, pages 236–245. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.sigdial-1.29. (Cited on pages 12 and 13.)

[75] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001. URL https://doi.org/10.1214/aos/1013203451. (Cited on page 31.)

[76] U. Gadiraju, J. Yang, and A. Bozzon. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT 2017, Prague, Czech Republic*, pages 5–14. ACM, 2017. URL https://doi.org/10.1145/3078714.3078715. (Cited on page 69.)

[77] C. Gao, W. Lei, X. He, M. de Rijke, and T.-S. Chua. Advances and challenges in conversational recommender systems: A survey. *AI Open*, 2(2666-6510):100–126, 2021. doi: https://doi.org/10.1016/j.aiopen.2021.06.002. URL https://www.sciencedirect.com/science/article/pii/S2666651021000164. (Cited on pages 1, 15, and 27.)

[78] C. Gao, S. Wang, S. Li, J. Chen, X. He, W. Lei, B. Li, Y. Zhang, and P. Jiang. CIRS: bursting filter bubbles by counterfactual interactive recommender system. *ACM Trans. Inf. Syst.*, 42(1):14:1–14:27, 2024. URL https://doi.org/10.1145/3594871. (Cited on page 29.)

[79] J. Gao, M. Galley, and L. Li. Neural approaches to conversational ai. *Foundations and Trends in Information Retrieval*, 13(2-3):127–298, 2019. URL https://doi.org/10.1561/1500000074. (Cited on page 12.)

[80] S. Ghazarian, J. Wei, A. Galstyan, and N. Peng. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-2310. (Cited on pages 12 and 49.)

[81] S. Ghazarian, R. M. Weischedel, A. Galstyan, and N. Peng. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. Technical report, 2020. URL https://doi.org/10.1609/aaai.v34i05.6283. (Cited on page 1.)

[82] F. Gilardi, M. Alizadeh, and M. Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056, 2023. URL https://doi.org/10.48550/arXiv.2303.15056. (Cited on page 69.)

[83] K. Gopalakrishnan, B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, and D. Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895, 2019. URL https://doi.org/10.21437/Interspeech.2019-3079. (Cited on pages 12 and 13.)

[84] J. R. Griffiths, F. C. Johnson, and R. J. Hartley. User satisfaction as a measure of system performance. *J. Libr. Inf. Sci.*, 39(3):142–152, 2007. URL https://doi.org/10.1177/0961000607080417. (Cited on page 1.)

[85] S. Guo, S. Zhang, W. Sun, P. Ren, Z. Chen, and Z. Ren. Towards explainable conversational recommender systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan,*, pages 2786–2795. ACM, 2023. URL https://doi.org/10.1145/3539618.3591884. (Cited on pages 70 and 74.)

[86] M. Hämäläinen and K. Al-Najjar. The great misalignment problem in human evaluation of NLP methods. *CoRR*, abs/2104.05361, 2021. URL https://arxiv.org/abs/2104.05361. (Cited on

page 92.)

[87] L. Han, K. Roitero, U. Gadiraju, C. Sarasua, A. Checco, E. Maddalena, and G. Demartini. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia*, pages 321–329. ACM, 2019. URL https://doi.org/10.1145/3289600.3291035. (Cited on page 69.)

[88] L. Han, K. Roitero, E. Maddalena, S. Mizzaro, and G. Demartini. On transforming relevance scales. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China*, pages 39–48. ACM, 2019. URL https://doi.org/10.1145/3357384.3357988. (Cited on page 78.)

[89] L. Han, E. Maddalena, A. Checco, C. Sarasua, U. Gadiraju, K. Roitero, and G. Demartini. Crowd worker strategies in relevance judgment tasks. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA*, pages 241–249. ACM, 2020. URL https://doi.org/10.1145/3336191.3371857. (Cited on page 69.)

[90] D. Harman. *Information Retrieval Evaluation*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2011. URL https://doi.org/10.2200/S00368ED1V01Y201105ICR019. (Cited on page 1.)

[91] H. Hashemi, H. Zamani, and W. B. Croft. Guided transformer: Leveraging multiple external sources for representation learning in conversational search. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China*, pages 1131–1140. ACM, 2020. URL https://doi.org/10.1145/3397271.3401061. (Cited on page 118.)

[92] S. H. Hashemi, K. Williams, A. El Kholy, I. Zitouni, and P. A. Crook. Measuring user satisfaction on smart speaker intelligent assistants using intent sensitive query embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1183–1192. ACM, 2018. URL https://doi.org/10.1145/3269206.3271802. (Cited on pages 12, 16, and 17.)

[93] A. Hassan. A semi-supervised approach to modeling web search satisfaction. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12. ACM, 2012. URL https://doi.org/10.1145/2348283.2348323. (Cited on page 17.)

[94] X. He, Z. Lin, Y. Gong, A. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, and W. Chen. Annollm: Making large language models to be better crowdsourced annotators. *CoRR*, abs/2303.16854, 2023. URL https://doi.org/10.48550/arXiv.2303.16854. (Cited on pages 69 and 152.)

[95] M. A. Hearst. Trends & controversies: Mixed-initiative interaction. *IEEE Intell. Syst.*, 14(5):14–23, 1999. URL https://doi.org/10.1109/5254.796083. (Cited on page 118.)

[96] M. Hossain and I. Kauranen. Crowdsourcing: a comprehensive literature review. *Strategic Outsourcing: An International Journal*, 8(1):2–22, 2015. (Cited on page 2.)

[97] P. S. Houts, C. C. Doak, L. G. Doak, and M. J. Loscalzo. The role of pictures in improving health communication: A review of research on attention, comprehension, recall, and adherence. *Patient Education and Counseling*, 61(2):173–190, 2006. URL https://www.sciencedirect.com/science/article/pii/S0738399105001461. (Cited on page 116.)

[98] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080, 2020. URL https://arxiv.org/abs/2003.11080. (Cited on page 153.)

[99] J. Huang and E. N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China*, pages 77–86. ACM, 2009. URL https://doi.org/10.1145/1645953.1645966. (Cited on pages 120 and 145.)

[100] L. Huang, Z. Ye, J. Qin, L. Lin, and X. Liang. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online, 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.742. (Cited on pages 12 and 49.)

[101] C. Hube, B. Fetahu, and U. Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12. ACM, 2019. URL https://doi.org/10.1145/3290605.3300637. (Cited on pages 2, 49, 68, 69, and 81.)

[102] D. Jannach, A. Manzoor, W. Cai, and L. Chen. A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5), 2021. URL https://doi.org/10.1145/3453154. (Cited on page 15.)

[103] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002. URL http://doi.acm.org/10.1145/582415.582418. (Cited on page 11.)

[104] V. Jeronymo, L. H. Bonifacio, H. Q. Abonizio, M. Fadaee, R. de Alencar Lotufo, J. Zavrel, and R. F. Nogueira. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *CoRR*, abs/2301.01820, 2023. URL https://doi.org/10.48550/arXiv.2301.01820. (Cited on page 90.)

[105] T. Ji, Y. Graham, G. J. F. Jones, C. Lyu, and Q. Liu. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland*, pages 6416–6437. Association for Computational Linguistics, 2022. URL https://doi.org/10.18653/v1/2022.acl-long.445. (Cited on page 2.)

[106] J. Jiang, A. Hassan Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 57–66. ACM, 2015. URL https://doi.org/10.1145/2684822.2685319. (Cited on pages 16 and 17.)

[107] J. Jiang, D. He, and J. Allan. Comparing in situ and multidimensional relevance judgments. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan*, pages 405–414. ACM, 2017. URL https://doi.org/10.1145/3077136.3080840. (Cited on page 70.)

[108] S. Jiang, S. Vakulenko, and M. de Rijke. Weakly supervised turn-level engagingness evaluator for dialogues. In *CHIIR '23: ACM SIGIR Conference on Human Information Interaction and Retrieval, Austin, TX, USA*, pages 258–268. ACM, 2023. URL https://doi.org/10.1145/3576840.3578319. (Cited on page 19.)

[109] Y. Jin, L. Chen, W. Cai, and P. Pu. Key qualities of conversational recommender systems: From users' perspective. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, HAI '21, page 93–102, New York, NY, USA, 2021. ACM. URL https://doi.org/10.1145/3472307.3484164. (Cited on pages 13 and 18.)

[110] H. Joko, S. Chatterjee, A. Ramsay, A. P. de Vries, J. Dalton, and F. Hasibi. Doing personal LAPS: LLM-augmented dialogue construction for personalized multi-session conversational search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA*, pages 796–806. ACM, 2024. URL https://doi.org/10.1145/3626772.3657815. (Cited on page 91.)

[111] P. Kadasi and M. Singh. Unveiling the multi-annotation process: Examining the influence of annotation quantity and instance difficulty on model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore*, pages 1371–1388. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.findings-emnlp.96. (Cited on page 58.)

[112] J. Kang, K. Condiff, S. Chang, J. A. Konstan, L. Terveen, and F. M. Harper. Understanding how people use natural language to ask for recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, page 229–237. ACM, 2017. URL https://doi.org/10.1145/3109859.3109873. (Cited on page 15.)

[113] G. Kazai. In search of quality in crowdsourcing for search engine evaluation. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland*, volume 6611 of *Lecture Notes in Computer Science*, pages 165–176. Springer, 2011. URL https://doi.org/10.1007/978-3-642-20161-5_17. (Cited on page 51.)

[114] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China*, pages 205–214. ACM, 2011. URL https://doi.org/10.1145/2009916.2009947. (Cited on page 58.)

[115] G. Kazai, J. Kamps, and N. Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 1941–1944, New York, NY, USA, 2011. ACM. URL https://doi.org/10.1145/2063576.2063860. (Cited on pages 2 and 49.)

[116] G. Kazai, J. Kamps, and N. Milic-Frayling. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 2583–2586. ACM, 2012. URL https://doi.org/10.1145/2396761.2398697. (Cited on pages 2, 49, and 59.)

[117] G. Kazai, J. Kamps, and N. Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 16(2):138–178, 2013. URL https://doi.org/10.1007/s10791-012-9205-0. (Cited on pages 2, 49, 51, 59, 69, and 152.)

[118] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, 3(1—2):1–224, 2009. URL https://doi.org/10.1561/1500000012. (Cited on pages 1, 2, 16, 17, and 122.)

[119] C. Kern, S. Eckman, J. Beck, R. Chew, B. Ma, and F. Kreuter. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore*, pages 14874–14886. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.findings-emnlp.992. (Cited on pages 2 and 58.)

[120] J. Kiesel, A. Bahrami, B. Stein, A. Anand, and M. Hagen. Toward voice query clarification. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA*, pages 1257–1260. ACM, 2018. URL https://doi.org/10.1145/3209978.3210160. (Cited on pages 116, 118, and 119.)

[121] H. Kim, Y. Choi, T. Yang, H. Lee, C. Park, Y. Lee, J. Y. Kim, and J. Kim. Using LLMs to investigate correlations of conversational follow-up queries with user satisfaction. *CoRR*, abs/2407.13166, 2024. URL https://doi.org/10.48550/arXiv.2407.13166. (Cited on page 91.)

[122] J. Y. Kim, J. Teevan, and N. Craswell. Explicit in situ user feedback for web search results. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy*, pages 829–832. ACM, 2016. URL https://doi.org/10.1145/2911451.2914754. (Cited on page 68.)

[123] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, page 193–202. ACM, 2014. URL https://doi.org/10.1145/2556195.2556220. (Cited on pages 16, 17, and 68.)

[124] S. Kiritchenko and S. M. Mohammad. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, Volume 2: Short Papers*, pages 465–470. Association for Computational Linguistics, 2017. URL https://doi.org/10.18653/v1/P17-2074. (Cited on page 92.)

[125] J. Kiseleva, K. Williams, A. Hassan Awadallah, A. C. Crook, I. Zitouni, and T. Anastasakos. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 45–54, New York, NY, USA, 2016. ACM. ISBN 9781450340694. URL https://doi.org/10.1145/2911451.2911521. (Cited on pages 13, 16, 17, and 31.)

[126] J. Kiseleva, K. Williams, J. Jiang, A. Hassan Awadallah, A. C. Crook, I. Zitouni, and T. Anastasakos. Understanding user satisfaction with intelligent assistants. In *CHIIR '16: ACM SIGIR Conference on Human Information Interaction and Retrieval*, page 121–130. ACM, 2016. URL https://doi.org/10.1145/2854946.2854961. (Cited on pages 12, 16, 23, 27, 37, and 49.)

[127] T. Kocmi and C. Federmann. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland*, pages 193–203. European Association for Machine Translation, 2023. URL https://aclanthology.org/2023.eamt-1.19. (Cited on page 92.)

[128] R. Koo, M. Lee, V. Raheja, J. I. Park, Z. M. Kim, and D. Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.29/. (Cited on page 2.)

[129] A. M. Krasakis, M. Aliannejadi, N. Voskarides, and E. Kanoulas. Analysing the effect of clarifying questions on document ranking in conversational search. pages 129–132, 2020. URL https://doi.org/10.1145/3409256.3409817. (Cited on pages 116 and 118.)

[130] J. Li, Y. Yang, Z. Wu, V. G. V. Vydiswaranm, and C. Xiao. Chatgpt as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies NAACL-HLT 2024, Volume 1 (Long Papers), Mexico City, Mexico, pages 2985–3004. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.naacl-long.165. (Cited on page 152.)

[131] M. Li, J. Weston, and S. Roller. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087, 2019. URL http://arxiv.org/abs/1909.03087. (Cited on pages 13 and 47.)

[132] M. Li, T. Shi, C. Ziems, M. Kan, N. F. Chen, Z. Liu, and D. Yang. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore*, pages 1487–1505. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.emnlp-main.92. (Cited on pages 59 and 152.)

[133] R. Li, S. E. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, Canada*, pages 9748–9758, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/800de15c79c8d840f4e78d3af937d4d4-Abstract.html. (Cited on pages 13, 20, 21, 50, 67, and 70.)

[134] Y. Li and Y. Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19, 2020. URL https://doi.org/10.1177/0022243719881113. (Cited on page 116.)

[135] Z. Li, C. Wang, P. Ma, D. Wu, S. Wang, C. Gao, and Y. Liu. Split and merge: Aligning position biases in LLM-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.621/. (Cited on page 2.)

[136] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out TSB 2004, Barcelona, Spain*, pages 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013. (Cited on pages 12, 49, and 92.)

[137] Y.-T. Lin and Y.-N. Chen. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the Workshop on NLP for Conversational AI NLP4ConvAI 2023, Toronto, Canada*, pages 47–58. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.nlp4convai-1.5. (Cited on page 92.)

[138] A. Lipani, B. Carterette, and E. Yilmaz. How am i doing?: Evaluating conversational search systems offline. *ACM Trans. Inf. Syst.*, 39(4):Article 51, 2021. (Cited on page 1.)

[139] C. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA*, pages 2122–2132. The Association for Computational Linguistics, 2016. URL https://doi.org/10.18653/v1/d16-1230. (Cited on pages 12, 16, and 92.)

[140] J. Liu and F. Han. Investigating reference dependence effects on user search interaction and satisfaction: A behavioral economics perspective. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China*, pages 1141–1150. ACM, 2020. URL https://doi.org/10.1145/3397271.3401085. (Cited on pages 27 and 37.)

[141] M. Liu, Y. Liu, J. Mao, C. Luo, M. Zhang, and S. Ma. "satisfaction with failure" or "unsatisfied success": Investigating the relationship between search success and user satisfaction. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1533–1542, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. URL https://doi.org/10.1145/3178876.3186065. (Cited on pages 19 and 68.)

[142] Y. Liu, Y. Chen, J. Tang, J. Sun, M. Zhang, S. Ma, and X. Zhu. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile*, pages 493–502. ACM, 2015. URL https://doi.org/10.1145/2766462.2767721. (Cited on page 68.)

[143] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore*, pages 2511–2522. Association for Computational Lin-

guistics, 2023. URL https://doi.org/10.18653/v1/2023.emnlp-main.153. (Cited on page 92.)

[144] Y. Liu, N. S. Moosavi, and C. Lin. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting*, pages 12688–12701. Association for Computational Linguistics, 2024. URL https://doi.org/10.18653/v1/2024.findings-acl.753. (Cited on page 100.)

[145] H. Lu, W. Ma, M. Zhang, M. de Rijke, Y. Liu, and S. Ma. Standing in your shoes: External assessments for personalized recommender systems. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada*, pages 1523–1533. ACM, 2021. URL https://doi.org/10.1145/3404835.3462916. (Cited on page 38.)

[146] X. Lu, M. Kelly, and E. F. Risko. Cognitive offloading increases false recall. In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual,*. cognitivesciencesociety.org, 2020. URL https://cogsci.mindmodeling.org/2020/papers/0618/index.html. (Cited on page 145.)

[147] E. Luger and A. Sellen. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA*, pages 5286–5297. ACM, 2016. URL https://doi.org/10.1145/2858036.2858288. (Cited on page 1.)

[148] Y. Ma, T. Kleemann, and J. Ziegler. Mixed-modality interaction in conversational recommender systems. In *Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with 15th ACM Conference on Recommender Systems (RecSys 2021), Online Event,*, volume 2948 of *CEUR Workshop Proceedings*, pages 21–37. CEUR-WS.org, 2021. URL https://ceur-ws.org/Vol-2948/paper2.pdf. (Cited on page 118.)

[149] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin. On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Trans. Inf. Syst.*, 35(3):19:1–19:32, 2017. URL https://doi.org/10.1145/3002172. (Cited on page 70.)

[150] J. Mao, Y. Liu, K. Zhou, J. Nie, J. Song, M. Zhang, S. Ma, J. Sun, and H. Luo. When does relevance mean usefulness and user satisfaction in web search? In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy*, pages 463–472. ACM, 2016. URL https://doi.org/10.1145/2911451.2911507. (Cited on page 50.)

[151] J. Mao, Y. Liu, N. Kando, C. Luo, M. Zhang, and S. Ma. Investigating result usefulness in mobile search. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France*, volume 10772 of *Lecture Notes in Computer Science*, pages 223–236. Springer, 2018. URL https://doi.org/10.1007/978-3-319-76941-7_17. (Cited on page 70.)

[152] R. E. Mayer and R. Moreno. Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1):43–52, 2003. URL https://doi.org/10.1207/S15326985EP3801_6. (Cited on page 120.)

[153] T. McDonnell, M. Lease, M. Kutlu, and T. Elsayed. Why is that relevant? Collecting annotator rationales for relevance judgments. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, Austin, Texas, USA*, pages 139–148. AAAI Press, 2016. URL https://doi.org/10.1609/hcomp.v4i1.13287. (Cited on page 72.)

[154] M. F. McTear. *Spoken dialogue technology - toward the conversational user interface*. Springer, 2004. URL http://www.springer.com/computer/ai/book/978-1-85233-672-1. (Cited on page 1.)

[155] S. Mehri and M. Eskénazi. USR: an unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online*, pages 681–707. Association for Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.acl-main.64. (Cited on page 92.)

[156] S. Mehri and M. Eskenazi. Unsupervised evaluation of interactive dialog with DialogGPT. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue SIGDIAL 2020, 1st virtual meeting*, pages 225–235. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.sigdial-1.28. (Cited on pages 1, 13, 18, 21, and 37.)

[157] S. Mehri and M. Eskenazi. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*

ACL 2020, Online, pages 681–707. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.acl-main.64. (Cited on pages 12, 13, 16, 17, 18, 19, 20, 21, 22, 23, 37, 38, 48, 49, and 66.)

[158] C. Meng, P. Ren, Z. Chen, Z. Ren, T. Xi, and M. de Rijke. Initiative-aware self-supervised learning for knowledge-grounded conversations. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada*, pages 522–532. ACM, 2021. URL https://doi.org/10.1145/3404835.3462824. (Cited on page 118.)

[159] N. Murrugarra-Llerena and A. Kovashka. Image retrieval with mixed initiative and multimodal feedback. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK*, page 310. BMVA Press, 2018. URL http://bmvc2018.org/contents/papers/0151.pdf. (Cited on page 118.)

[160] C. Niu, X. Wang, X. Cheng, J. Song, and T. Zhang. Enhancing dialogue state tracking models through LLM-backed user-agents simulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand*, pages 8724–8741. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.473. (Cited on page 91.)

[161] J. Novikova, O. Dušek, and V. Rieser. Rankme: Reliable human ratings for natural language generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HLT 2018, Volume 2 (Short Papers), New Orleans, Louisiana*, pages 72–78. Association for Computational Linguistics, 2018. URL https://aclanthology.org/N18-2012. (Cited on page 49.)

[162] S. Nowak and S. M. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2010, Philadelphia, Pennsylvania, USA*, pages 557–566. ACM, 2010. URL https://doi.org/10.1145/1743384.1743478. (Cited on page 120.)

[163] H. O'Brien, A. W. Cole, A. Kampen, and K. Brennan. The effects of domain and search expertise on learning outcomes in digital library use. In *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany*, pages 202–210. ACM, 2022. URL https://doi.org/10.1145/3498366.3505761. (Cited on page 145.)

[164] O. Ogundepo, T. R. Gwadabe, C. E. Rivera, J. H. Clark, S. Ruder, D. I. Adelani, B. F. P. Dossou, A. A. Diop, C. Sikasote, G. Hacheme, H. Buzaaba, I. Ezeani, R. Mabuya, S. Osei, C. Emezue, A. N. Kahira, S. H. Muhammad, A. Oladipo, A. T. Owodunni, A. L. Tonja, I. Shode, A. Asai, T. O. Ajayi, C. Siro, S. Arthur, M. Adeyemi, O. Ahia, A. Anuoluwapo, O. Awosan, C. Chukwuneke, B. Opoku, A. Ayodele, V. Otiende, C. Mwase, B. Sinkala, A. N. Rubungo, D. A. Ajisafe, E. F. Onwuegbuzia, H. Mbow, E. Niyomutabazi, E. Mukonde, F. I. Lawan, I. S. Ahmad, J. O. Alabi, M. Namukombo, C. E. Mbonu, M. Phiri, N. Putini, N. Mngoma, P. A. Amuok, R. N. Iro, and S. Adhiambo. Cross-lingual open-retrieval question answering for african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore*, pages 14957–14972. Association for Computational Linguistics, 2023. URL https://doi.org/10.18653/v1/2023.findings-emnlp.997. (Cited on page 154.)

[165] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. URL https://doi.org/10.48550/arXiv.2303.08774. (Cited on pages 51 and 105.)

[166] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html. (Cited on page 153.)

[167] P. Owoicho, I. Sekulic, M. Aliannejadi, J. Dalton, and F. Crestani. Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. pages 632–642, 2023. URL https://doi.org/10.1145/3539618.3591683. (Cited on pages 118 and 138.)

[168] A. Paivio. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 45(3):255, 1991. URL https://doi.org/10.1037/h0084295. (Cited on page 144.)

[169] A. Panickssery, S. R. Bowman, and S. Feng. LLM evaluators recognize and favor their own gen-

erations. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html. (Cited on page 2.)

[170] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318. Association for Computational Linguistics, 2002. URL https://doi.org/10.3115/1073083.1073135. (Cited on pages 1, 12, 49, and 92.)

[171] K. Park, M. Cha, and E. Rhim. Positivity bias in customer satisfaction ratings. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France*, pages 631–638. ACM, 2018. URL https://doi.org/10.1145/3184558.3186579. (Cited on page 55.)

[172] M. Parmar, S. Mishra, M. Geva, and C. Baral. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia*, pages 1771–1781. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.eacl-main.130. (Cited on page 48.)

[173] W. Pei, Z. Yang, M. Chen, and C. Yue. Quality control in crowdsourcing based on fine-grained behavioral features. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW2):442:1–442:28, 2021. URL https://doi.org/10.1145/3479586. (Cited on pages 2 and 49.)

[174] C. Qu, L. Yang, W. B. Croft, J. R. Trippas, Y. Zhang, and M. Qiu. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA*, pages 989–992. ACM, 2018. URL https://doi.org/10.1145/3209978.3210124. (Cited on page 120.)

[175] C. Qu, L. Yang, W. B. Croft, Y. Zhang, J. R. Trippas, and M. Qiu. User intent prediction in information-seeking conversations. In *CHIIR '19: ACM SIGIR Conference on Human Information Interaction and Retrieval, Glasgow, Scotland, UK*, pages 25–33. ACM, 2019. URL https://doi.org/10.1145/3295750.3298924. (Cited on page 17.)

[176] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html. (Cited on page 118.)

[177] F. Radlinski and N. Craswell. A theoretical framework for conversational search. In *CHIIR '17: ACM SIGIR Conference on Human Information Interaction and Retrieval, Oslo, Norway*, pages 117–126. ACM, 2017. URL https://doi.org/10.1145/3020165.3020183. (Cited on pages 1, 89, 91, and 115.)

[178] F. Radlinski, K. Balog, B. Byrne, and K. Krishnamoorthi. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue SIGDIAL 2019, Stockholm, Sweden*, pages 353–360. Association for Computational Linguistics, 2019. URL https://aclanthology.org/W19-5941. (Cited on page 12.)

[179] H. A. Rahmani, C. Siro, M. Aliannejadi, N. Craswell, C. L. A. Clarke, G. Faggioli, B. Mitra, P. Thomas, and E. Yilmaz. LLM4Eval: Large language model for evaluation in IR. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA*, pages 3040–3043. ACM, 2024. URL https://doi.org/10.1145/3626772.3657992. (Cited on page 153.)

[180] H. A. Rahmani, E. Yilmaz, N. Craswell, B. Mitra, P. Thomas, C. L. A. Clarke, M. Aliannejadi, C. Siro, and G. Faggioli. LLMJudge: LLMs for relevance judgments. In *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024) co-located with 10th International Conference on Online Publishing (SIGIR 2024), Washington D.C., USA*, volume 3752 of *CEUR Workshop Proceedings*, pages 1–3. CEUR-WS.org, 2024. URL https://ceur-ws.org/Vol-3752/paper8.pdf. (Cited on page 152.)

[181] S. Rao and H. D. III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, Volume 1: Long Papers*, pages 2737–2746. Association for Computational Linguistics, 2018. URL https://aclanthology.

org/P18-1255/. (Cited on pages 94 and 118.)

[182] A. Rastogi, X. Zang, S. Sunkara, R. Gupta, and P. Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8689–8696, Apr. 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6394. (Cited on page 12.)

[183] R. Ren, P. Qiu, Y. Qu, J. Liu, W. X. Zhao, H. Wu, J.-R. Wen, and H. Wang. Bases: Large-scale web search user simulation with large language model based agents. *ArXiv*, abs/2402.17505, 2024. URL https://api.semanticscholar.org/CorpusID:268032631. (Cited on page 91.)

[184] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. In *Recommender Systems Handbook*, pages 1–35. Springer, 2011. URL https://doi.org/10.1007/978-0-387-85820-3_1. (Cited on page 15.)

[185] K. Roitero, E. Maddalena, G. Demartini, and S. Mizzaro. On fine-grained relevance scales. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA*, pages 675–684. ACM, 2018. URL https://doi.org/10.1145/3209978.3210052. (Cited on pages 66, 67, and 70.)

[186] K. Roitero, M. Soprano, S. Fan, D. Spina, S. Mizzaro, and G. Demartini. Can the crowd identify misinformation objectively?: The effects of judgment scale and assessor's background. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China*, pages 439–448. ACM, 2020. URL https://doi.org/10.1145/3397271.3401112. (Cited on pages 48, 49, 51, and 69.)

[187] K. Roitero, E. Maddalena, S. Mizzaro, and F. Scholer. On the effect of relevance scales in crowdsourcing relevance assessments for information retrieval evaluation. *Inf. Process. Manag.*, 58(6):102688, 2021. URL https://doi.org/10.1016/j.ipm.2021.102688. (Cited on pages 48 and 49.)

[188] C. Rzepka and B. Berger. User interaction with ai-enabled systems: A systematic review of IS research. In *Proceedings of the International Conference on Information Systems - Bridging the Internet of People, Data, and Things, ICIS 2018, San Francisco, CA, USA*. Association for Information Systems, 2018. URL https://aisel.aisnet.org/icis2018/general/Presentations/7. (Cited on page 1.)

[189] S. Saha, O. Levy, A. Celikyilmaz, M. Bansal, J. Weston, and X. Li. Branch-solve-merge improves large language model evaluation and generation. 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*, pages 8352–8370. Association for Computational Linguistics, 2024. URL https://doi.org/10.18653/v1/2024.naacl-long.462. (Cited on page 96.)

[190] S. Santhanam, A. Karduni, and S. Shaikh. Studying the effects of cognitive biases in evaluation of conversational agents. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA*, pages 1–13. ACM, 2020. URL https://doi.org/10.1145/3313831.3376318. (Cited on pages 48, 49, 58, 68, and 81.)

[191] T. Saracevic. Relevance reconsidered. In *Information Science: Integration in Perspectives. Proceedings of the Second Conference on Conceptions of Library and Information Science, Copenhagen (Denmark)*, 1996. (Cited on page 18.)

[192] A. Schmitt and S. Ultes. Interaction quality: Assessing the quality of ongoing spoken dialog interaction by experts - and how it relates to user satisfaction. *Speech Commun.*, 74:12–36, 2015. URL https://doi.org/10.1016/j.specom.2015.06.003. (Cited on page 48.)

[193] A. Schmitt, S. Ultes, and W. Minker. A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system. In *Proceedings of the International Conference on Language Resources and Evaluation LREC 2012, Istanbul, Turkey*, pages 3369–3373. European Language Resources Association (ELRA), 2012. URL https://aclanthology.org/L12-1157. (Cited on page 30.)

[194] A. See, S. Roller, D. Kiela, and J. Weston. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HLT 2019, Volume 1 (Long and Short Papers), Minneapolis, Minnesota*, pages 1702–1723. Association for Computational Linguistics, 2019. URL https://aclanthology.org/N19-1170. (Cited on pages 12 and 17.)

[195] I. Sekulic, M. Aliannejadi, and F. Crestani. User engagement prediction for clarification in search. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual*

Event, Part I, volume 12656 of *Lecture Notes in Computer Science*, pages 619–633. Springer, 2021. URL https://doi.org/10.1007/978-3-030-72113-8_41. (Cited on pages 92, 118, and 138.)

[196] I. Sekulic, M. Aliannejadi, and F. Crestani. Towards facet-driven generation of clarifying questions for conversational search. In *ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada*, pages 167–175. ACM, 2021. URL https://doi.org/10.1145/3471158.3472257. (Cited on page 92.)

[197] I. Sekulic, M. Aliannejadi, and F. Crestani. Analysing utterances in LLM-based user simulation for conversational search. *ACM Trans. Intell. Syst. Technol.*, 15(3):62:1–62:22, 2024. URL https://doi.org/10.1145/3650041. (Cited on pages 91, 93, and 100.)

[198] T. Sellam, D. Das, and A. P. Parikh. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online*, pages 7881–7892. Association for Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.acl-main.704. (Cited on page 92.)

[199] C. Siro and T. O. Ajayi. Evaluating the robustness of machine reading comprehension models to low resource entity renaming. In *Proceedings of the 4th Workshop on African Natural Language Processing, AfricaNLP@ICLR 2023, Kigali, Rwanda*, 2023. URL https://openreview.net/pdf?id=GfnDjdcnYT. (Cited on page 154.)

[200] C. Siro, M. Aliannejadi, and M. de Rijke. Understanding user satisfaction with task-oriented dialogue systems. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain*, pages 2018–2023. ACM, 2022. URL https://doi.org/10.1145/3477495.3531798. (Cited on pages 13, 14, 16, 17, 18, 19, 20, 23, 26, 27, 48, 49, 68, 70, and 79.)

[201] C. Siro, M. Aliannejadi, and M. de Rijke. Understanding and predicting user satisfaction with conversational recommender systems. *ACM Transactions on Information Systems*, 42(2):Article 55, sep 2023. URL https://doi.org/10.1145/3624989. (Cited on pages 48, 66, 68, 69, 70, 78, and 94.)

[202] C. Siro, M. Aliannejadi, and M. de Rijke. Rethinking the evaluation of dialogue systems: Effects of user feedback on crowdworkers and LLMs. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA*, pages 1952–1962. ACM, 2024. URL https://doi.org/10.1145/3626772.3657712. (Cited on pages 2 and 96.)

[203] C. Siro, M. Aliannejadi, and M. de Rijke. Context does matter: Implications for crowdsourced evaluation labels in task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico*, pages 1258–1273. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-naacl.80. (Cited on pages 66, 70, and 94.)

[204] C. Siro, Y. Yuan, M. Aliannejadi, and M. de Rijke. AGENT-CQ: Automatic generation and evaluation of clarifying questions for conversational search with llms. *CoRR*, abs/2410.19692, 2024. (Cited on page 152.)

[205] C. Siro, Z. Abbasiantaeb, Y. Yuan, M. Aliannejadi, and M. de Rijke. Do images clarify? A study on the effect of images on clarifying questions in conversational search. In *CHIIR '25: ACM SIGIR Conference on Human Information Interaction and Retrieval, Melbourne, Australia*, page 273–291. ACM, 2025. (Cited on pages 2 and 153.)

[206] K. Sun, S. Moon, P. Crook, S. Roller, B. Silvert, B. Liu, Z. Wang, H. Liu, E. Cho, and C. Cardie. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HLT 2021, Online*, pages 1570–1583. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.naacl-main.124. (Cited on pages 19 and 37.)

[207] W. Sun, S. Zhang, K. Balog, Z. Ren, P. Ren, Z. Chen, and M. de Rijke. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada*, pages 2499–2506. ACM, 2021. URL https://doi.org/10.1145/3404835.3463241. (Cited on pages 13, 16, 17, 21, 26, 30, 31, 47, and 66.)

[208] Y. Sun and Y. Zhang. Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA*, pages 235–244. ACM, 2018. URL https://doi.org/10.1145/3209978.3210002. (Cited on page 1.)

[209] A. Talmor, O. Yoran, A. Catav, D. Lahav, Y. Wang, A. Asai, G. Ilharco, H. Hajishirzi, and J. Berant. Multimodalqa: Complex question answering over text, tables and images. 2021. URL https://openreview.net/forum?id=ee6W5UgQLa. (Cited on page 118.)

[210] Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346, 2024. (Cited on page 2.)

[211] L. Tavakoli, J. R. Trippas, H. Zamani, F. Scholer, and M. Sanderson. Mimics-duo: Offline & online evaluation of search clarification. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain*, pages 3198–3208. ACM, 2022. URL https://doi.org/10.1145/3477495.3531750. (Cited on page 92.)

[212] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. URL https://doi.org/10.48550/arXiv.2302.13971. (Cited on page 105.)

[213] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. URL https://doi.org/10.48550/arXiv.2307.09288. (Cited on page 59.)

[214] M. A. Turk. Multimodal interaction: A review. *Pattern Recognit. Lett.*, 36:189–195, 2014. URL https://doi.org/10.1016/j.patrec.2013.07.003. (Cited on pages 118 and 120.)

[215] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 225–231. ACM, 2001. URL https://doi.org/10.1145/383952.383992. (Cited on page 11.)

[216] P. Vakkari, M. Völske, M. Potthast, M. Hagen, and B. Stein. Modeling the usefulness of search results as measured by information use. *Inf. Process. Manag.*, 56(3):879–894, 2019. URL https://doi.org/10.1016/j.ipm.2019.02.001. (Cited on page 70.)

[217] S. Vakulenko, E. Kanoulas, and M. de Rijke. An analysis of mixed initiative and collaboration in information-seeking dialogues. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China*, pages 2085–2088. ACM, 2020. URL https://doi.org/10.1145/3397271.3401297. (Cited on page 118.)

[218] A. Venkatesh, C. Khatri, A. Ram, F. Guo, R. Gabriel, A. Nagar, R. Prasad, M. Cheng, B. Hedayatnia, A. Metallinou, R. Goel, S. Yang, and A. Raju. On evaluating and comparing open domain dialog systems. *CoRR*, abs/1801.03625, 2018. URL https://api.semanticscholar.org/CorpusID:57189198. (Cited on pages 12, 13, 16, 17, 18, and 19.)

[219] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98/EACL '98, page 271–280. Association for Computational Linguistics, 1997. URL https://doi.org/10.3115/976909.979652. (Cited on page 16.)

[220] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7. (Cited on page 153.)

[221] J. Wang, D. I. Adelani, S. Agrawal, R. Rei, E. Briakou, M. Carpuat, M. Masiak, X. He, S. Bourhim, A. Bukula, M. Mohamed, T. Olatoye, H. Mokayed, C. Mwase, W. Kimotho, F. Yuehgoh, A. Anuoluwapo, J. Ojo, S. H. Muhammad, S. Osei, A. Omotayo, C. Chukwuneke, P. Ogayo, O. Hourrane, S. E. Anigri, L. Ndolela, T. Mangwana, S. A. Mohamed, A. Hassan, O. O. Awoyomi, L. Alkhaled, S. S. Al-Azzawi, N. A. Etori, M. Ochieng, C. Siro, S. Njoroge, E. Muchiri, W. Kimotho, L. N. W. Momo, D. Abolade, S. Ajao, T. P. Adewumi, I. Shode, R. Macharm, R. N. Iro, S. S. Abdullahi, S. E. Moore, B. Opoku, Z. Akinjobi, A. Abeeb, N. C. Obiefuna, O. R. Ogbu, S. Brian, V. A. Otiende, C. E. Mbonu, S. T. Sari, and P. Stenetorp. Afrimte and africomet: Enhancing COMET to embrace under-

resourced african languages. 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*, pages 5997–6023. Association for Computational Linguistics, 2024. URL https://doi.org/10.18653/v1/2024.naacl-long.334. (Cited on page 154.)

[222] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic*, pages 4195–4205. Association for Computational Linguistics, 2021. URL https://doi.org/10.18653/v1/2021.findings-emnlp.354. (Cited on pages 2 and 69.)

[223] S. Wang, H. Kim, N. Janaka, K. Yue, H. Nguyen, S. Zhao, H. Liu, and K. Le. "what's this?": Understanding user interaction behaviour with multimodal input information retrieval system. In *Adjunct Proceedings of the 26th International Conference on Mobile Human-Computer Interaction, MobileHCI 2024, Melbourne, VIC, Australia*, pages 3:1–3:7. ACM, 2024. URL https://doi.org/10.1145/3640471.3680230. (Cited on page 118.)

[224] X. Wang, K. Zhou, J. Wen, and W. X. Zhao. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA*, pages 1929–1937. ACM, 2022. URL https://doi.org/10.1145/3534678.3539382. (Cited on page 15.)

[225] X. Wang, S. MacAvaney, C. Macdonald, and I. Ounis. Generative query reformulation for effective adhoc search. *CoRR*, abs/2308.00415, 2023. URL https://doi.org/10.48550/arXiv.2308.00415. (Cited on page 138.)

[226] Z. Wang and Q. Ai. Controlling the risk of conversational search via reinforcement learning. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia*, pages 1968–1977. ACM / IW3C2, 2021. URL https://doi.org/10.1145/3442381.3449893. (Cited on pages 116 and 118.)

[227] Y. Wei, Z. Wang, Y. Lu, C. Xu, C. Liu, H. Zhao, S. Chen, and Y. Wang. Editable scene simulation for autonomous driving via collaborative LLM-agents. *CoRR*, abs/2402.05746, 2024. URL https://doi.org/10.48550/arXiv.2402.05746. (Cited on page 91.)

[228] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain*, pages 132–141. ACM, 2009. URL https://doi.org/10.1145/1498759.1498819. (Cited on pages 120 and 122.)

[229] J. D. Williams and S. J. Young. Characterizing task-oriented dialog using a simulated ASR chanel. In *8th International Conference on Spoken Language Processing, INTERSPEECH-ICSLP 2004, Jeju Island, Korea*, pages 185–188. ISCA, 2004. URL https://doi.org/10.21437/Interspeech.2004-114. (Cited on page 19.)

[230] R. E. Wright. Logistic regression. In *Reading and Understanding Multivariate Statistics*, pages 217–244. American Psychological Association, 1995. (Cited on page 31.)

[231] C. Wu, S. C. H. Hoi, R. Socher, and C. Xiong. TOD-BERT: pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online*, pages 917–929. Association for Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.emnlp-main.66. (Cited on page 47.)

[232] J. Xu, Y. Wang, D. Tang, N. Duan, P. Yang, Q. Zeng, M. Zhou, and X. Sun. Asking clarification questions in knowledge-based question answering. 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 2019, Hong Kong, China*, pages 1618–1629. Association for Computational Linguistics, 2019. URL https://doi.org/10.18653/v1/D19-1172. (Cited on page 118.)

[233] W. Xu, G. Zhu, X. Zhao, L. Pan, L. Li, and W. Wang. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.826/. (Cited on page 2.)

[234] X. Yao, G. Bouma, and Y. Zhang. Semantics-based question generation and implementation. *Dialogue Discourse*, 3(2):11–42, 2012. URL http://dad.uni-bielefeld.de/index.php/dad/article/view/1439. (Cited on page 90.)

[235] Y. Yeh, M. Eskénazi, and S. Mehri. A comprehensive assessment of dialog evaluation metrics. Technical report, 2021. URL https://arxiv.org/abs/2106.03706. (Cited on page 1.)

[236] S. Yoon, Z. He, J. Echterhoff, and J. McAuley. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HLT 2024, Volume 1: Long Papers, Mexico City, Mexico*, pages 1490–1504. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.naacl-long.83. (Cited on page 93.)

[237] Y. Yuan, C. Shi, R. Wang, L. Chen, F. Jiang, Y. You, and W. Lam. Mcqueen: a benchmark for multimodal conversational query rewrite. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates*, pages 4834–4844. Association for Computational Linguistics, 2022. URL https://doi.org/10.18653/v1/2022.emnlp-main.320. (Cited on page 118.)

[238] Y. Yuan, C. Siro, M. Aliannejadi, M. de Rijke, and W. Lam. Asking multimodal clarifying questions in mixed-initiative conversational search. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore*, pages 1474–1485. ACM, 2024. URL https://doi.org/10.1145/3589334.3645483. (Cited on pages 92, 116, 118, 120, 138, 144, 146, and 153.)

[239] H. Zamani, S. T. Dumais, N. Craswell, P. N. Bennett, and G. Lueck. Generating clarifying questions for information retrieval. In *WWW '20: The Web Conference 2020, Taipei, Taiwan*, pages 418–428. ACM / IW3C2, 2020. URL https://doi.org/10.1145/3366423.3380126. (Cited on pages 89, 91, 92, 94, 97, 110, 111, 118, and 119.)

[240] H. Zamani, G. Lueck, E. Chen, R. Quispe, F. Luu, and N. Craswell. MIMICS: A large-scale data collection for search clarification. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland*, pages 3189–3196. ACM, 2020. URL https://doi.org/10.1145/3340531.3412772. (Cited on page 118.)

[241] H. Zamani, B. Mitra, E. Chen, G. Lueck, F. Diaz, P. N. Bennett, N. Craswell, and S. T. Dumais. Analyzing and learning from user interactions for search clarification. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China*, pages 1181–1190. ACM, 2020. URL https://doi.org/10.1145/3397271.3401160. (Cited on pages 89, 92, 116, and 118.)

[242] H. Zamani, J. R. Trippas, J. Dalton, and F. Radlinski. Conversational information seeking. *Found. Trends Inf. Retr.*, 17(3-4):244–456, 2023. URL https://doi.org/10.1561/1500000081. (Cited on page 89.)

[243] C. Zhang, Y. Chen, L. F. D'Haro, Y. Zhang, T. Friedrichs, G. Lee, and H. Li. Dynaeval: Unifying turn and dialogue level evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing ACL-IJCNLP 2021, Volume 1: Long Papers, Online*, pages 5676–5689. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.acl-long.441. (Cited on page 17.)

[244] C. Zhang, G. Lee, L. F. D'Haro, and H. Li. D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29(2329-9290):2502–2516, 2021. URL https://doi.org/10.1109/TASLP.2021.3074012. (Cited on page 12.)

[245] S. Zhang and K. Balog. Evaluating conversational recommender systems via user simulation. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA*, pages 1512–1520. ACM, 2020. URL https://doi.org/10.1145/3394486.3403202. (Cited on pages 13 and 65.)

[246] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics ACL 2018, Volume 1: Long Papers, Melbourne, Australia*, pages 2204–2213. Association for Computational Linguistics, 2018. URL https://aclanthology.org/P18-1205. (Cited on pages 17 and 19.)

[247] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr. (Cited on page 92.)

[248] W. Zhang, L. Li, D. Cao, and T. Liu. Exploring implicit feedback for open domain conversation generation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on*

Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, pages 547–554. AAAI Press, 2018. URL https://doi.org/10.1609/aaai.v32i1.11253. (Cited on pages 65 and 68.)

[249] Y. Zhang and X. Chen. Explainable recommendation: A survey and new perspectives. *Found. Trends Inf. Retr.*, 14(1):1–101, 2020. URL https://doi.org/10.1561/1500000066. (Cited on page 70.)

[250] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast, QLD, Australia*, pages 83–92. ACM, 2014. URL https://doi.org/10.1145/2600428.2609579. (Cited on page 70.)

[251] Y. Zhao, J. Wu, X. Wang, W. Tang, D. Wang, and M. de Rijke. Let me do it for you: Towards LLM empowered recommendation via tool learning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA*, pages 1796–1806. ACM, 2024. URL https://doi.org/10.1145/3626772.3657828. (Cited on page 93.)

[252] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html. (Cited on pages 92 and 94.)

[253] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J. Wen, and J. Yu. Improving conversational recommender systems via knowledge graph based semantic fusion. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA*, pages 1006–1014. ACM, 2020. URL https://doi.org/10.1145/3394486.3403143. (Cited on page 15.)

[254] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, and J. Wen. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online)*, pages 4128–4139. International Committee on Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.coling-main.365. (Cited on page 15.)

[255] K. Zhou, X. Wang, Y. Zhou, C. Shang, Y. Cheng, W. X. Zhao, Y. Li, and J. Wen. Crslab: An open-source toolkit for building conversational recommender system. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL 2021 - System Demonstrations, Online*, pages 185–193. Association for Computational Linguistics, 2021. URL https://doi.org/10.18653/v1/2021.acl-demo.22. (Cited on page 39.)

[256] S. Zhuang, X. Ma, B. Koopman, J. Lin, and G. Zuccon. PromptReps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval. *CoRR*, abs/2404.18424, 2024. URL https://doi.org/10.48550/arXiv.2404.18424. (Cited on page 91.)

[257] J. Zou and E. Kanoulas. Learning to ask: Question-based sequential bayesian product search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China*, pages 369–378. ACM, 2019. URL https://doi.org/10.1145/3357384.3357967. (Cited on page 118.)

[258] J. Zou, M. Aliannejadi, E. Kanoulas, M. S. Pera, and Y. Liu. Users meet clarifying questions: Toward a better understanding of user interactions for search clarification. *ACM Trans. Inf. Syst.*, 41(1): 16:1–16:25, 2023. URL https://doi.org/10.1145/3524110. (Cited on pages 116, 118, and 119.)

Summary

Conversational systems aim to help people interact with machines through natural language. These systems are increasingly important in our daily lives, from recommender systems suggesting personalized content to search systems helping users find information through dialogue. While traditional evaluation relies on metrics like clicks, evaluating conversational systems requires understanding complex human-system interaction patterns. This thesis investigated their evaluation through five research questions organized around two themes: understanding the evaluation of task-based conversational systems and advancing clarification in conversational search.

Under the first theme, we began by examining user satisfaction in CRS. Our analysis showed that satisfaction manifests differently at turn and dialogue levels – item relevance determines immediate satisfaction, while engagement and task completion influence overall satisfaction. Using these dialogue aspects as predictive features demonstrated their effectiveness in identifying successful and problematic interactions, confirming the importance of considering multiple interaction dimensions (Chapter 2). Given the reliance on crowdsourced labels as ground truth for evaluating conversational systems, we then investigated factors affecting annotation reliability. We found that dialogue context is essential for reliable evaluation, yet excessive dialogue context reduces annotator performance for aspects like usefulness. We demonstrated how LLMs can serve as annotation assistants, generating concise summaries that balance information richness with cognitive load (Chapter 3). We then examined how user feedback through follow-up utterances influences assessment quality, identifying complementary strengths between human and LLM evaluators. We showed that humans excel at capturing subjective aspects like usefulness in the presence of feedback from the user utterance, while LLMs provide consistent objective assessments (Chapter 4).

In our second theme, we investigated how CS systems can understand user needs through clarifying questions. We developed AGENT-CQ, demonstrating that LLMs can generate effective clarifying questions and provide reliable evaluation through simulated crowdsourcing. Our temperature-variation approach produced high-quality questions that improved retrieval performance compared to baselines (Chapter 5). We then analyzed how visual elements affect clarifying questions, finding that while users prefer image-enhanced questions, their effectiveness varies with task type and user expertise (Chapter 6).

Our findings indicate three key research directions. First, the development of hybrid evaluation frameworks that combine human and LLM capabilities while addressing challenges in task allocation and disagreement resolution. Second, the creation of context-aware systems that can detect and adapt to feedback signals from user utterances. Third, the establishment of more inclusive evaluation benchmarks representing broader linguistic and demographic diversity.

This thesis advances our understanding of evaluating and improving conversational systems through new methodologies for assessing satisfaction, frameworks for automated evaluation, and insights into multimodal interactions. These contributions lay the groundwork for developing more effective and user-centric conversational systems.

Samenvatting

Conversatiesystemen streven ernaar mensen te helpen communiceren met machines via natuurlijke taal. Deze systemen worden steeds belangrijker in ons dagelijks leven, van aanbevelingssystemen die gepersonaliseerde inhoud suggereren tot zoeksystemen die gebruikers helpen informatie te vinden via dialoog. Terwijl traditionele evaluatie vaak gebaseerd is op statistieken zoals klikken, vereist de evaluatie van conversatiesystemen inzicht in complexe interactiepatronen tussen mens en systeem. Deze thesis onderzoekt de evaluatie van conversatiesystemen aan de hand van vijf onderzoeksvragen, georganiseerd rond twee thema's: het begrijpen van de evaluatie van taakgerichte conversatiesystemen en het verbeteren van verduidelijking in conversatiegebaseerd zoeken.

Binnen het eerste thema begonnen we met het onderzoeken van gebruikerstevredenheid in conversatie-gebaseerde aanbevelingssystemen. Onze analyse toonde aan dat tevredenheid zich op verschillende niveaus manifesteert: op beurt- en dialoogniveau. De relevantie van aanbevelingen bepaalt de onmiddellijke tevredenheid, terwijl betrokkenheid en taakvoltooiing de algehele tevredenheid beïnvloeden. Het gebruik van deze dialoogaspecten als voorspellende kenmerken bleek effectief in het identificeren van zowel succesvolle als problematische interacties, wat het belang bevestigt van het in overweging nemen van meerdere interactiedimensies (Hoofdstuk 2). Aangezien de evaluatie van conversatiesystemen sterk afhankelijk is van door crowdsourcing verkregen labels als grondwaarheid, onderzochten we vervolgens factoren die de betrouwbaarheid van annotaties beïnvloeden. We ontdekten dat dialoogcontext essentieel is voor betrouwbare evaluatie, maar dat een overdaad aan context de prestaties van annotatoren vermindert. We toonden aan hoe *LLMs* kunnen dienen als annotatieassistenten door beknopte samenvattingen te genereren die informatierijkdom en cognitieve belasting in balans brengen (Hoofdstuk 3). Vervolgens onderzochten we hoe gebruikersfeedback in de vorm van vervolguitingen de beoordelingskwaliteit beïnvloedt. We identificeerden complementaire sterke punten tussen menselijke en LLM-evaluatoren: mensen blinken uit in het vastleggen van subjectieve aspecten wanneer ze feedback krijgen, terwijl LLMs consistente en objectieve beoordelingen bieden (Hoofdstuk 4).

Binnen het tweede thema onderzochten we hoe conversatiegebaseerde zoeksystemen gebruikersbehoeften beter kunnen begrijpen door middel van verhelderende vragen. We ontwikkelden AGENT-CQ en toonden aan dat LLMs effectieve verhelderende vragen kunnen genereren en betrouwbare evaluaties kunnen leveren via gesimuleerde crowdsourcing. Onze temperatuurvariatie-aanpak resulteerde in hoogwaardige vragen die de zoekprestaties verbeterden ten opzichte van basislijnen (Hoofdstuk 5). Vervolgens analyseerden we hoe visuele elementen van invloed zijn op verhelderende vragen. We ontdekten dat gebruikers de voorkeur geven aan vragen met afbeeldingen, maar dat de effectiviteit hiervan afhangt van het type taak en de expertise van de gebruiker (Hoofdstuk 6).

Onze bevindingen wijzen op drie belangrijke onderzoeksrichtingen. Ten eerste de ontwikkeling van hybride evaluatiekaders die menselijke en LLM-capaciteiten combineren, met aandacht voor uitdagingen zoals taakverdeling en meningsverschillen. Ten tweede de creatie van contextbewuste systemen die signalen uit gebruikersuitingen kunnen detecteren en zich daaraan kunnen aanpassen. Ten derde de totstandbrenging

van meer inclusieve evaluatiebenchmarks die een bredere taalkundige en demografische diversiteit vertegenwoordigen.

Dit proefschrift draagt bij aan een beter begrip van hoe conversatiesystemen geëvalueerd en verbeterd kunnen worden, door nieuwe methodologieën voor het meten van tevredenheid, kaders voor geautomatiseerde evaluatie en inzichten in multimodale interacties te ontwikkelen. Deze bijdragen leggen de basis voor de ontwikkeling van effectievere en gebruiksvriendelijkere conversatiesystemen.