



Evaluating Engaging Clarification Questions in Information Retrieval

*A thesis submitted in fulfilment of the requirements for the
degree of Doctor of Philosophy*

Leila Tavakoli

MSc in Software Engineering, University of Tehran, Iran

Bsc in Software Engineering, IAUSTB, Iran

School of Computing Technologies
College of Science, Technology, Engineering and Maths
RMIT University
Australia

August 2023

Declaration

I certify that except where due acknowledgement has been made, this research is that of the author alone; the content of this research submission is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

In addition, I certify that this submission contains no material previously submitted for award of any qualification at any other university or institution, unless approved for a joint-award with another institution, and acknowledge that no part of this work will, in the future, be used in a submission in my name, for any other qualification in any university or other tertiary institution without the prior approval of the University, and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of any published works contained within this thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my research submission to be made available on the web, via the University's digital research repository, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Leila Tavakoli

MSc in Software Engineering, University of Tehran, Iran

Bsc in Software Engineering, IAUSTB, Iran

02 August 2023

"To my parents and loving husband, whose endless love, encouragement, and belief in me have fueled my pursuit of knowledge and helped me overcome every obstacle along the way."

Acknowledgements

Undertaking a PhD during the unprecedented times of the Covid-19 pandemic presented an extraordinary set of challenges and uncertainties that shaped every aspect of my journey. Throughout this challenging period, I have been fortunate to receive invaluable support and encouragement from a multitude of individuals who played integral roles in guiding me towards the successful completion of my research. As I reflect upon this transformative experience, I am profoundly grateful to all those who stood by me, offering their support, wisdom, and understanding.

Foremost, I wish to convey my utmost appreciation to my supervisors, Professor Mark Sanderson, Professor Falk Scholer, Doctor Hamed Zamani, Doctor Johanne Trippas, and Professor Bruce Croft, whose invaluable guidance, profound comprehension, constant encouragement, and assistance made this endeavour a reality. Without their presence, none of this work would have been feasible.

Mark, your support and mentorship have been instrumental in shaping my professional and personal development throughout this journey. Since our initial correspondence in 2018, your invaluable guidance, insightful advice, and exemplary conduct as a mentor have far exceeded my expectations. Beyond the realm of research, you have imparted valuable lessons, teaching me the importance of embracing diverse perspectives and approaching challenges from multiple angles. It is through your mentorship that I have grown not only as a researcher but also as an individual.

Falk, I was fortunate to have had you as my supervisor throughout these past four years. Your insights, valuable feedback, and constructive criticism have been pivotal in shaping and refining my ideas, significantly deepening my comprehension of various facets of my field of study.

Hamed, I am grateful for the privilege of having you as my supervisor. Your friendly and compassionate support has shaped the quality of my work. Whenever faced with daunting research challenges, your guidance and input have instilled in me a sense of confidence and assurance that they were indeed conquerable.

I hold the appreciation for the support provided by Johanne. Your feedback and support throughout the final year of my PhD have been truly invaluable. I am sincerely grateful for your continuous availability and dedication to adding value to my research.

I also consider myself incredibly fortunate to have had the privilege of being mentored

by Professor Bruce Croft as my supervisor.

I appreciate Dr Damiano Spina, who served as the examiner for all three milestones of my work. Your knowledge and thought-provoking questions during our weekly group meetings (CHIRE) and milestone sessions have consistently broadened my horizons.

Being a member of the CIDDA Human Information REtrieval (CHIRE) has been an extraordinary experience that has greatly enriched my journey. I am deeply grateful to Prof. Mark for providing me with the opportunity to serve as the organiser of these meetings. Additionally, I would like to express my appreciation to all the members for their support. Marwah, Nuha, Oleg, and Sachin, I am truly grateful for your friendship and support throughout this endeavour.

I am grateful to Professor Zahir Tari and Professor Zhifeng Bao for extending the invitation to serve as the HDR representative of the IR group during the second year of my PhD journey, which coincided with the challenging period of the Covid-19 pandemic.

I am appreciative of the help and support I received from A/Professor Jeffery Chan, who was the panel chair in my milestones and the Higher Degrees by Research Manager of our discipline during the course of my PhD.

I would like to extend my deepest gratitude to my beloved husband, Amir, for the unwavering love, support, and understanding he has shown me throughout my entire PhD journey. Your presence as someone who has gone through this very journey has been invaluable. Thank you for being my rock, my confidant, and my biggest cheerleader. Your love and your support have made all the difference in my journey, and I am forever grateful for your presence in my life.

I would like to seize this opportunity to convey my sincerest appreciation to my parents for their constant backing, affection, and motivation during both my PhD endeavour and my lifelong pursuit of education. Dad, from the very beginning, you instilled in me a thirst for knowledge, nurturing my curiosity and guiding me towards intellectual growth. Mom, your sacrifices have allowed me to pursue my dreams, and I am forever grateful for the countless opportunities you have provided. Words cannot adequately express the depth of my appreciation for the immeasurable impact you have had on my life and academic journey. Thank you, Mom and Dad, for being my pillars of strength, my role models, and my unwavering support system.

I would like to express my heartfelt gratitude to my brother, Mohammad, who has been a true friend throughout my PhD journey and every step of my life and academic pursuits.

I would like to extend my heartfelt appreciation to my beloved Mother-in-law, Ferdows, and my dear sisters, Nadia and Delara, for their support throughout these years. A special word of gratitude goes to my dear sisters-in-law, Sarvenaz and Ayda, and also my brothers-in-law, Reza, Shahin, and Aidin. Having you as part of my family has been a blessing, and I consider myself incredibly fortunate to have you all in my life.

Lastly, I want to express my eternal gratitude to my niece, Rose, and my nephews, Ario and Leo. Your laughter, innocence, and love have provided a source of inspiration and motivation, reminding me to find joy in the smallest moments.

Contents

| | |
|--|-------------|
| Declaration | iii |
| Acknowledgements | vi |
| List of Figures | xi |
| List of Tables | xiii |
| Abstract | 1 |
| 1 Introduction | 5 |
| 1.1 Motivation | 6 |
| 1.2 Contributions | 10 |
| 1.3 Thesis Structure | 12 |
| 2 Background | 15 |
| 2.1 Human-generated Clarification Questions | 16 |
| 2.2 Clarification Models | 17 |
| 2.2.1 Clarification Selection Models | 17 |
| 2.2.2 Clarification Generation Models | 19 |
| 2.3 Clarification Datasets | 21 |
| 2.3.1 CQA Clarification Datasets | 22 |
| 2.3.2 Search Clarification Datasets | 22 |
| 2.4 Online and Offline Evaluation in Information Retrieval | 24 |
| 3 Useful Clarification Questions in Community Question Answering Forums | 29 |
| 3.1 Introduction | 29 |
| 3.2 Methodology | 31 |
| 3.2.1 Identifying Potential Clarifications | 32 |
| 3.2.2 Annotation and Data Sampling | 33 |
| 3.3 Results and Analysis | 38 |

| | | |
|----------|--|-----------|
| 3.3.1 | User Engagement and Clarification | 39 |
| 3.3.2 | Useful Clarifications | 41 |
| 3.3.3 | Clarification Types and Patterns | 44 |
| 3.4 | Discussion | 50 |
| 3.5 | Summary | 51 |
| 4 | Asking Engaging Clarification Question in Search Engines: Task Formulation and Limitations | 53 |
| 4.1 | Introduction | 53 |
| 4.2 | Methodology | 55 |
| 4.2.1 | Task Formulation | 56 |
| 4.2.2 | Experimental Results | 58 |
| 4.2.3 | Discussion: Limitations of Existing Resources | 64 |
| 4.3 | Summary | 66 |
| 5 | Introducing MIMICS-Duo: A Dataset for Online and Offline Evaluation of Search Clarification | 67 |
| 5.1 | Data Sampling from MIMICS-ClickExplore | 68 |
| 5.2 | Task Design | 69 |
| 5.3 | Pilot Tasks | 71 |
| 5.4 | Quality Assurance and Attention Measures | 72 |
| 5.5 | Crowd-sourcing | 72 |
| 5.6 | AMT Workers' Feedback | 73 |
| 5.6.1 | MIMICS-Duo Dataset Analysis | 74 |
| 5.7 | Research Enabled by MIMICS-Duo | 76 |
| 5.8 | Summary | 78 |
| 6 | Online and Offline Evaluation in Search Clarification | 81 |
| 6.1 | Introduction: Current Practice and Knowledge Gap | 82 |
| 6.2 | Methodology | 84 |
| 6.2.1 | Dataset | 84 |
| 6.2.2 | Experimental Design | 86 |
| 6.2.3 | Data Analysis | 86 |
| 6.2.4 | Evaluation Metrics | 89 |
| 6.3 | Results | 91 |
| 6.3.1 | Overall Practice in Designing Online and Offline Evaluations in Search Clarification | 92 |
| 6.3.2 | Impact of Query Length on the Relationship Between Online and Offline Evaluations | 95 |
| 6.3.3 | Impact of Uncertainty on the Relationship Between Online and Offline Evaluations | 95 |
| 6.3.4 | The Most vs. the Least Engaging Panes | 98 |
| 6.3.5 | Manual Clarification Pane Inspection | 99 |
| 6.4 | Discussion | 100 |
| 6.5 | Summary | 102 |

| | | |
|----------|--|------------|
| 7 | Understanding Modality Preferences in Clarification Questions | 105 |
| 7.1 | Introduction | 106 |
| 7.2 | Experimental Design | 108 |
| 7.3 | Results | 111 |
| 7.4 | Discussion and Summary | 115 |
| 8 | Conclusions and Future Work | 117 |
| 8.1 | Thesis Contributions | 118 |
| 8.2 | Discussion and Summary | 120 |
| 8.3 | Future Directions | 122 |
| | Bibliography | 125 |
| | Appendix A Instructions and Examples of Crowd-sourcing Tasks | 141 |
| | Appendix B Publications | 165 |
| | Appendix C Ethics Approval Letter | 167 |

List of Figures

| | | |
|------|---|----|
| 3.1 | A question posted on <i>Stack Exchange</i> . (<i>Asker</i> : who posted the initial question, <i>Responder</i> : another user provided the answer, <i>Accepted Answer</i> : an answer among all provided answers by other users that were chosen by the Askers.) | 30 |
| 3.2 | The flow of study of the study in this chapter. | 31 |
| 3.3 | Clarification receiving an informative answer. | 34 |
| 3.4 | Clarification receiving an uninformative answer. | 35 |
| 3.5 | A useless answer due to a misunderstanding. | 36 |
| 3.6 | A clarification asked to eliminate ambiguity. | 37 |
| 3.7 | A non-valuable clarification question. | 37 |
| 3.8 | A clarification that enhances the accepted answer. | 38 |
| 3.9 | Percentage of posts with an accepted answer, grouped by Answerer. . . . | 39 |
| 3.10 | Number of posts with an accepted answer grouped by the % of questions answered by the Asker. | 40 |
| 3.11 | Fraction of answered clarifications per question in the Quantitative Finance site. | 40 |
| 3.12 | Fraction of answered clarifications per question in English Language site. . | 41 |
| 3.13 | Fraction of answered clarifications per question in Science Fiction site. . . | 41 |
| 3.14 | Characteristics of clarification questions answered either by the Asker or a Responder. | 43 |
| 3.15 | Distribution of clarifications by type. | 45 |
| 3.16 | The probability of a clarification question that has a certain type, given that the clarification question is answered by the Asker. | 46 |
| 3.17 | The probability of a clarification question being answered by the Asker, given a particular clarification question type. | 46 |
| 3.18 | The probability of a post having a particular clarification question type, given that the post has an accepted answer. | 47 |
| 3.19 | The probability of a post having an accepted answer, given the presence of a clarification question of a specific type. | 47 |
| 3.20 | Popular clarification patterns grouped by type. | 49 |

| | | |
|------|--|-----|
| 3.21 | The top ten and bottom ten patterns by point-wise <i>KL-Divergence</i> , ($P(x)$) is popularity distributions of the patterns of the useful clarification questions and ($Q(x)$) is popularity distributions of the non-useful clarification questions. | 49 |
| 4.1 | Example of query and clarification pane (i.e., A clarification question plus candidate answers). | 54 |
| 4.2 | Performance of LTR models after mapping the <i>Engagement Level</i> considering the related search- and video-based features. | 61 |
| 4.3 | Performance of LTR models after mapping the <i>Engagement Level</i> without considering the related search- and video-based features. | 62 |
| 5.1 | An overview of the three steps of the data collection. | 69 |
| 5.2 | Quality Label vs. Engagement Level. | 76 |
| 5.3 | Mean values of different aspect labels for clarification panes with various overall quality. | 77 |
| 6.1 | Two examples of clarification pane rank lists based on the <i>Engagement Level</i> and <i>Coverage</i> labels for a query. | 86 |
| 6.2 | The prompt template used to feed the GPT model. | 88 |
| 6.3 | Variations of <i>Overall Quality</i> (OQ), <i>Coverage</i> (Cov), <i>Diversity</i> (Div) and the number of candidate answers (# Ans) in the MECPs when compared to the LECPs. (§: means the percentage of the MECPs that have higher <i>Overall Quality</i> than the LECPs is significantly different, Student's t-test, $p<0.05$). | 99 |
| 7.1 | An example of Task II (T vs. MM). | 109 |
| 7.2 | Questionnaire template. | 110 |

List of Tables

| | | |
|-----|---|----|
| 3.1 | The analysed sites of <i>Stack Exchange</i> | 32 |
| 3.2 | Sample Size (Number of investigated Potential clarification question). . . | 38 |
| 3.3 | Who answers clarification questions. | 39 |
| 3.4 | Percentage of Actual Clarification Questions. | 42 |
| 4.1 | Statistics of the datasets. | 57 |
| 4.2 | LTR Features and Feeding Inputs. | 58 |
| 4.3 | Performance of LTR models based on P@1, trained and tested on <i>MIMICS-ClickExplore</i> . The ground truth is the ranked lists based on the <i>Engagement Level</i> | 59 |
| 4.4 | Performance of LTR models based on P@1, trained and tested on <i>MIMICS-Manual</i> . The ground truth is the ranked lists based on the <i>Overall Quality</i> label. | 60 |
| 4.5 | Performance of LTR models based on P@1, trained and tested on <i>MIMICS-ClickExplore</i> , without considering the related search- and video-based features. The ground truth is the ranked lists created using the <i>Engagement Level</i> | 60 |
| 4.6 | Performance of LTR models based on P@1, trained and tested on <i>MIMICS-Manual</i> , without considering the related search- and video-based features. The ground truth is the ranked lists created using the <i>Overall Quality</i> label. | 61 |
| 4.7 | Performance of LTR models trained on <i>MIMICS-ClickExplore</i> . (Significance test results are explained in the text.) | 63 |
| 4.8 | Performance of LTR models trained on <i>MIMICS-Manual</i> . (Significance test results are explained in the text.) | 64 |
| 4.9 | The top 10 features with the highest weight learned by <i>RankBoost</i> from <i>MIMICS-ClickExplore</i> and <i>MIMICS-Manual</i> | 64 |
| 5.1 | Statistics of MIMICS-Duo dataset. | 68 |
| 5.2 | Distribution of the quality label of clarification panes and their candidate answers. | 75 |
| 5.3 | Distribution of the characteristics label of clarification panes. | 77 |

| | | |
|-----|---|-----|
| 5.4 | Correlations between online and offline measures. (<i>Cov</i> , <i>Div</i> , <i>Und</i> , <i>IO</i> , <i>OQ</i> , <i>EL</i> , <i>OR</i> and <i>#Ans.</i> stand for <i>Coverage</i> , <i>Diversity</i> , <i>Understandability</i> , <i>Importance Order</i> , <i>Engagement Level</i> , <i>Offline Rating</i> and <i>Number of Candidate Answers</i> , respectively.) | 78 |
| 6.1 | Relationships between the ranked lists of clarification panes created by the <i>Engagement Level</i> and created by offline labels. | 92 |
| 6.2 | Evaluation of three GPT-3.5 configurations across varying temperature settings and five LTR models, utilising offline labels to generate ranked lists of clarifications. | 94 |
| 6.3 | Impact of the query length on relationships between the ranked lists of clarifications created by the <i>Engagement Level</i> and created by offline labels. (Short Query: 126 queries with 415 query-clarification pairs; Long Query: 180 queries with 619 query-clarification pairs.) | 96 |
| 6.4 | Impact of the <i>Impression Level</i> on relationships between the ranked lists of clarifications created by the <i>Engagement Level</i> and created by offline labels. | 97 |
| 6.5 | Impact of the <i>Impression Level</i> on the performance of three GPT-3.5 configurations across varying temperature settings.) | 98 |
| 6.6 | Examples queries and their most and least engaging clarification panes. | 100 |
| 6.7 | Examples queries with online and offline labels. | 100 |
| 7.1 | Pairwise preference for clarification modality (%) | 112 |
| 7.2 | Motivations behind user preference (%). | 113 |
| 7.3 | Comparison of human-collected and computer-generated search clarification question images. | 114 |

Abstract

Information-seeking systems for natural language questions often encounter a range of grammatically complex queries presented in unpredictable ways. Users often need to rephrase their questions in order to obtain a satisfactory answer, which can be both demanding and time-consuming. One solution to this challenge involves prompting clarification questions when a query is intricate or ambiguous. It is widely acknowledged that if a search system can ask clarification questions to better understand the user’s intention, the chances of retrieving a satisfactory answer are higher. While clarification plays a vital role in conversational and interactive information-seeking systems, previous studies have indicated that users are not easily engaged with these clarification questions despite their positive impact. To improve the performance of such models, it is crucial to employ evaluation methods that take into account user behaviour and the characteristics of engaging clarification questions.

Currently, there is limited understanding of clarification questions from a user’s perspective, particularly what makes a clarification question engaging. This understanding is crucial since a clarification question is only valuable when the user actively engages with it. To address these knowledge gaps, we conduct a series of experiments to analyse user behaviour when interacting with clarifications on various information-seeking system platforms. Our initial analysis focuses on human-generated clarification questions to gain insights into how they are employed to disambiguate queries and better understand information needs. By identifying the most useful clarification questions, we analyse their characteristics in terms of types and patterns, comparing them with non-useful clarifications. Our analysis reveals that the most useful clarification questions exhibit consistent patterns across different topics.

Next, we expand our study to clarification questions in search engines by examining the *MIMICS* dataset, the only available dataset containing real search clarifications, including

information about user engagement and the quality of clarification questions. This research phase aims to investigate the task of identifying the most engaging clarification question from multiple clarifications generated for a given query in a search engine. In cases where multiple clarification questions are available, we frame this task as a learning-to-rank (LTR) problem, utilising various information such as the query itself, clarification questions, candidate answers, and search engine results page (SERP) information. Furthermore, we demonstrate the scarcity of query-clarification pairs with both online and offline evaluations in the dataset, which impedes drawing robust conclusions regarding the impact of online and offline evaluations on search clarification and identifying the most engaging clarification panes from a user’s perspective. Our experiments unveil the limitations of the *MIMICS* dataset in search clarification, motivating us to introduce a new search clarification dataset called *MIMICS-Duo* in the subsequent phase.

Building upon the *MIMICS*, *MIMICS-Duo* facilitates multi-dimensional evaluation of search clarification. This dataset encompasses 306 search queries accompanied by multiple clarifications, fine-grained annotations on clarification questions (including quality and aspect labels) and offline ratings. Using the *MIMICS-Duo* dataset, we explore further the task of identifying the most engaging clarification question for a given query and extensively investigate the relationship between online and offline evaluations, an area that has been largely unexplored in the existing literature. In contrast to the prevailing belief that offline evaluations are inadequate for supporting online evaluations, we observe that offline evaluations align with online evaluations when it comes to identifying the most engaging clarification question among multiple clarifications generated for a given query. We further investigate the impact of the query length and the low uncertainty in the online evaluation on the relationship between offline and online evaluations.

In addition, we explore the impact of human labelling on improving the performance of Large Language and LTR models in identifying the most engaging clarification questions from the user’s point of view. This is achieved by incorporating offline evaluations as input features. We show that LTR models do not outperform individual offline labels. However, GPT stands out as the top performer, surpassing all Learning-to-Rank models and offline labels.

Finally, we explore how recent advancements in technology in terms of implementing different modalities in search clarification can enhance user engagement with the clarification questions. *Multi-modal* clarification approach involves incorporating multiple media types, such as text and image, to refine and enhance search results. We investigate user preferences regarding the modality of clarification and demonstrate that, in most cases, users prefer multi-modal clarifications over those using only one modality. Additionally, we explore the task of automatically generating corresponding images and show that text-

to-image generation systems like Stable Diffusion can be utilised to generate multi-modal clarification questions.

In conclusion, this research focuses on understanding what makes a clarification question engaging from a user's perspective, emphasising the need for user engagement to derive value from these questions. Overall, these findings contribute to the advancement of information-seeking systems and provide insights into user behaviour and the characteristics of engaging clarification questions.

Keywords: Clarification Question, Community Question Answering, Conversational Information Seeking, Asynchronous Conversation, Ranking Clarification Question, Multi-Modal Clarification Question, Learning-to-Rank, Large Language Model, Text-to-Image Model, Clarification Dataset

Chapter 1

Introduction

Search engines employ complex algorithms to provide appropriate results based on a given query. However, the queries themselves can be unclear, disorganised, and filled with irrelevant information, posing a challenge for computers to accurately interpret the intended significance. While search engines have improved over time, users may still need to scan multiple result pages or reformulate their queries to find the information they need. To address this challenge, one approach is to ask clarification questions (CQs) to clarify user information needs. Recent studies have shown that this approach can provide functional and emotional benefits for users [122]. Clarification questions are particularly useful in limited bandwidth interfaces, such as small-screen devices and speech-only conversational agents. However, it is important to balance the benefits of CQs with the risk of frustrating the user by asking too many or unsatisfactory questions, which can harm the success of the search system. The literature has discussed the advantages of implementing CQs in various fields, including dialogue systems [57, 32, 6], community question answering [14, 89, 62, 104], conversational search systems [4, 122, 32, 132], and speech recognition [100].

Understanding how people interact in information-seeking systems is crucial for enhancing confidence in retrieved information [4] and developing successful information-seeking systems [82]. In fact, it is considered one of the final goals of information retrieval [83]. The generating and asking of CQs have been an area of interest for several years, with initial works conducted by van Beek et al. in 1993 [109] and followed by Moldovan and Harabagiu [77], Voorhees [110], and De Boni and Manandhar [29] for open-domain question answering systems. However, users are often reluctant to engage with CQs, and there

is a lack of understanding of what makes a clarification engaging from a user’s perspective. To improve the performance of information-seeking systems by generating and selecting CQs that are likely to engage the users more proactively, we investigate CQs from various perspectives.

1.1 Motivation

Clarification is the process of clarifying and refining the user’s search intent to produce more relevant results. While clarification has become a core component of information-seeking systems [128], previous research has shown that even though CQs receive positive engagement, users are not easily engaged with them [126, 124]. Despite increased attention to search clarification [14, 88, 4, 57], there needs to be more research into enhancing user interaction with CQs. A lack of knowledge on what makes a clarification helpful and engaging from a user’s perspective motivates us to study CQs in the context of two different information-seeking systems, including search engines and community question-answering (CQA) forums. This research aims to shed light on how users perceive and interact with clarifications in different contexts and to provide valuable insights into the effectiveness of CQs in both traditional information retrieval platforms and community-driven knowledge-sharing environments. Understanding what factors contribute to a clarification being helpful and engaging will have practical implications for the design and implementation of information-seeking systems. It could lead to the development of more effective and user-friendly CQ strategies, which in turn can improve search accuracy and user satisfaction with search results.

Evaluating the effectiveness of CQs is essential to understand their impact on user search experiences. Two primary evaluation methods are commonly used: manual human judgements (offline evaluations) and actual user interaction data such as click-through rate (CTR) analysis (online evaluations). However, there is an ongoing debate in the literature regarding the consistency between online and offline evaluations in assessing retrieval quality [27, 33, 38, 94]. Previous research has shown that online and offline assessments can lead to different results in the context of information retrieval quality. This divergence is attributed to various factors, including user behaviour, system biases, and experimental settings. Despite this existing body of work, it remains uncertain to what extent this divergence applies specifically to the evaluation of search clarification. The uncertainty surrounding the consistency between online and offline evaluations in the context of search clarification creates a significant gap in our understanding. Addressing this gap is crucial for several reasons:

- **Validity of Evaluation Metrics:** Understanding whether online and offline evaluations

align or diverge in the context of CQs is essential to validate the evaluation metrics used to assess retrieval quality. This knowledge is fundamental for researchers and developers who rely on evaluation methods to make informed decisions about the effectiveness of their CQ strategies.

- **Optimising Search Systems:** Identifying the factors contributing to the divergence between online and offline evaluations of CQs can provide valuable insights into the strengths and limitations of each evaluation approach. This knowledge can be leveraged to optimise information-seeking systems for better search performance and user satisfaction.
- **User-Centric Design:** By uncovering the discrepancies between online and offline evaluations, researchers can gain a deeper understanding of how users interact with CQs in real-world scenarios. This knowledge can inform user-centric design principles, allowing for the creation of more user-friendly CQ mechanisms.
- **Advancement of Information Retrieval Research:** Addressing the controversy surrounding online and offline evaluations in the context of search clarification can contribute to the broader field of information retrieval research. It can lead to methodological improvements and better align evaluation practices with real-world user needs and behaviours.

This research aims to resolve the uncertainty regarding the consistency between online and offline evaluations of CQs in information retrieval. However, effective evaluation of search clarification methods requires suitable resources that reflect the diverse aspects of user interactions and system performance. Currently, the available search clarification datasets have limitations in capturing the full spectrum of evaluation dimensions. These datasets are either based on user interaction signals, such as click-through rate (CTR), obtained from online sources [125] or created through a manual annotation process (offline dataset) [4]. Both types of datasets have inherent drawbacks that hinder a comprehensive evaluation of search clarification methods:

- **Insufficiency of Online Datasets:** Online datasets derived from user interaction signals provide valuable insights into real-world user behaviour. However, they often lack the diversity needed for a thorough evaluation. Since these datasets are generated from actual user interactions, they might be influenced by the current system's limitations or biases. Consequently, they may not adequately cover various scenarios and user intentions, limiting their effectiveness in evaluating the robustness and generalisability of search clarification models.

- **Limitations of Offline Datasets:** On the other hand, offline datasets created through manual annotation processes are valuable for controlled experiments and benchmarking purposes. However, they may not fully capture the complexities of real-world user interactions, making it challenging to assess the practical utility of search clarification methods in actual information-seeking systems. Moreover, the manual annotation process itself can introduce subjectivity and biases, potentially affecting the reliability of the evaluations.
- **Deployment Pipeline Challenges:** The deployment of clarification models in information retrieval systems involves a pipeline that combines offline evaluation with manual annotations and online evaluation through A/B testing based on real user interactions. The shortcomings of the available datasets create a barrier in this deployment pipeline. Without comprehensive datasets that encompass a wide range of user intents and interactions, it becomes challenging to effectively assess the efficacy and impact of CQs in real-world settings.

A multi-dimensional evaluation, combining both online and offline evaluation approaches, is essential to gain a comprehensive understanding of how CQs perform across various scenarios and user contexts. By conducting research to develop more diverse and representative datasets for search clarification, this study aims to: (i) improve the reliability and validity of evaluation metrics for search clarification methods, (ii) enable a more accurate assessment of system performance and user satisfaction with CQs, and (iii) facilitate the development and optimisation of information retrieval systems by leveraging real-world user behaviours and intentions.

Finally, in the context of search clarification, previous studies have primarily focused on the use of text-based interactions. However, in real-world conversational information-seeking systems, user interactions often involve multiple modalities, such as text and images. The recent Alexa Prize TaskBot Challenge [1] has highlighted the significance of *multi-modal* interactions in shaping the user experience [31]. Consequently, understanding the impact of *multi-modal* clarification on user interaction in search engines has become a compelling area of investigation. Several factors contribute to the importance of studying *multi-modal* clarification in search engines:

- **Realistic User Interactions:** In today's information-seeking landscape, users are increasingly engaging with search engines using various modalities, including text-based queries, voice commands, and images. To create effective information retrieval systems, it is essential to adapt to and support these diverse user interaction patterns. Therefore, exploring how *multi-modal* clarification can enhance user interactions is critical for creating more realistic and user-friendly search experiences.

- **Leveraging Rich Information Sources:** Images and other non-textual modalities can provide additional context and information that text alone may not capture. By incorporating *multi-modal* clarification, search engines can leverage this rich source of information to better understand and refine user queries, leading to more relevant search results and enhanced user satisfaction.
- **Emerging Research Frontier:** While there is increasing interest in multi-modal AI and human-computer interactions, there is still limited research on how *multi-modal* clarification specifically impacts user interaction in search engines. Addressing this research gap can contribute to the growing body of knowledge in multi-modal information retrieval and conversational search systems.

We aim to investigate the extent to which *multi-modal* clarification enhances user interaction in search engines. By exploring the integration of text and image-based interactions, the research seeks to identify the potential benefits and challenges of incorporating *multi-modal* clarification strategies. The findings can inform the design and development of more effective and inclusive information retrieval systems that cater to diverse user preferences and interaction styles. The motivation behind this research lies in the increasing prevalence of *multi-modal* interactions in conversational information-seeking systems and the need to understand how *multi-modal* clarification can enhance user engagement in search engines.

In this research, we first analyse the CQs that are generated by humans on the *Stack Exchange* platform, a CQA (Community Question Answering) website, in Chapter 3. The analysis provides insights into how these questions are used to remove ambiguity and improve understanding of information needs. We extract a set of CQs from the posts on the platform and define a new taxonomy for annotating these questions and their responses. The usefulness of the CQs is evaluated based on whether they add any new information to the original post and the accepted answer. We also identify the characteristics of the useful CQs in terms of their types and patterns.

Next, we explore clarification questions in a search engine using *MIMICS* dataset, the only available dataset in search clarification, in Chapter 4. We focus on the task of identifying the most engaging clarification question (MECQ) from several questions generated for a particular query in Microsoft *Bing* using several learning-to-rank (LTR) models. We show that the *MIMICS* dataset has several limitations to performing this study that motivate us to develop a multi-dimensional dataset, known as *MIMICS-Duo*, for search clarification in Chapter 5.

Introducing the *MIMICS-Duo* dataset provides us with the opportunity to investigate the relationship between online and offline evaluations in search clarification in Chapter 6. We examine the effectiveness of an Oracle clarification selection model that has access to every

offline label in predicting the online label. Additionally, we explore how combinations of offline labels (such as using random forests and neural networks) affect online evaluation. We also investigate the impact of query length and uncertainty in collected online labels on the correspondence between offline and online labels. We further investigate whether implementing offline labels as input features can enhance the performance of Large Language Models (LLMs) and LTR models in identifying the MECPs.

Finally, we study the user preference over clarification modality in Chapter 7 and evaluate the performance of several text-to-image generation models in creating relevant images for *text-only* CQs to make *multi-modal* clarification questions.

1.2 Contributions

This thesis enhances the current knowledge of clarification questions in information-seeking systems by answering the following research questions. We first analyse the CQs in a community question answering (CQA) forum in Chapter 3 to answer two research questions of (i) what CQs are more useful in terms of helping the Asker to get a correct answer? And (ii) what are the characteristics of such CQs (i.e., type and pattern)? Our key contributions from answering the first two research questions are:

- Presenting a new taxonomy to investigate the usefulness of clarification questions.
- Examining the relationship between posts with accepted answers and different types of answerers.
- Detecting useful and non-useful clarification questions.
- Extracting and identifying the types and patterns of useful and non-useful clarification questions.

The research presented in this chapter resulted in the following publications:

- Tavakoli, L., Zamani, H., Scholer, F., Croft, W. B., & Sanderson, M. (2022). Analysing clarification in asynchronous information-seeking conversations. *Journal of the Association for Information Science and Technology*, 73(3), 449-471.

We then investigate the clarifications in a search engine to address the third and fourth research questions: (iii) Can SERP feature help us identify the most engaging clarification question from a user's Perspective? and (iv) Is there any relationship between online and offline evaluations in search clarification using the *MIMICS* dataset, the only search clarification dataset? We show that this research question cannot be addressed using the available clarification dataset. The observed limitations motivate us to introduce a new

search clarification question dataset, which is one of our main contributions. This newly created dataset contains a series of quality and rating annotations in addition to CTR information for 1,034 query-clarification pairs. The methodological outcomes of creating this dataset include proposing a new data collection setup, questionnaires, crowd-sourcing framework and validation process to collect various characteristics labels for the CQs. This dataset helps us establish the relationships between different aspects of clarification panes, which can be used for further improvement of generating and asking clarification models.

The research presented in Chapters 4 and 5 resulted in the following publication:

- Tavakoli, L., Trippas, J. R., Zamani, H., Scholer, F., & Sanderson, M. (2022, July). MIMICS-Duo: Offline & Online Evaluation of Search Clarification. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 3198-3208).

We explore the created dataset to address the next three research questions as follows: (v) What are the best overall practices in designing offline evaluation methodologies for search clarification that correspond with online evaluation? (vi) Does query length impact the relationship between online and offline evaluations in search clarification? And (vii) Does uncertainty in the online evaluation impact the relationship between online and offline evaluation? The contributions of this research questions above can be listed as follows:

- Initially, we examine the performance of an Oracle clarification selection model that has access to individual offline labels and its correlation with the online label. Moving beyond the assumption of independence among offline labels, we explore their combination through methods like random forests and neural networks, aiming to determine if such combinations align with online evaluations. Furthermore, we leverage a Large Language Model (LLM) to predict user engagement with the clarification in an online setting, taking into account the available offline labels.
- Inspired by the findings of Zamani et al. [127]), which demonstrated variations in user behaviour between short and long queries, we proceed to partition our dataset accordingly. Our objective is to examine the extent to which offline labels align with online evaluations, specifically for short and long queries.
- To manage uncertainty in online evaluation, we utilise the concept of impression count, which refers to the number of times a clarification question (CQ) is presented to users during A/B testing. A higher impression count corresponds to a more reliable and less uncertain online label, as it is based on the click-through rate observed.

The last research question in this study that we aim to explore is: (viii) do users prefer *multi-modal* clarification questions over uni-modal (i.e., textual or visual)? We analyse three

different modalities of (i) *textual*, (ii) *visual*, and (iii) *multi-modal* (i.e., a combination of the two) for randomly sampled CQs from the *MIMICS-Manual* dataset that is introduced in this thesis and investigate various influential factors on user preference. Finally, we investigate whether generating corresponding images to the clarification panes can be automated using text-to-image generation models. The quality and the relevance of generated images, in addition to user preferences over human-collected and computer-generated images, are investigated through manual annotation. The contribution of this study, presented in Chapter 7, can be summarised as follows:

- **Exploration of *Multi-modal* Search Clarification:** The study addresses the research gap in understanding the impact of *multi-modal* interactions on user experience in search engines. Previous studies on search clarification have focused solely on text-based interactions, while this study investigates the potential benefits of incorporating visual elements alongside textual content.
- **User Preference Analysis:** Through a crowd-sourced user study, the research examines user preferences for three different modalities of clarification questions (CQs): *textual*, *visual*, and *multi-modal* (combining text and image). The study analyses the impact of these modalities on user preferences and explores the factors influencing those preferences.
- **Fine-grained Analysis:** The research analyses various factors related to user preferences, including image quality, image/text clarity, the relevance of text and image, and other image aspects. Furthermore, the research explores the feasibility of automating the process of generating images for *multi-modal* clarification questions (CQs) by leveraging text-to-image generation models.

1.3 Thesis Structure

The thesis consists of seven further chapters to cover the conducted research and address the research questions mentioned beforehand. The background on human- and computer-generated CQs in questions-answering forums, search engines and information-seeking systems in terms of ranking, generating and asking CQs are discussed thoroughly in Chapter 2. This chapter also provides a detailed review of previous studies on online and offline evaluations of retrieval quality.

Chapter 3 focuses on characteristics of useful CQs in CQA forums. We propose a new taxonomy to investigate the usefulness of CQs. We then study the types and patterns of useful and non-useful CQs in this chapter.

In Chapter 4, we investigate the task of identifying the MECQ for a given query using various learning-to-rank (LTR) models. The CQs collected in the *MIMICS* dataset, and their online and offline corresponding labels are used for training and testing the models. The limitations of this dataset in offline and online evaluations in search clarification are also discussed.

In Chapter 5, a new search clarification dataset, called *MIMICS-Duo*, is introduced. This new dataset overcomes the limitations of the *MIMICS* dataset. This dataset contains a series of quality and rating labels in addition to user engagement level, an online label. We present the details of the data collection and the design of the experiments. We analyse the properties of the dataset and show how this newly-developed dataset can help to establish the relationships between different aspects of clarification panes.

The next stage of this research is to present the extensive analyses that we conducted on the *MIMICS-Duo* dataset. In Chapter 6, we investigate formulations of offline labelling and their relations with online evaluation based on click-through rate.

In Chapter 7, we introduce the *MIMICS-MM* dataset and investigate the impact of clarification modality on user preference when interacting with the CQs. This chapter presents the design of the experiments, text-to-image generation modelling that is employed for generating the visual aspect of the clarification panes, human annotation process and related analyses and discussions.

Finally, Chapter 8 presents the outcome and conclusion of this research and provides some recommendations for future work.

Chapter 2

Background

As information retrieval systems continue to grow in complexity, ensuring accurate and effective communication between users and machines becomes increasingly crucial. One common method for achieving this is through the use of clarification questions (CQs) that help clarify user intent and refine search queries.

A Clarification Question (CQ) is a type of question that aims to address confusion or complexity in a particular subject matter. It is used to reduce misunderstandings and ensure clarity. Lately, the practice of asking CQs has gained attention in various domains of information retrieval (IR), such as search engines, question-answering (QA) forums, and conversational search. The purpose is to clarify the user's information needs when their query or question lacks clear intent. However, users often do not actively engage with CQs, which hinders the effectiveness of generating and asking CQs. This is because users may have different intentions even when asking the same query, making it ineffective to generate a single generic CQ for a given context.

Search engines regularly fail to understand users' complex information needs, and retrieved results for those complex needs are often not satisfactory [4]. In such cases, users often have to reformulate their queries multiple times due to the complexity of their information needs. Asking CQs can improve user satisfaction as it helps the system to better determine the intent of the user who submits the question [22, 131, 134]. The benefit of using CQs has been investigated in several fields such as dialogue systems [32, 6], community question answering (CQA) [14, 89, 62], conversational search systems [4, 122, 32, 132], and speech recognition [100].

In this chapter, we explore human- and model-generated clarification questions from various angles (Sections 2.1, 2.2) as well as the clarification datasets (Section 2.3) and evaluation techniques used to assess retrieval quality (Section 2.4). We also provide a summary of previous works and the gaps which will be addressed in this study. By understanding the state of the art in CQ research, we can improve the accuracy and effectiveness of information retrieval systems and ultimately enhance the user experience.

2.1 Human-generated Clarification Questions

Clarification questions are an essential tool in information retrieval, allowing users to refine their search queries and obtain more accurate results. In this section, we will explore the role of human-generated CQs in community question-answering forums and the ways in which they can improve search performance. By understanding the importance of human-generated CQs, we can better design systems that effectively bridge the gap between users and machines.

One of the first studies on the importance of CQs was conducted by Conrad and Schober [26]. They showed that CQs helped minimise misinterpretation in a household telephone survey, which led to compromising data quality. In studies conducted to understand the characteristics of CQs, Kato et al. [54] and Braslavski et al. [14] investigated human-generated CQs in Social Q&A and *Stack Exchange* sites (*Stack Exchange* is an online question-answering website containing posts about various topics, comments on posts and answers to the post and comments.). Kato et al. [54] investigated the relationship between CQs and dialogue outcomes with respect to the specificity of the posted question and the requested clarification in a social QA system. They classified the CQs into seven types of *Check*, *More Info*, *General*, *Selection*, *Confirmation*, *Experience* and *Other*, and observed that the most common CQ type is *Check*. They developed a question classifier to provide clarifications in the case of under-specified requests. Braslavski et al. [14] analysed user behaviour and types of CQs in *Stack Exchange* to explore the problem of predicting the specific subject of a CQ. They divided the questions into seven types of *More Info*, *Check*, *Reason*, *General*, *Selection*, *Experience* and *Not a CQ*. They also investigated three-word question starting patterns of common CQs. They found that CQs vary in topic and format and mainly depend on the content and individual characteristics of users. Later, Qu et al. [82] analysed the distribution, co-occurrence, and flow pattern of user intent in information-seeking conversations. They classified twelve classes of intent, including the original question, repeat the question, CQ, further details, follow-up question, information request, potential answer, positive feedback, negative feedback, greetings/gratitude, junk and others. Their research led to finding some frequently occurring intent patterns during

information seeking. In another study on CQs in *Stack Exchange* sites, Kumar et al. [62] ranked CQs using natural language inference and defined a CQ as good when the answer to the clarification led to a resolution of the under specification in the question posed in the original post from the Asker.

Among all discussed studies, the research conducted by Braslavski et al. [14], which focused on human-generated CQs, was the closest research to our work that will be presented in Chapter 3. However, our work is substantially different. We first classify CQs based on the type of answerer to focus on those that engage the Asker more. We then define a new taxonomy to investigate the usefulness of CQs from different aspects. The findings of this research help us to determine useful and non-useful CQs. This is a more detailed analysis compared to Braslavski et al. [14]. They classified CQs into several types and presented common patterns in general, while we define different types regardless of their answers, identify popular patterns of the useful CQs, and compare useful with non-useful CQs.

2.2 Clarification Models

In an effective information retrieval system, every user query would be met with a perfectly tailored set of CQs, refining the user’s intent and leading to the most accurate results. However, given the sheer volume of potential questions and the variability of user behaviour, it is often impractical to present every possible prompt. Clarification question selection models aim to address this challenge by identifying the most effective and relevant prompts for a given query. In this section, we will explore various approaches to Generating and selecting CQs. We will also consider the factors that can influence the performance of the models and the evaluation methods used to assess the effectiveness of different models. By understanding the strengths and limitations of CQ models, we can develop systems that strike the right balance between precision and efficiency.

2.2.1 Clarification Selection Models

Investigation of selection models for CQ has been initiated by Kiyota et al. [59]. They proposed a dialogue-based QA system utilising a large text knowledge base. The system was designed to navigate users to the desired answer by asking CQs using dialogue cards and description extraction of each retrieved text. When a user asked a vague question, the system asked the user a CQ. This process continued until the question was clarified. Their study showed that CQs are required for any dialogue with a search system. A few years later, Rao [87], Rao and Daumé III [88] and Zhang et al. [131] built neural network models based on the theory of expected value of perfect information (EVPI) proposed by Avriel

and Williams [8] to teach machines to ask CQs when there is uncertainty. EVPI had two components of Answer Modelling and Utility Calculator and helped them calculate which question was most likely to elicit an answer that would make the post more informative.

Zhang et al. [131] proposed the *System Ask - User Respond* paradigm for conversational search. They developed a multi-memory network architecture and trained their model on a large-scale dataset in e-commerce. The system was capable of asking CQs from users directly to understand user needs. They performed their experiments on the Amazon e-commerce scenario based on real-world user purchase datasets and found out that their approach outperformed state-of-the-art product search and recommendation baselines. Later, Aliannejadi et al. [4] worked on the task of asking CQs in open-domain information-seeking conversational systems. They reported that their neural question selection model was able to outperform the LTR baselines significantly and enhanced user satisfaction by asking questions that can address users' information needs. Ranking CQs on question-answering forums such as *Stack Exchange* was first investigated by Kumar et al. [62]. They assumed a helpful CQ should increase the probability of the correct answer to a given post. They showed that BERT representations pre-trained on the Stanford Natural Language Inference (SNLI) corpus [13] and the Multi-Genre NLI (MultiNLI) corpus [?] can achieve very high performance on the task.

Recent work on CQ selection systems has been conducted by Ou and Lin [78], Zamani et al. [127], Wu et al. [117], Kumar et al. [62], Sekulić et al. [95]. Ou and Lin [78] proposed a CQ selection system consisting of response understanding, candidate question recalling and CQ ranking. They fine-tuned an ELECTRA model to obtain a better understanding of users' responses and used an improved BM25 model to recall the candidate questions. In the ranking stage, they reconstructed the training dataset and introduced two models based on ELECTRA. They finally summed up the output probabilities of the models and chose the question with the highest probability as the CQ. Zamani et al. [127] analysed large-scale user interactions with CQs in a major web search engine and proposed a model for learning representation for CQs-based on user interaction. They successfully used the model for re-ranking automatically generated CQs for a given query. In other studies on clarification systems, Wu et al. [117] proposed the Predicting, Explaining, and Rectifying Failed Questions (PERQ) framework. They designed an interactive system that identified ambiguities in failed questions and requested minimal clarification actions from users. In addition, Sekulić et al. [95] modelled search clarification prediction as a user engagement problem. They focused on the task of predicting user engagement levels on clarification panes. They proposed a BERT-based model and showed that it outperformed traditional ML models. They also showed that information such as titles and text snippets of retrieved documents is beneficial in the task of predicting user engagement.

2.2.2 Clarification Generation Models

Insights about the importance of CQs in conversational search systems have led to models to generate clarification in conversational search to understand user intent better. In this section, we will explore the key concepts and techniques behind CQ generation models and examine their potential applications and limitations.

Generating CQs is a fairly new field of research in information retrieval and has recently attracted attention (e.g., [89, 122, 4, 72, 132]). One of the first attempts conducted by Quarteroni and Manandhar [84] was to design an interactive question-answering system capable of follow-up and asking CQs. In another study, Deits et al. [30] improved the performance of natural language communication between humans and robots by enabling a robot to engage in clarification dialogue with a human. Coden et al. [25] discussed challenges in interacting with users to ask CQs for entity identification. They investigated three types of CQ approaches: type-based, example-based, and usage-based, for automatically generating such questions. Interestingly, they concluded that no one method worked all the time, and an ensemble of methods may have the best performance. Zhang et al. [131] proposed the *System Ask - User Respond* paradigm for conversational search. They developed a multi-memory network architecture and trained their model on a large-scale dataset in e-commerce. The system was capable of asking CQs from users directly to understand user needs. They performed their experiments on the Amazon e-commerce scenario based on real-world user purchase datasets and found out that their approach outperformed state-of-the-art product search and recommendation baselines. To resolve to generate generic and bland CQs that cannot elicit useful information, Cao et al. [17] proposed a model that could produce questions with various levels of specificity. They trained a classifier that annotated a CQ with its level of specificity to the given context. They showed that a trained CQ generation model could generate questions with various levels of specificity. Rao and Daumé III [89] proposed an adversarial training approach for generating CQs. This sequence-to-sequence model updated a context with an answer to the CQ. They found that their model produced more useful and specific questions compared to previous models, including models trained using maximum likelihood objective and trained using utility reward-based reinforcement learning. This research inspired other research groups such as Zamani et al. [123], Hashemi et al. [43], Shwartz et al. [97] and Dhole [32] to focus on the design of clarification systems.

Zamani et al. [123] focused on the task of generating clarification for open-domain search and proposed various models for asking CQs, and they examined their models using human annotation. They showed that the Query Clarification Maximisation approach performed better than others. Hashemi et al. [43] proposed a multi-source attention network

and applied it to conversational search tasks for utilising user responses to CQs. They focused on conversations with only one CQ and a multi-turn setting. Their evaluation showed that their models, which implemented guided transformers, substantially outperformed state-of-the-art baselines. Shwartz et al. [97] proposed an unsupervised framework based on self-talk to generate natural language CQs and their corresponding answers. They generated multiple CQs considering (i) concatenating one of several question prefixes, curated for each task, and (ii) generating five questions for each prefix using Nucleus sampling. They noticed several shortcomings of using pre-trained learning models as knowledge providers, including (i) insufficient coverage, (ii) insufficient precision, and (iii) limited reasoning capabilities. Their empirical results demonstrated that the self-talk procedure they proposed substantially improved the performance of zero-shot language model baselines and outperformed models that obtained knowledge from external knowledge bases. In another study to resolve intent ambiguities in dialogue systems, Dhole [32] used an existing question generator and a sentence similarity model to generate discriminative questions. Dhole [32] presented a method to disambiguate queries that are ambiguous between two intents. They stated that the proposed method could take advantage of any question-generating system with no need for annotated data of CQs. They showed that the model improves the overall accuracy of classifying the user's intent.

Generating CQs in the conversational search was the focus of studies conducted by Rosset et al. [92], Zamani et al. [122], Zamani et al. [127], Majumder et al. [72] and Zhang and Zhu [132]. Rosset et al. [92] first defined an evaluation metric, *usefulness* to measure whether the suggestions provide valuable information for the next step or not. Then, they developed two suggestion systems, a BERT-based ranker and a GPT-2-based generator. Both together were trained with weak supervision signals transferring past users' search behaviours in search sessions. Zamani et al. [122] attempted to generate CQs for a search engine. They proposed supervised and reinforcement learning models to generate CQs learned from weak supervision data. They also investigated methods for generating candidate answers for each CQ. Human evaluation of the generated CQs and candidate answers demonstrated the effectiveness of their proposed solutions. Zamani et al. [127] again analysed large-scale user interactions with CQs in the *Bing* search engine in terms of the user engagements received by CQs based on different properties of search queries. They also proposed a model for learning representation for CQs, considering user interaction data as implicit feedback. Their evaluation of both the click data and human-labelled data demonstrated the high quality of the proposed method. Zhang and Zhu [132] believed that due to the possibility of various user intents, generating one generic CQ per context cannot be successful. Therefore, they proposed the task of diverse CQ generation and addressed the challenge of specificity. To do so, they introduced a model named Keyword

Prediction and Conditioning Network (KPCNet), which predicted keywords focusing on the specific aspects of the question. Then, they developed keyword selection methods to produce multiple keyword groups for generation diversity. Their analyses showed that the model improved the quality of the CQ generation. To improve generating CQs, Majumder et al. [72] followed another strategy. They first identified what is the missing proposal by taking a difference between the global (i.e., collecting previous similar contexts) and the local (i.e., the available context for a product) views and then trained a model to identify what is useful. Finally, they generated a question about it using a fine-tuned BART model.

We showed that while previous works investigated CQs in CQA sites and search engines from various aspects, including characteristics of CQs, proposing selecting and generating CQs, they are still in the early stages with limited success in engaging users due to a lack of understanding of what makes a CQ engaging from a user’s perspective. In fact, to successfully make users interact with a CQ, a useful and engaging CQ needs to be asked of the user when the user information need is not clear. We will extensively investigate the relationship between user engagement with the clarification characteristics in this research.

Moreover, despite the growing interest in search clarification and its exploration from various perspectives, there is a noticeable research gap regarding user preferences and perceptions of different modalities in search clarification. The literature review also reveals that studies on multi-modality in information retrieval (IR) have mainly overlooked the field of search clarification. For example, Yang et al. [119] introduced an online video recommendation system incorporating multi-modal fusion and relevance feedback. Zha et al. [129] proposed Visual Query Suggestion (VQS) for image search, Altinkaya and Smeulders [7] developed a model for stuttering detection, Srinivasan and Setlur [98] explored utterance recommendations for visual analysis, Pantazopoulos et al. [80] integrated computer vision and conversational systems for socially assistive robots, and Ferreira et al. [34] presented TWIZ, a multi-modal conversational task wizard. However, none of these works specifically addressed the challenges and techniques related to multi-modal clarification questions in the context of search systems. Hence, there is a significant research gap in this area, highlighting the need for further exploration and development.

2.3 Clarification Datasets

Clarification question datasets play a crucial role in the development and evaluation of CQ generation models. These datasets consist of a set of questions that aim to clarify ambiguous or unclear statements and are used to train and test machine learning models to generate similar questions. As the demand for more accurate and effective natural language processing systems continues to grow, the availability of high-quality CQ datasets

has become increasingly important. This section will explore the key characteristics and properties of CQ datasets. For research datasets, we can divide resources into two main categories: CQA clarification datasets (e.g., collected human-generated CQs on *Stack Exchange*) and search clarification datasets (e.g., collected model-generated CQs on *Bing*). In this section, we discuss the available clarification datasets.

2.3.1 CQA Clarification Datasets

Among the clarification datasets in CQA, Rao [87], Braslavski et al. [14] and Rao and Daumé III [88] created clarification question dataset using *Stack Exchange* platform. Rao [87] and Rao and Daumé III [88] extracted a total of 37,000 and 77,000 triples (post: the initial unedited post, question: the comment containing a question, answer: an edit made to the post after the question or the author’s response to the question in the comments section), respectively, from *Stack Exchange* three domains of *askubuntu*, *unix* and *superuser*. Their dataset was later used by Kumar et al. [62] for ranking CQs via natural language inference. In another study, Braslavski et al. [14] used two *Stack Exchange* sites of *Home Improvements* (DIY) and *Arqade* (GAMES) to build their dataset of about 83,000 questions. Another large-scale dataset, called ClarQ, extracted from 173 different *Stack Exchange* domains, was presented by [61]. They proposed a bootstrapping framework to employ a neural network for classifying CQs. One of the latest clarification datasets was introduced by Min et al. [75]. They constructed a dataset from an open-domain QA benchmark containing diverse types of ambiguity, which are not normally visible from the prompt question alone.

Some other researchers such as Rao and Daumé III [89], Zhang and Zhu [132], and Majumder et al. [72] investigated generating CQ models on the Amazon Review dataset [74, 73]. In this dataset, context is a product description, including the product title, and the question is a CQ asked about the product, and the answer is the seller’s or other users’ reply to the question.

Although CQA datasets provided valuable opportunities to investigate the CQs from various aspects, they are of limited use in asking and generating CQs in search clarification. These datasets record human interactions with human-generated CQs, while in a conversational search system, a human interacts with a machine. The nature of the query and the information needed is also different in search clarification compared to a community forum.

2.3.2 Search Clarification Datasets

Several search clarification datasets have been created over the last few years by Xu et al. [118], Aliannejadi et al. [4], Penha et al. [81], Aliannejadi et al. [5], Zamani et al. [125] and Aliannejadi et al. [6]. Xu et al. [118] created a clarification dataset, CLAQUA, of 40,000

open-domain examples to enable systems to ask CQs in open-domain question answering. This dataset supported three tasks: giving a question, checking whether clarification is needed; if yes, generating a CQ, then predicting answers based on user feedback. Aliannejadi et al. [4] collected a CQ dataset through crowd-sourcing named *Qulac*. This dataset was built on top of the TREC Web Track 2009-2012 data and contained over 10,000 question-answer pairs for 198 TREC topics with 762 facets. Inspired by *Qulac*, Aliannejadi et al. [5, 6] crowd-sourced new datasets to study CQs that were suitable for conversational settings and in open domain dialogues focusing on single and multi-turn conversations. Penha et al. [81] created a dataset that focused on the interaction between an agent and a user, including CQs. The researchers presented a conceptual model and provided baseline results for conversation response ranking and user intent prediction tasks.

The *MIMICS* dataset, which is used widely in this thesis, was introduced by Zamani et al. [125]. It is the largest search clarification dataset extracted from *Bing* search engine. Each clarification was generated by a *Bing* production algorithm and contained a CQ and up to five candidate answers. Compared to other datasets, *MIMICS* contains realistic queries, is comprehensive and covers a wide range of clarification types, including user interaction signals. *MIMICS* also contains search engine results pages (SERPs) of up to ten retrieved documents, including a title, URL, and snippet for each query. The full *MIMICS* data collection consists of three datasets:

1. *MIMICS-Click*, which includes over 400,000 unique queries, their associated clarification panes, and the corresponding aggregated user interaction signals. Each data point in *MIMICS-Click* includes a query-clarification pair, its impression level (low, medium, or high), its engagement level (between 0 and 10, and the engagement level of 0 means there was no click on the CQ), and the conditional click probability for each individual candidate answer.
2. *MIMICS-clickExplore*, which includes over 60,000 unique queries, their multiple clarification panes, and user interaction signal similar to the *MIMICS-click* dataset.
3. *MIMICS-Manual*, which includes over 2,000 unique search queries with their multiple CQs, quality labels were manually labelled by at least three trained annotators (A quality label of 2 (Good), 1 (Fair), or 0 (Bad) has been assigned to each CQ), and landing SERPs.

Available CQ datasets are either created based on the user interaction signals, such as click-through rate or collected through manual annotation. We show in Chapter 4 that while the *MIMICS* dataset, as the largest search clarification dataset, enhanced our understanding of user interaction with CQs and inspired several other studies (e.g., [95, 112,

71, 44]), it is not yet sufficient for training and evaluating search clarification methods and models. It does not support online-offline evaluations in search clarification to understand the relationship between user engagement and clarification characteristics. We aim to resolve this shortcoming in search clarification by introducing the *MIMICS-Duo* dataset in Chapter 5. *MIMICS-Duo* is a balanced dataset that benefits from user interaction signals while providing insightful information about the characteristics of CQs, enabling extensive online-offline evaluations in search clarification.

2.4 Online and Offline Evaluation in Information Retrieval

To understand what makes a CQ engaging from a user’s point of view, the relationships between various characteristics of the CQs, labelled by human judgement, and user click-through rate, known as a signal for user engagement, need to be investigated. Such studies are known as online evaluations, and we review the previous works on this topic now.

Online and offline evaluation are two widely used methods for evaluating the effectiveness of information retrieval systems. While offline evaluations are performed on pre-collected datasets, online evaluations involve testing the system in real time using actual users. Both approaches have their advantages and disadvantages, and the choice of which method to use depends on various factors, such as the type of system being evaluated and the available resources. In this section, we will discuss the key concepts and techniques behind online and offline evaluations in information retrieval and examine their strengths and weaknesses.

As mentioned above, there are two approaches in general to evaluate retrieval quality: (1) manual judgements of the relevance of documents to queries provided by trained experts (offline evaluation) [24] and (2) user behaviour observations (implicit feedback) when presenting the search results (online evaluation) [19].

The effectiveness of using expert judgements, which is also known as offline evaluation in quality retrieval analysis, has been proved before [111]. Offline evaluations are often used before deploying new ranking policies, which help to run A/B testing (i.e., a randomised experiment that usually involves two variants (A and B), shown to users, and statistical analysis is used to determine which variation performs better) [60] more safely and intelligently [23, 64]. However, such an evaluation has two limitations. First, expert judgements may not be capable of reflecting the actual relevance and cannot reliably estimate the user’s actual information need simply based on the query issued and inaccurately reflect user utility [18, 2]. This comes from the fact that different users may

issue the same textual query while having different information needs or intents [105]. Second, the cost associated with conducting offline evaluations, such as hiring experts or setting up infrastructure, is typically substantial. Additionally, the process of offline evaluations usually takes a considerable amount of time to complete, which can range from days to weeks or even longer. These factors make the benefits of offline evaluations limited for many organisations or projects, as the expenses and time required may be too burdensome. Consequently, alternative evaluation methods that are more cost-effective and faster, such as online evaluations, are often preferred. These online metrics are based on observable user behaviour [55, 19] and include CTR (Click Through Rate) and the ranks of clicked documents [53] as well as their extensions (e.g., UCTR (i.e., a binary value representing click) [23], PLC (i.e., number of clicks divided by the position of the lowest click) [39]), dwell time including query dwell time, time to first click, the average of click dwell time [121, 48], query reformulations, response times, how the session was terminated (e.g., by closing the browser window or by typing a new Internet address) [35], mouse movement and per-topic reading time [58].

Retrieval performance metrics based on implicit feedback directly from the users, known as online evaluations, can be grouped into two classes of absolute metrics and pairwise preferences [69]. Contrary to absolute metrics that provide an overall assessment of the retrieval performance based on predefined criteria, pairwise preference methods such as interleaving assume that the better of two (or more) options can be identified based on user behaviour. For example, clicked results are preferred over results previously skipped in the ranking [51]. Despite the enormous value of click-through data, it is inherently biased and very noisy [113]. There are multiple sources of bias, including position bias [52], presentation bias (e.g., the position of results in the ranking) [101], and trust bias [79]. Such noisy data may lead to biased training data that negatively affects the downstream applications [49]. There are also some other factors, such as educational level, intelligence, and familiarity with the IR system, that impact the decision of user satisfaction and the click-through data [3, 46], making the data difficult to interpret. This is in agreement with observations by Zheng et al. [133] that click-through data and relevance do not always correlate, and CTR should be used with precaution.

Offline evaluations provide a low-cost methodology to predict the performance of models and insight into whether it is worth testing on the more expensive online evaluation. However, substantial discrepancies between the offline and online evaluations have been identified in the literature. Cremonesi et al. [27], Garcin et al. [38], Ekstrand et al. [33], Garcin et al. [38], Said and Bellogín [94] identified several inconsistencies when investigating recommendation methods using online and offline evaluations. Yi et al. [120] investigated the performance of predictive models for search advertising using online and

offline evaluation metrics and showed that some offline metrics like AUC (the Area Under the Receiver Operating Characteristic Curve) and RIG (Relative Information Gain) could be misleading and result in a discrepancy in online and offline metrics. Such discrepancy was also observed and stated by Beel et al. [11] and Beel and Langer [10]. In another study, Garcin et al. [38] investigated news recommenders and showed that in an offline setting, recommending popular stories is a winning strategy, but in an online setting, it was the poorest.

Online evaluations can also be misleading. Zheng et al. [133] and later Garcin et al. [38] showed that CTR, an adopted and widely accepted metric in online evaluations, overestimates the impact of popular items. In fact, recommending items with higher CTR does not necessarily imply higher relevance of two items, and factors like item popularity, item serendipity or the placement/order of recommendations may also influence a user's click behaviour.

Chen et al. [20] meta-evaluate a series of existing online and offline metrics to study how well they predict actual search user satisfaction in different search scenarios. They showed both types of evaluation noticeably correlate with user satisfaction, but they reflect satisfaction from different perspectives and for different search tasks. They observed a strong correlation between top-weighted offline metrics and user satisfaction in homogeneous search (i.e. ten blue links), whereas online metrics outperform offline metrics when vertical results are federated. They also understood that incorporating mouse hover information into existing online evaluation metrics better aligns with search user satisfaction than click-based online metrics. Liu and Yu [66] believed users often seek different goals at different search moments, which may evaluate system performances differently. Therefore, it would be difficult to achieve real-time adaptive search evaluation and recommendation. They meta-evaluated a series of online and offline evaluation metrics under varying states based on a user study dataset. Their results showed that the performance of query-related and online features had large variations across different task states. However, offline evaluation metrics, in general, had stronger correlations with user satisfaction.

In another study, Rossetti et al. [93] showed that with the same set of users, the ranking of algorithms based on offline accuracy measurements contradicts the results from the online study. Later, a comparison of online and offline assessments for Query Auto Completion was carried out by Bampoulidis et al. [9], and it showed a large potential for significant bias if the raw data used in an online experiment is re-used for offline evaluations. It is worth noting that a lack of correlation between offline and online evaluations in voice shopping traffic and Web image search was also reported by Zhang et al. [130] and Ingber et al. [47].

While prior works have offered insight into how well online and offline evaluations correlate in retrieval quality, there is no extensive study on this controversial topic in

search clarification. The only available study was conducted by Zamani et al. [125], who examined the MIMICS dataset and investigated correlations between online and offline evaluations using a single offline label. They concluded that no correlation was observed between the two evaluation methods. The focus of our study is to investigate the relationship between online and offline evaluations in terms of ranking multiple clarification panes and identifying the most engaging clarification pane for a given query (the main of any clarification selection model). Furthermore, we group the query-clarification pairs based on the query length and impression level for a more detailed study. Furthermore, we investigate the impact of offline labels on the improvement of the performance of LTR and Large Language models in identifying the most engaging clarification questions. The study described in Chapter 6 aims to extensively investigate user engagement with model-generated CQs in a search engine through online and offline evaluation approaches to fill the current gap.

Chapter 3

Useful Clarification Questions in Community Question Answering Forums

Community Question-Answering (CQA) forums record asynchronous exchanges between people seeking to solve complex information needs, including those that a web search may have failed to answer [68, 96]. This is because human-driven online QA services allow users to seek different forms of information, ranging from factual information to personal opinions or advice through human-to-human interactions [21]. However, the questions that are asked on information-seeking forums such as *Stack Exchange*, *Quora*, or *Yahoo! Answers* are sometimes complex or ambiguous for other users. In such cases, other users usually ask several clarification questions (CQs) to clarify an Asker's information need. Our initial investigation of CQA forums shows that a high percentage of CQs are left unanswered, suggesting that not all clarifications are necessarily useful. As the first step toward understanding what makes a CQ engaging from a user's point of view, this chapter analyses human-generated CQs on the *Stack Exchange* platform, to provide insights into how they are used to disambiguate and provide a better understanding of information needs.

3.1 Introduction

In this chapter, we present our own approach to classifying CQs with respect to the type of *Answerer* (i.e., Asker: who posted the initial question or a Responder: other users) by extracting a set of CQs from posts (the initial questions posted by the Askers). Figure 3.1,

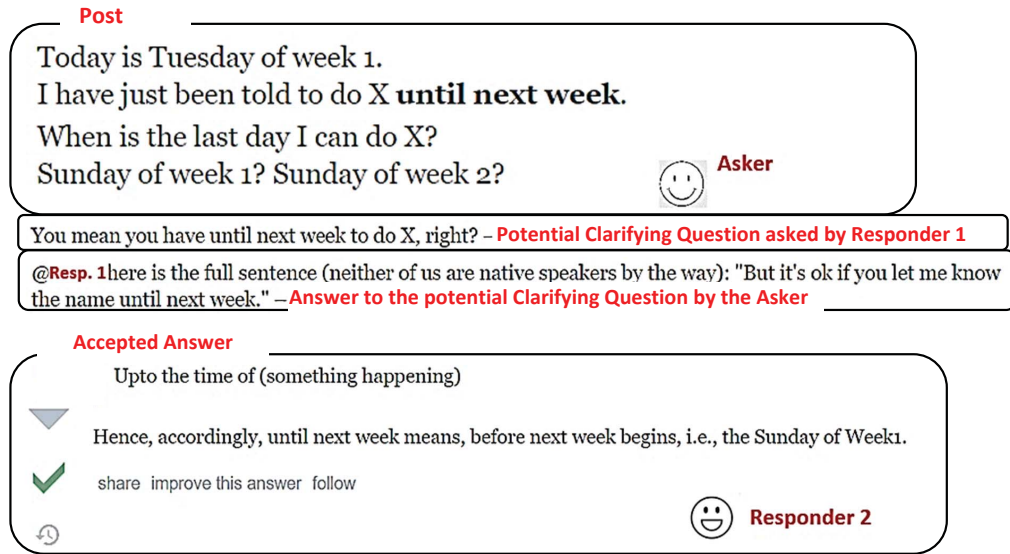


Figure 3.1: A question posted on *Stack Exchange*. (*Asker*: who posted the initial question, *Responder*: another user provided the answer, *Accepted Answer*: an answer among all provided answers by other users that were chosen by the Askers.)

shows an initial question posted by an *Asker*, followed by a CQ from a *Responder*. The interaction led to an *accepted answer* (i.e., an answer among all provided answers by other users that are chosen by the Asker). We use this terminology throughout the chapter. We aim to answer the following research questions:

- What clarification questions are more useful (in terms of helping the Asker to get a correct answer)?
- What are the characteristics of useful clarification questions?

We first investigate the *usefulness* of clarifications based on manual annotation. We propose new definitions for *useful* and *non-useful* clarification questions. We investigate the CQs in terms of whether they add any information to the post and the accepted answer, which is the answer chosen by the Asker. After identifying which CQs are more useful, we investigate the characteristics of these questions in terms of their types and patterns. Non-useful CQs are also identified, and their patterns are compared with useful clarifications. Figure 3.2 shows the steps taken in this chapter to explore the CQs in CQA forums.

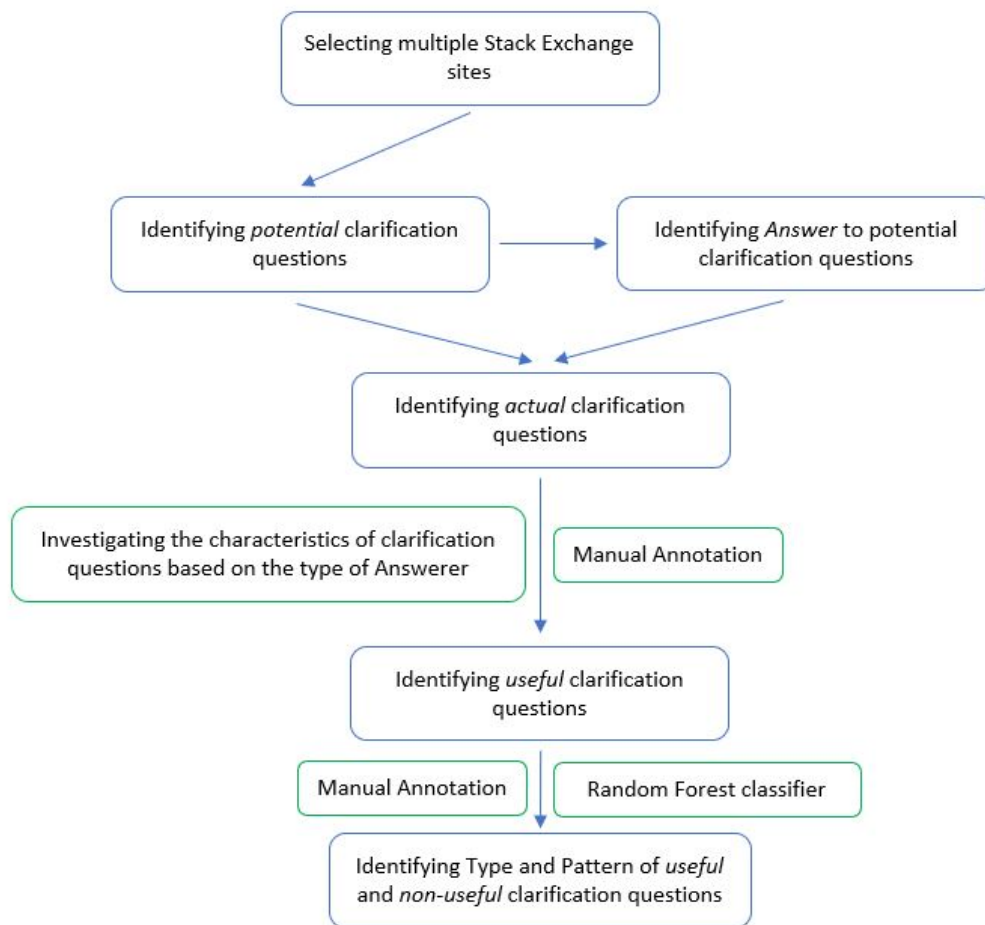


Figure 3.2: The flow of study of the study in this chapter.

3.2 Methodology

To investigate our research questions and better understand what CQs are useful in terms of helping the Asker of a post obtain an accepted answer, we investigate publicly available data from *Stack Exchange*¹ covering a period from July 2009 to September 2019. We investigate three sites that had the highest number of posts² in each of three different *Stack Exchange* categories: Business³ (which holds three sites), Culture/Recreation⁴ (46 sites) and Life/Arts⁵ (26 sites). Table 3.1 reports the details of the three chosen sites: Quantitative Finance (*QF*), English Language and Usage (*EL*), and Science Fiction and Fantasy (*SF*).

The data is first processed by identifying posts, potential CQs, and their answers within

¹<https://archive.org/details/stackexchange>

²As stated by Shah et al. [96] and Liu et al. [67], such popularity may indicate that users cannot satisfy their information needs using web search engines.

³<https://stackexchange.com/sites#business>

⁴<https://stackexchange.com/sites#cultureandrecreation>

⁵<https://stackexchange.com/sites#lifeandarts>

Table 3.1: The analysed sites of *Stack Exchange*.

| Category | Site | # of Posts |
|--------------------|---|------------|
| Business | <i>Quantitative Finance (QF)</i> | 13,187 |
| Culture/Recreation | <i>English Language and Usage (EL)</i> | 107,266 |
| Life/Arts | <i>Science Fiction and Fantasy (SF)</i> | 55,959 |

the data. Second, questions are classified with respect to the type of Answerer. Third, the relationship between the type of Answerer and the post is investigated.

We showed in Chapter 2 that information-seeking community forums have been studied widely over the last two decades [65, 63]. However, human-to-human interaction in such platforms in terms of interaction with CQs is a new area of interest. This is particularly helpful as it can enhance our knowledge of developing information-seeking systems with higher performance. We showed that there is some research on generating and asking CQs and some studies on CQs in QA forums. While generating and asking a CQ is important, getting an answer for the CQ can even be more important [103]. This means a CQ helps the search system only when the user provides an informative answer. To improve the performance of information-seeking systems by generating CQs that engage the users more actively (useful CQs), we attempt to find characteristics of such clarifications in terms of types and patterns in this study.

The aim of this chapter is to understand why some CQs lead to more engagement from Askers. This is because we discovered that many CQs are left unanswered, and such CQs are generally not useful in helping to better understand an Asker’s information need. We can not understand this without investigating the answer to the CQs at the same time. Therefore, we classify CQs with respect to the type of *Answerer*. This is, in contrast, to Braslavski et al. [14] who only studied the CQ itself.

3.2.1 Identifying Potential Clarifications

To identify *potential clarification questions*, we collect comments from within posts that contain at least one sentence ending with a question mark, regardless of the question content. Sentences containing question marks that appear in the form of a quotation are ignored (e.g. *Note Swiss German might say: “Wie isch ire Name?”*). A question is also disregarded if it is part of a hyperlink (e.g. a link to a web page is provided in the form of a question). We exclude any questions submitted by the Asker, assuming that the person who submitted a post would not ask for clarifications. In addition, if the question starts with “@username”, it should be Asker’s username. This is to ensure that the question is directed to the Asker and is not part of a conversation between other users.

In order to identify the *answer* to a CQ, the following criteria have to be met:

- The comment starts with “@username”, which is the name of the user who asked the CQ.
- The comment is submitted after the CQ, based on timestamps.
- The user who asks the CQ did not comment between the CQ and the provided answer to that CQ (this maximises the likelihood of the comment being a response to the CQ).

3.2.2 Annotation and Data Sampling

To investigate the characteristics of CQs based on the type of Answerer (i.e., the Asker, a Responder, or unanswered questions), we conduct manual annotation to answer the following questions:

- Is a *potential* CQ as defined above an *actual* CQ?
- Does the CQ have an informative answer?
- Does answering the CQ add any value to the post overall? (The focus of this attribute is the CQ itself, regardless it has an answer or not.)
- Does answering the CQ add any information to the accepted answer of the post?

Three annotators, one who also acted as *coordinator*, carried out the labelling. The coordinator met the other two annotators to explain the labelling strategy and the annotation procedure (the *guidelines*) to provide a common ground for everyone. The coordinator collected all annotations, aggregated them, and identified any disagreements. Next, the coordinator met the annotators to get their feedback and discuss any challenges that they encountered. The annotators discussed the labels with disagreement. As a result of the discussions, the guidelines were sometimes refined, CQs with disagreements were re-labelled, and those new labels were aggregated once again. 363 posts out of a total of 557 sampled posts were initially agreed. When the guideline was discussed and amended, in a few cases (17 posts) where there was still disagreement, majority voting was used to obtain the final label. This means the coordinator recorded the related label if the agreement score was greater than or equal to 66.67%. The developed annotation guidelines and the four steps are summarised below:

- ***Actual clarification question:*** To determine the usefulness of a CQ, it is a prerequisite to ensure that a potential CQ is an actual CQ. A potential question is considered an actual question (hereafter, simply called a CQ) if it is on the topic of the post, if it appears to be clear, and does not contain: (i) sarcastic/humorous questions and

rhetorical questions [14] and (ii) comments which provide a solution or give a hint for the post in the form of a question (e.g. “*Why don’t you just try a backtest ...?*”). This type of question does not generally look for an answer. In contrast to Kato et al. [54], we do not consider such a question as a CQ.

To identify a CQ accurately, an investigation of potential CQs, any accompanying sentences, and the post submitted by Asker is required. For those potential CQs that are actual CQs, the following three attributes are assessed.

- **Informative answer to the CQ:** At this stage, we classify an answer to a CQ as informative or non-informative. This part of the study is essential because a CQ needs to have an informative answer to help the Askers with their posts. An answer to a CQ is informative when it responds to the CQ or a portion of it (Figure 3.3).

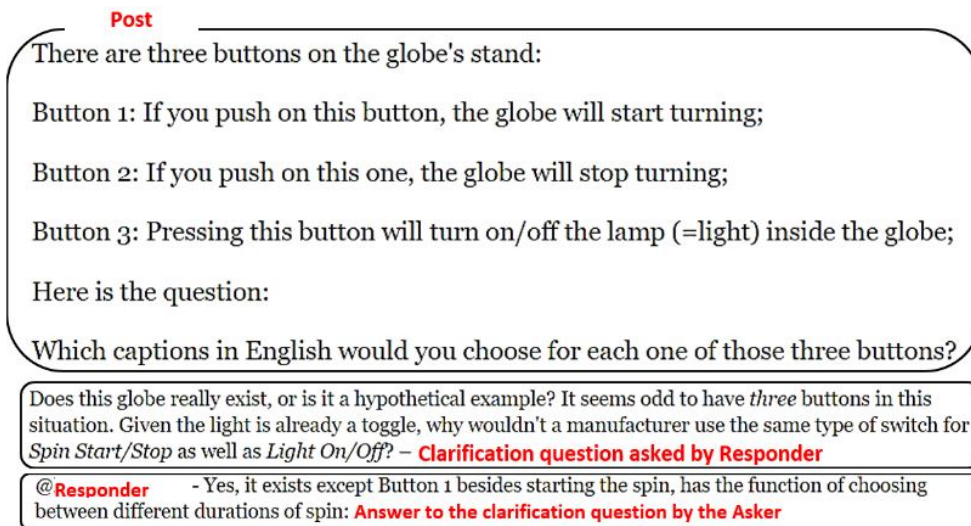


Figure 3.3: Clarification receiving an informative answer.

There are some conditions when the answer to the CQ is not informative: (i) the CQ has accompanying sentences, and the Asker responds to these sentences rather than the CQ itself (Figure 3.4) and (ii) the CQ receives a relevant but incorrect answer when the Asker misunderstands the CQ (Figure 3.5).

- **Valuable for the post:** A CQ can be relevant to a post, but it does not necessarily add value to it. Here, we consider a CQ as valuable for the post if it attempts to resolve ambiguity or to eliminate any incompleteness in the post (Figure 3.6). In contrast, Figure 3.7 indicates a CQ asking something that is not about the post and therefore does not add value to the post. To evaluate this attribute, the CQ and the post need to be considered together.
- **Valuable for the accepted answer:** We consider a CQ valuable for the accepted answer if it improves an accepted answer or if answering the CQ is necessary to

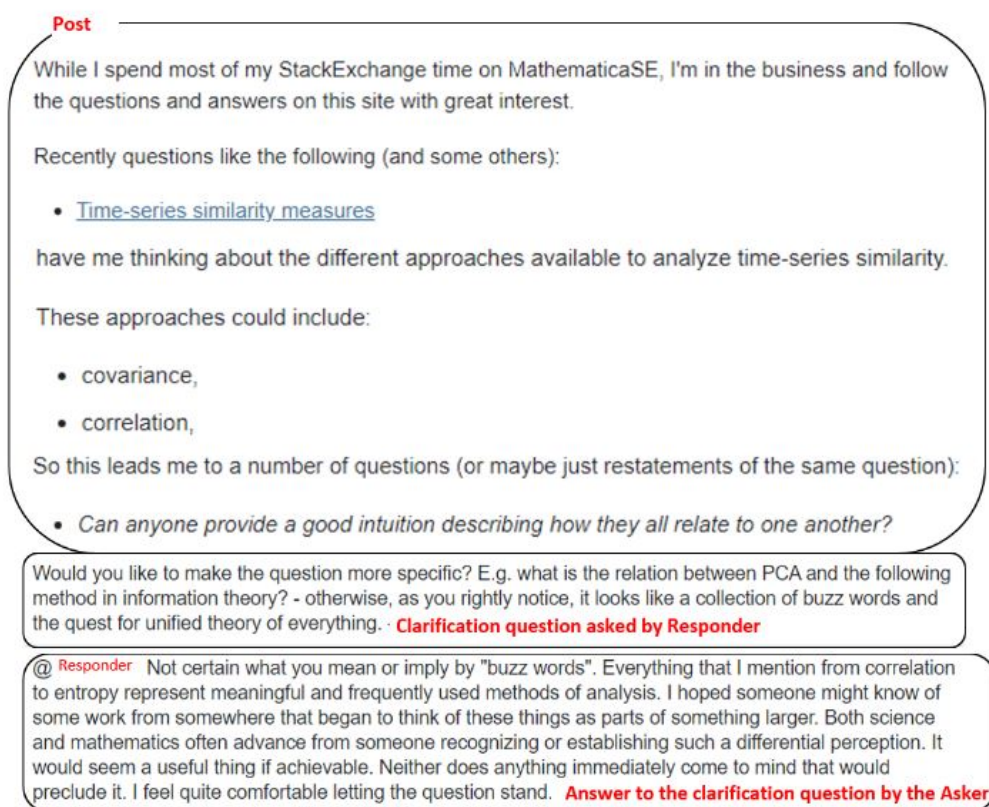


Figure 3.4: Clarification receiving an uninformative answer.

produce an accepted answer for the post. To be considered valuable for the accepted answer, the CQ needs to meet the following criteria: (i) the CQ has an informative answer, (ii) the CQ is labelled as valuable for the post. This is because the category of *valuable for the accepted answer* is a subclass of *valuable for the post*, and (iii) the post has an accepted answer.

The CQ and its answer, the post and the accepted answer need to be considered together to label this attribute. Since a post usually has some introductory parts or details, in order for a CQ to be valuable for an accepted answer, the question needs to address the main focus of the post, in contrast to being valuable for the post, which can target any aspect of the post. Moreover, the answer to the CQs needs to improve the accepted answer. Figure 3.8 shows an example of a CQ that is valuable for both the post and the accepted answer.

When the annotation guidelines are finalised, all CQs of the three sites are randomly sampled to assess the labels with respect to the type of Answerer. To ensure that the samples are representative of their constituent site, each sample size is estimated based on a finite population with a confidence level of 90% and an error margin of 10%. To give every potential CQ an equal chance of being selected, a simple random sampling approach with a

Post

Why is "hopefully" treated so mercilessly?

Asked 10 years ago Active 2 years ago Viewed 2k times

▲ Is the word "hopefully" unjustly treated? We don't like the sentence:

7 "Hopefully, my ship is just over the horizon and due in real soon now."

▼ But we don't mind saying:

📌 "Happily, the tree fell on that eyesore shed."

1 "Sadly, the swallows have not returned."

🕒 Why is "hopefully" so unwelcome at the adverb party?

ObJoke: From The New Yorker a few years ago, to illustrate proper use of "hopefully:"

Dad (shaving): Ouch! Damn!
 Son: What's wrong?
 Dad: I cut my chin!
 Son (hopefully): Off?

Share Improve this question Follow

edited Dec 8 '13 at 14:19 asked May 5 '11 at 23:35

tchrist ♦ 124k 47 334 525 **Asker**

Where do you see that hopefully is "despised" and, so, less used? (if I got what you mean)... – **Responder 1**
 '11 at 23:43 ✓ **Clarification question asked by Responder 1**

@. **Res 1** -- yes, despised. And rejected. – **Asker** May 5 **An incorrect answer to the clarification question provided by the Asker**

I think Alenanno is asking *where* it is despised. As in, why do you think it is? – **Responder 2**

Responder 2 is trying to clarify the Responder's 1 intent

Figure 3.5: A useless answer due to a misunderstanding.

random number generator is used. In total, 557 potential CQs are sampled across the three sites. Table 3.2 indicates the samples of potential CQs (taken from each domain), which are used to assess four different taxonomies based on the type of Answerer.

Types and Patterns of Clarification Questions

To investigate what the characteristics of useful CQs are, we now analyse the type (based on user intent) and the pattern (represented as a trigram of words) of CQs. We compare the most common patterns of useful CQs with the patterns of non-useful CQs. This analysis provides insight for developing models for generating CQs in information-seeking systems.

To recognise the patterns and types of CQs, we first perform a manual annotation on 376 comments, which are answered by the Askers in the *QF* site and contain at least one CQ. The annotation follows the same procedure described earlier. Two annotators perform the classification to reduce the impact of personal judgement; disagreements are discussed between the annotators, and a coordinator makes the final decision. and we iteratively update the annotation guidelines based on the discussions between the annotators and the

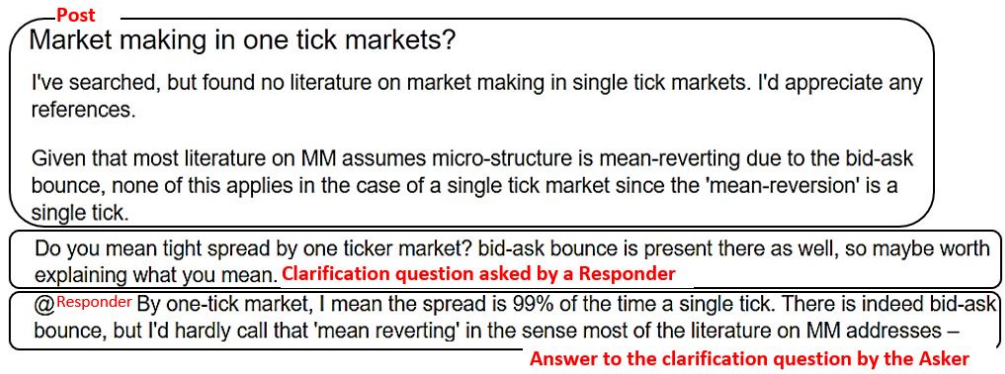


Figure 3.6: A clarification asked to eliminate ambiguity.

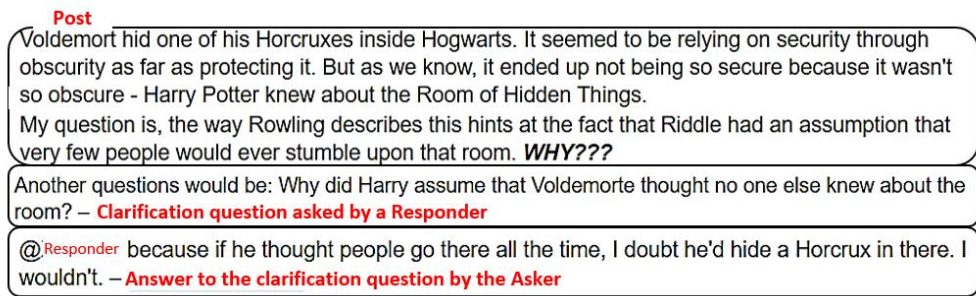


Figure 3.7: A non-valuable clarification question.

coordinator.

As the large size of the complete dataset makes complete manual labelling infeasible, the remaining clarifications (answered by Responders or left unanswered) in the *QF* site, and all clarifications in the two other sites, are labelled automatically using a Random Forest classifier. We use TF-IDF weighted bag-of-word features, which include n -grams ($n \in \{1, 2, 3, 4, 5\}$). To provide a more meaningful representation, some patterns such as *could you please give*, *can you please give*, *could you give*, and *can you give* are compiled. The classifier is trained on 80% of the aforementioned annotated dataset collected from the *QF* site. Each data point includes a CQ and its type and pattern. It is worth noting that there are some questions that do not adhere to any specific patterns, such as “*Perhaps question an example of one of your long sentences?*”. In order to detect such CQs, these pattern-less clarifications are included in the training data. However, these CQs are excluded from the subsequent analysis. To foster this research, we open-source our implementation and annotated data.⁶

To measure the accuracy of the classifier, we use the remaining 20% of the annotated data as test data. Our model achieves an accuracy of 78.21% on the held-out test set. To further verify the quality of the automatic annotation process, 10% of the results are cross-

⁶https://github.com/Leila-Ta/Clarification_CQA/

Post

Who was more likely to die, a Death Eater or Order of the Phoenix member?

Asked 9 years, 3 months ago · Active 2 years, 10 months ago · Viewed 4k times

My [thesis](#) is that Peter Pettigrew chose the Death Eaters since it was safer - specifically, you had a lot more chances of **not getting killed** compared to staying on James Potter's side.

Based on canon, is that supported?

Who died more, Death Eaters or Order of the Phoenix?

Share · Improve this question · Follow

asked Jan 21 '12 at 17:07
Asker

Is death the only outcome you care about or would being sentenced to Azkaban also be a concern? I don't recall if Voldemort had control over the dementors in the first war or not so this is just me wondering out loud. – **Clarification question asked by a Responder 1**

@ **Res 1** death only. – **Answer to the clarification question by the Asker**

(New! [list of deaths](#), from HP wiki. Analysis at the end.) **Accepted Answer**

30 I'm going to say that it was generally safer to be a Death Eater because the good guys will at least try to capture them, while the Death Eaters won't hesitate to kill. Consider the three groups fighting:

✓ First, the Order of the Phoenix. They were probably the most powerful of the groups, but also the most principled. I'm pretty certain they didn't kill the death eaters they fought unless absolutely necessary.

...

Share · Improve this answer · Follow

edited Jan 22 '12 at 0:07

answered Jan 21 '12 at 18:42
Responder 2

Figure 3.8: A clarification that enhances the accepted answer.

Table 3.2: Sample Size (Number of investigated Potential clarification question).

| Type of Answerer | Quantitative Finance | English Language and Usage | Science Fiction and Fantasy | Total |
|------------------|----------------------|----------------------------|-----------------------------|-------|
| Asker | 58 | 67 | 67 | 192 |
| Responder | 30 | 66 | 66 | 162 |
| Unanswered | 67 | 68 | 68 | 203 |

checked by the annotators. Annotators are also asked to edit any pattern that is labelled mistakenly and to add these to the list of detected patterns and types.

3.3 Results and Analysis

In this section, we present the results and analyses of user engagement with CQs, the usefulness of CQs and clarification types and patterns.

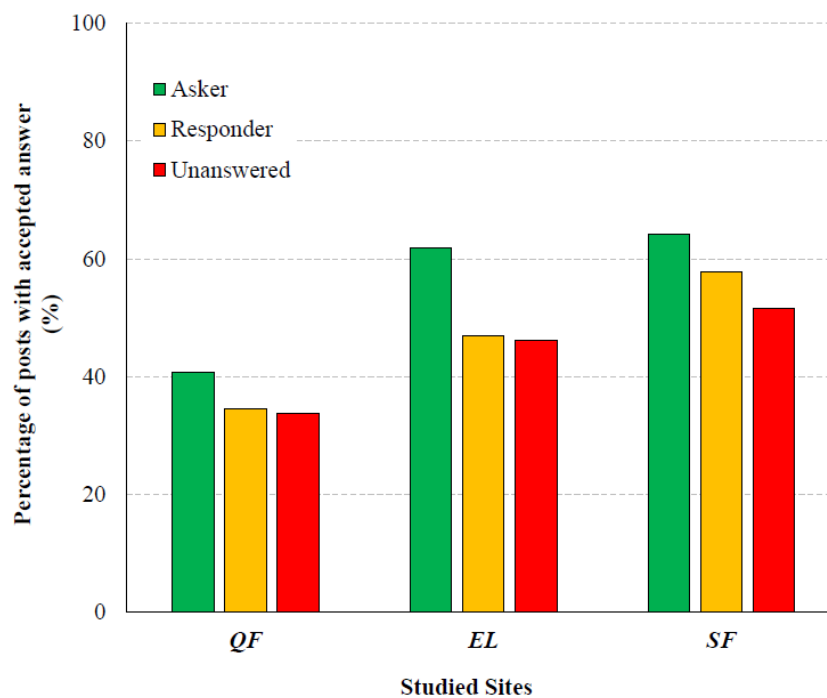
Table 3.3: Who answers clarification questions.

| Type of Answerer | Quantitative Finance | English Language and Usage | Science Fiction and Fantasy |
|-------------------|----------------------|----------------------------|-----------------------------|
| Asker | 376 (8.7%) | 4065 (6.46%) | 3144 (9.04%) |
| Responder | 42 (0.97%) | 2027 (3.22%) | 2100 (6.04%) |
| Asker & Responder | 3 (0.07%) | 100 (0.16%) | 167 (0.48%) |
| Unanswered | 3905 (90.39%) | 56971 (90.48%) | 29707 (85.4%) |

3.3.1 User Engagement and Clarification

To understand the utility of CQs, we analyse questions with respect to the Answerer. Table 3.3 shows the majority of clarifications (around 90%) are unanswered. Of those that are answered, the Asker is the most likely to reply. The high percentage of clarifications with no answer confirms the importance of investigating the properties of the CQs that engage Askers. The results shown in Table 3.3 are all statistically significant for these three sites and also for the types of the Answerer (two-tailed z-test, $p < 0.00001$).

Considering those posts where CQs are answered by only one type of Answerer, we find that when CQs are answered by the Asker, there is a higher chance of the post gaining an accepted answer (Figure 3.9). The percentage is significantly higher than the other two Answerer types in *EL* and *SF* sites (two-tailed z-test, $p = 0.05614$ for *QF*; $p < 0.00001$ for *EL*; $p < 0.00001$ for *SF*).

**Figure 3.9:** Percentage of posts with an accepted answer, grouped by Answerer.

Next, we examine the relationship between the percentage of CQs answered by the

Asker per post with the number of posts with an accepted answer (Figure 3.10). We divide the percentage of the CQs answered by the Asker into four bins. As can be seen, there are similar trends across all three sites: the greater the fraction of CQs answered by the Asker, the more posts obtain an accepted answer. For posts with an accepted answer, we examine the relationship between the number of asked CQs and the rate at which Asker answers those questions. Figures 3.11, 3.12 and 3.13 show similar trends across the three sites. We can see the maximum asked potential CQs are different in investigated sites (a maximum of 8, 20 and 12 potential CQs are asked per post in the *QF*, *EL* and *SF* sites, respectively).

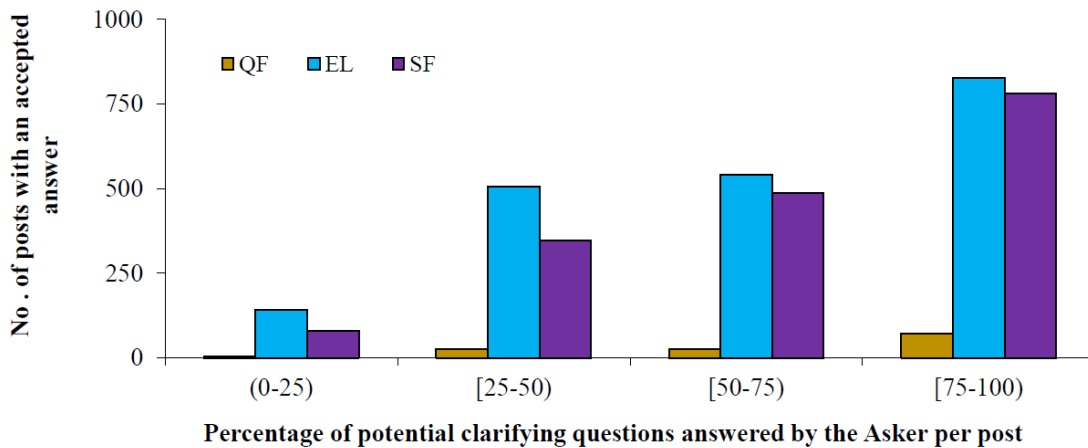


Figure 3.10: Number of posts with an accepted answer grouped by the % of questions answered by the Asker.

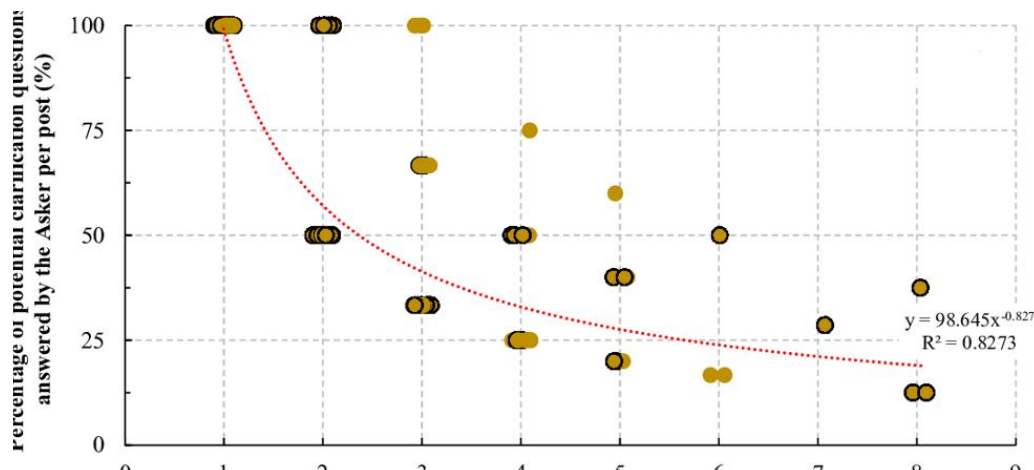


Figure 3.11: Fraction of answered clarifications per question in the Quantitative Finance site.

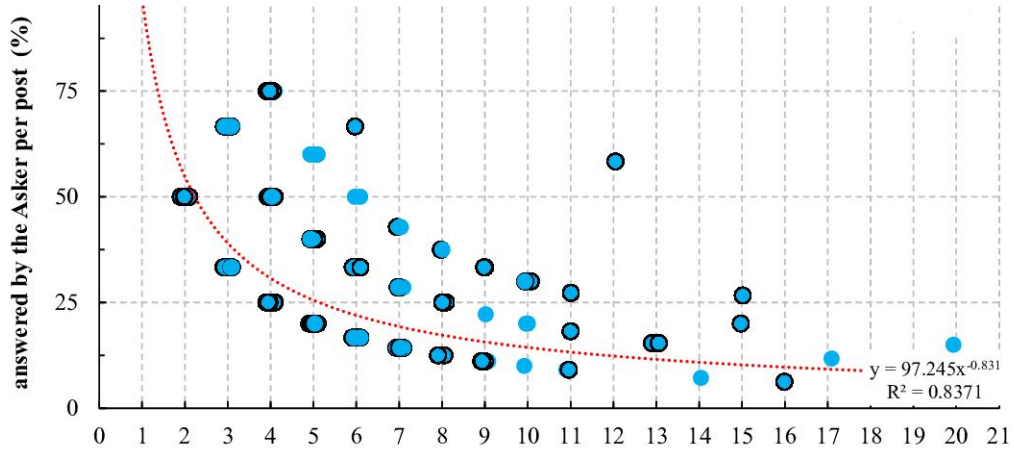


Figure 3.12: Fraction of answered clarifications per question in English Language site.

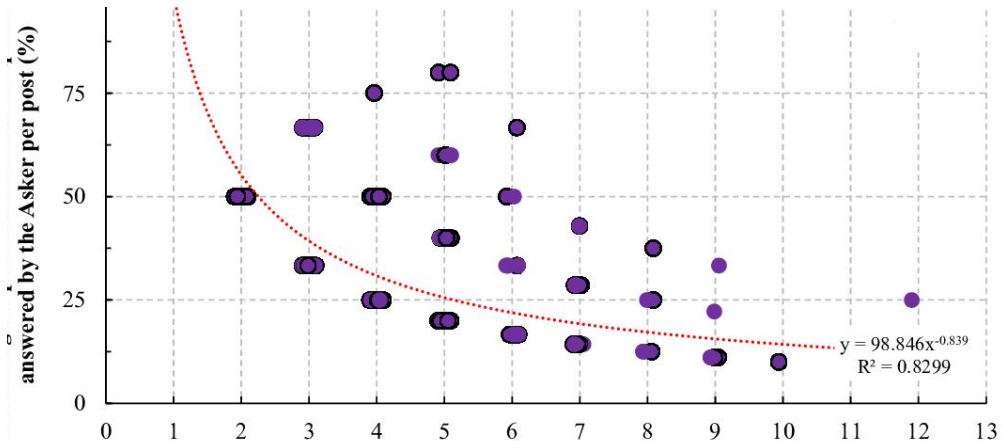


Figure 3.13: Fraction of answered clarifications per question in Science Fiction site.

3.3.2 Useful Clarifications

Our analysis indicates that more than 79% of posts with an accepted answer have three or fewer potential CQs answered by the Asker, regardless of the topic. The mean number of answered potential CQs per post is 1.13, with a mode value of 1 for all sites. Our findings showed that potential CQs answered by the Asker increase the chance of a post obtaining an accepted answer. Therefore, for the rest of this analysis, we consider a potential CQ as useful if it is answered by the Asker and the post receives an accepted answer. We consider a potential CQ as non-useful if it is left unanswered, but the post still obtained an accepted answer, i.e. answering the question was unnecessary to address Asker's information need.

We also examine the elapsed time between posting a CQ, the CQ receiving an answer, and the post obtaining an accepted answer to investigate if there is any relationship between them. However, we have not observed any noticeable trend. This is perhaps because *Stack*

Exchange is an online platform, and the users are located worldwide, so periods of activity can vary substantially.

In the previous subsection, we started to investigate what potential CQs are useful in terms of helping a post obtain an accepted answer. In this section, we study clarification usefulness using the manual annotation described in Section 3.2.2. We study what percentage of potential clarifications are actual CQs. We further investigate the potential CQs that have informative answers or help the post to achieve an accepted answer.

Table 3.4 indicates which potential CQs were labelled as actual clarifications. We can see that those answered by Askers are more likely to be actual CQs compared to those answered either by a Responder or left unanswered. Table 3.4 also shows that the average number of actual CQs varies across sites (71.6% in *QF* and 59.7% in *SF*, and 46.2% in *EL*); the differences are statistically significant (two-tailed z-test, $p < 0.00001$ for *SF* and *EL* sites; $p = 0.0198$ for *QF* and *SF* sites; and $p = 0.0067$ for *EL* and *SF* sites). The low average number of actual CQs in the *EL* site in comparison to the other two suggests that most of the CQs in this site might be a comment containing either sarcastic/humorous questions, rhetorical questions, off-topic questions as explained in Section 3.2.2. This again suggests that the nature of a topic and /or the nature of the user's interest in those topics could influence user behaviour regarding asking and responding to CQs.

Table 3.4: Percentage of Actual Clarification Questions.

| Answerer | <i>QF</i> | <i>EL</i> | <i>SF</i> |
|------------|-----------|-----------|-----------|
| Asker | 79.3 | 55.2 | 64.2 |
| Responder | 66.7 | 40.9 | 54.5 |
| Unanswered | 68.7 | 42.6 | 60.3 |
| Average | 71.6 | 46.2 | 59.7 |

We next investigate the usefulness of CQs in terms of their having an informative answer and being valuable for the post and the accepted answer. In all four of the graphs in Figure 3.14 we see that more than 90% of CQs answered either by the Asker or a Responder have an informative answer. This shows that CQs normally receive informative answers regardless of the topic. Figure 3.14 also demonstrates that the CQs answered by Askers are more valuable for the posts compared to clarifications answered either by a Responder or left unanswered. Moreover, such CQs are more valuable for the accepted answer in comparison to those answered by a Responder. This highlights the importance of those CQs which are answered by Askers. Findings from Table 3.3 in addition to the result presented in Figure 3.14 show that although Responders contribute more to answer potential CQs in the *EL* and *SF* sites compared to the *QF* site, their contribution in answering the CQs are less valuable for the accepted answers.

The results from the manual annotation helped us to get a better understanding of how

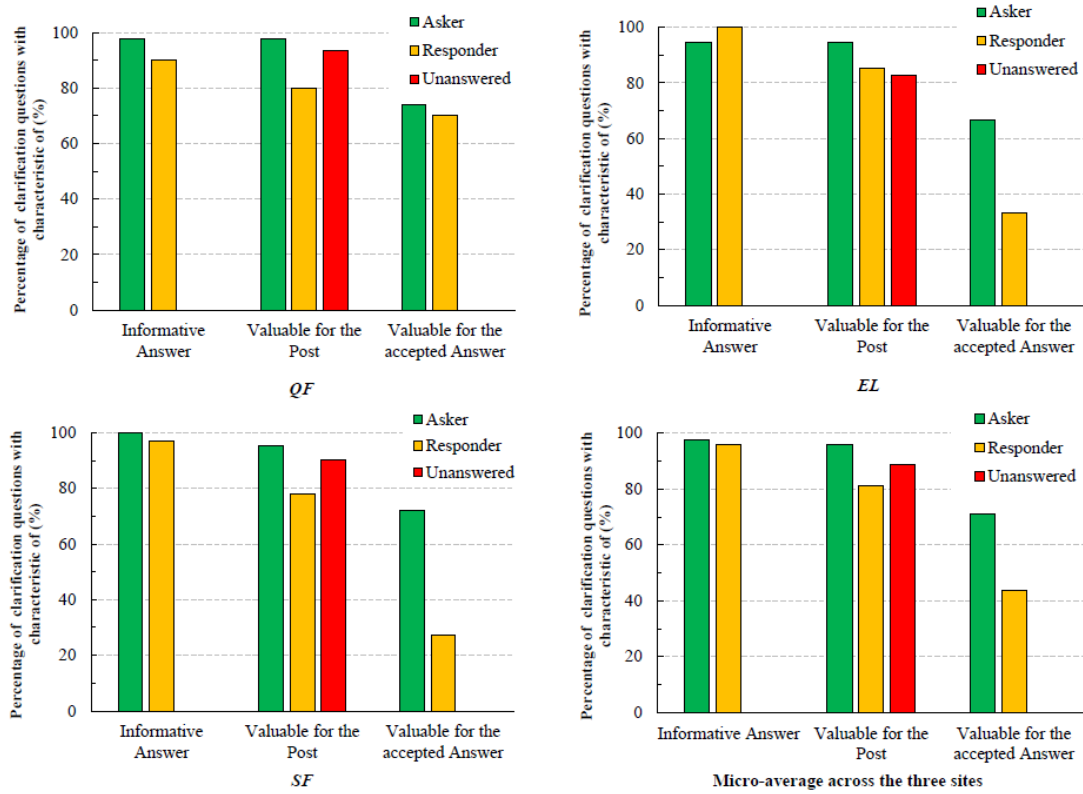


Figure 3.14: Characteristics of clarification questions answered either by the Asker or a Responder.

a CQ can be considered as useful in obtaining an accepted answer or non-useful as below:

- **Useful Clarification Questions:** We consider a CQ useful if it is answered by the Asker, has an informative answer, and is valuable for the post and the accepted answer.
- **Non-Useful Clarification Questions:** We consider a CQ as non-useful if it is left unanswered and is not valuable for the post, but the post still receives an accepted answer. This is because such CQs are not only lacking in value for the post but, moreover, answering them is not necessary to address the information need originally posted by the Asker.

The definition of useful and non-useful CQs can be applied in other studies on CQs and on other platforms. This is because the usefulness is defined based on the successful interaction with the user who submits the post and the answer to the clarification in this study, whether it is useful for the post to obtain an accepted answer or not.

3.3.3 Clarification Types and Patterns

Now, we analyse the extracted types and patterns of CQs based on the manual and automated annotation processes described earlier. The six identified types of CQs are presented below:

- **Ambiguity/Incompleteness:** The CQ asks about an unclear part of the post. The post is either ambiguous or some information is missing, which means further clarification or more details are required. In such cases, asking a CQ may lead to the post being revised (e.g. *“How much money did you assume to start with?”*)
- **Confirmation:** The Responder may ask a CQ to confirm her/his perception about a piece of information in the post. In some other cases, the Responder may want to emphasise important information in the post and confirm it with the Asker (e.g. *“Does it have to be a single word?”*).
- **General:** The CQ is a general question that does not refer to any specific part of the post. Such CQs can often be asked from all posts (e.g. *“Would you like to make the question more specific?”*).
- **Incorrectness:** When the Responder thinks there is wrong information in the post, this type of CQ is asked by the Responder to resolve the problem before providing an answer (e.g. *“Are you sure it is 62 and not 66?”*).
- **Paraphrasing:** The Responder attempts to paraphrase the post by asking CQs to make it more digestible and to understand the post correctly (e.g. *“Are you asking how to write an exchange simulator?”*).
- **Suggestion:** The Responder asks CQs to draw the Asker’s attention to a specific point, which can sometimes be a solution in the form of a suggestion, a reference, or an example (e.g. *“Can the book “Monte Carlo simulation in financial engineering” by Glasserman help you?”*).

Figure 3.15 shows the percentage of the type of the asked CQs in each site. It is evident that all three sites show similar trends for all types except *Suggestion* in the *SF* site. We can see that *Ambiguity* and *Confirmation* are the most common, followed by *General* and *Suggestion*. It is not unusual to find out that the type of *Ambiguity* is the most common type. This is because the CQs are mainly asked to eliminate any ambiguity or lack of information in the post.

To answer whether all CQs are useful (**RQ1**), we learnt in subsection 3.3.2 that a CQ is useful if it is answered by the Asker and helps the post to obtain an accepted answer. Now, as a first step toward answering **RQ2**, which is characterising useful CQs, we investigate if

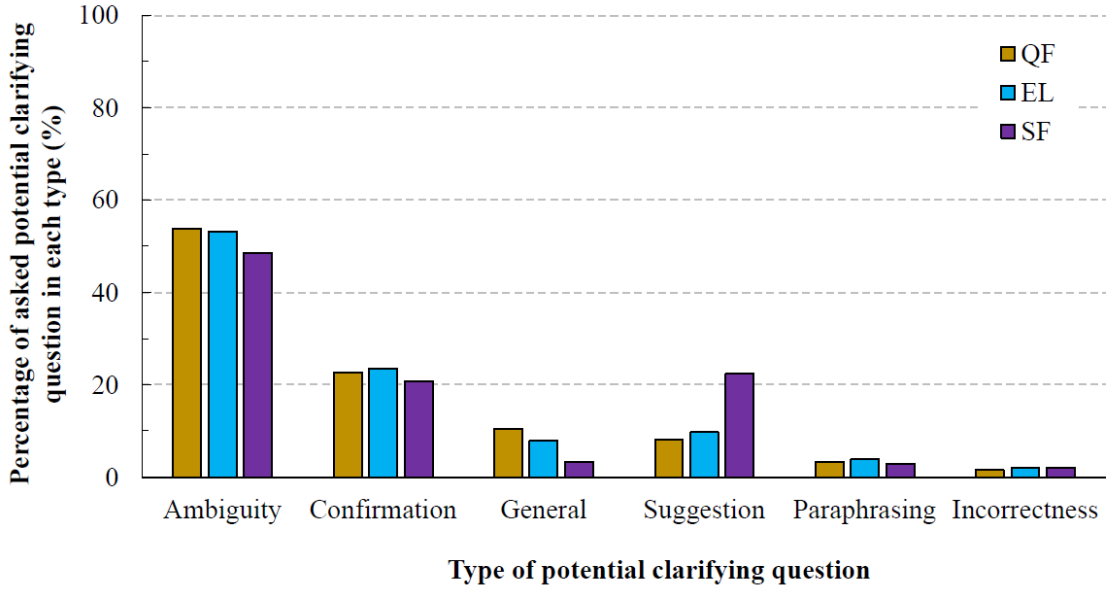


Figure 3.15: Distribution of clarifications by type.

there is a relationship between the type of CQ, Asker interaction, and the likelihood of a post obtaining an accepted answer. We define a series of measures as follows:

- $ClarQ_{Type}$: The type of clarification question.
- $ClarQ_{Asker}$: A clarification question is answered by the Asker.
- $P_{ClarQ_{Type}}$: A post contains a particular clarification type.
- P_{AccAns} : A post contains an accepted answer.

We graph the following conditional probabilities:

- $P(ClarQ_{Type} | ClarQ_{Asker})$: The conditional probabilities in Figure 3.16 show the *Ambiguity* type is most answered by Askers, however, *Ambiguity* is the commonest type of clarification question.
- $P(ClarQ_{Asker} | ClarQ_{Type})$: Figure 3.17 indicates that relative to the number of clarifications of each type, the probability of the Asker answering clarifications is generally even across each site. There are some site-specific variations, however.
- $P(P_{ClarQ_{Type}} | P_{AccAns}, ClarQ_{Asker})$: Figure 3.18 shows that given that a post with an accepted answer has a clarification question, which has been answered by the Asker, the clarification has a high chance to be of type *Ambiguity*. However, this is because clarification questions of this type are asked more.

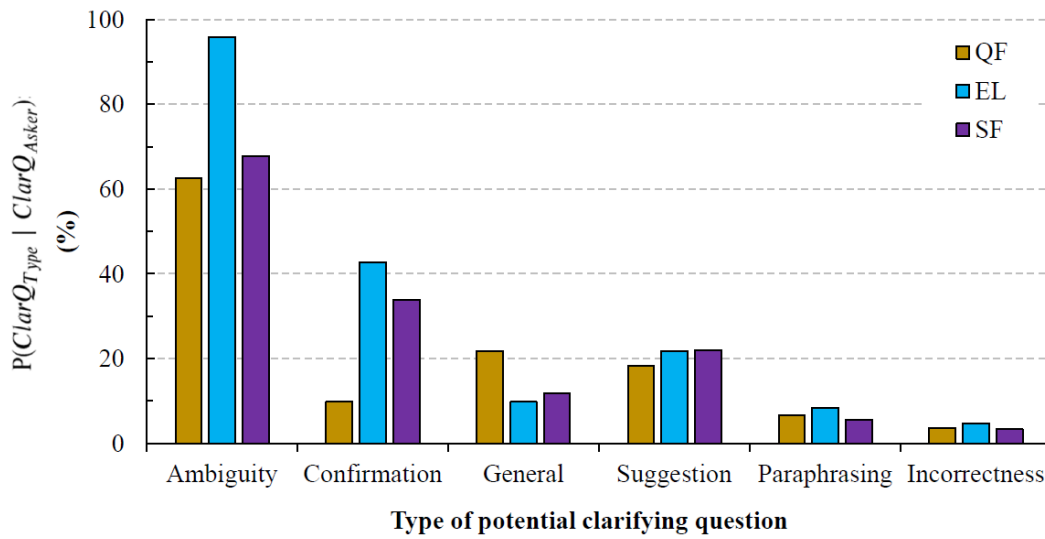


Figure 3.16: The probability of a clarification question that has a certain type, given that the clarification question is answered by the Asker.

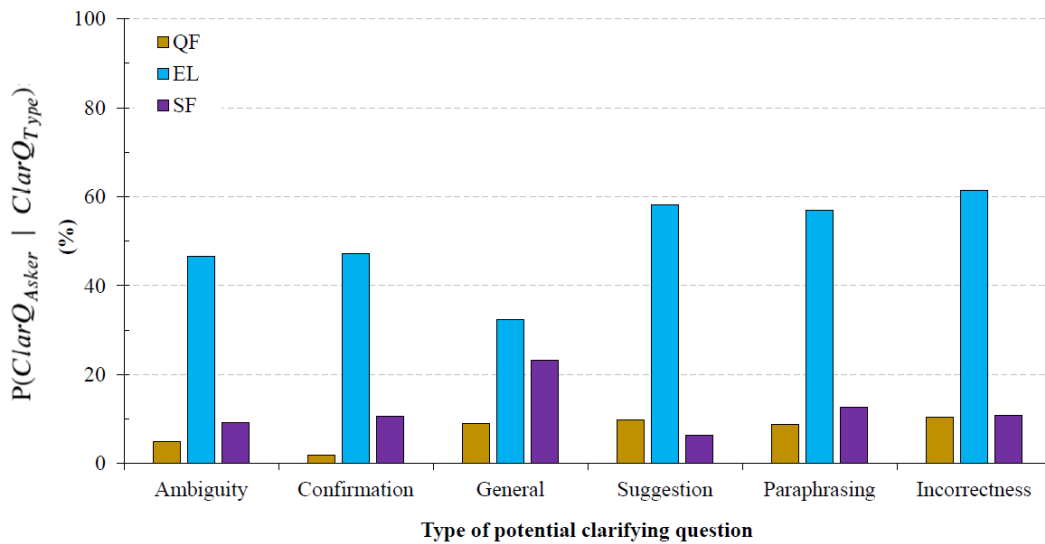


Figure 3.17: The probability of a clarification question being answered by the Asker, given a particular clarification question type.

- $P(P_{AccAns} | P_{ClarQ_{Type}}, ClarQ_{Asker})$: The graph of this conditional probability (Figure 3.19) shows that clarification type is less important to a post having an accepted answer if the Asker answers a clarification question. We also see that when clarification questions are answered in the *EL* and *SF* sites, regardless of the type, the post gets an accepted answer almost always.

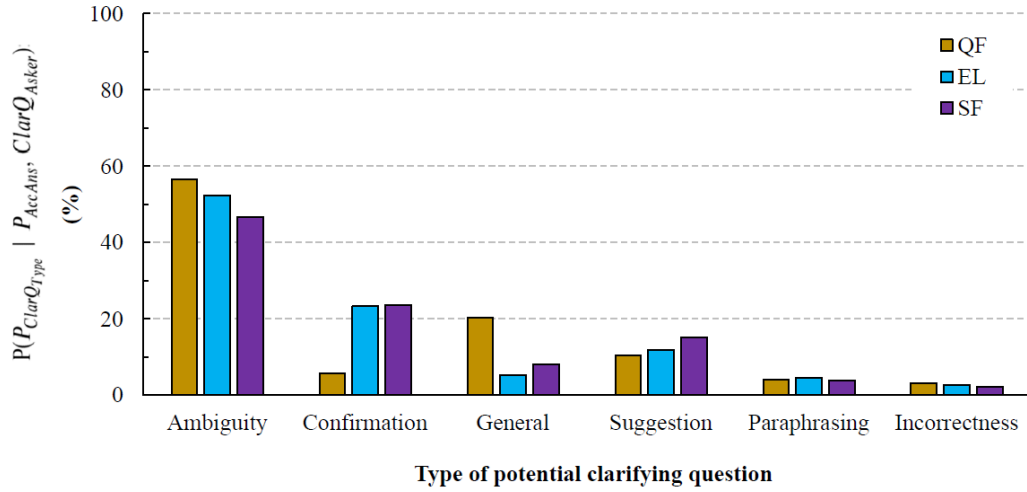


Figure 3.18: The probability of a post having a particular clarification question type, given that the post has an accepted answer.

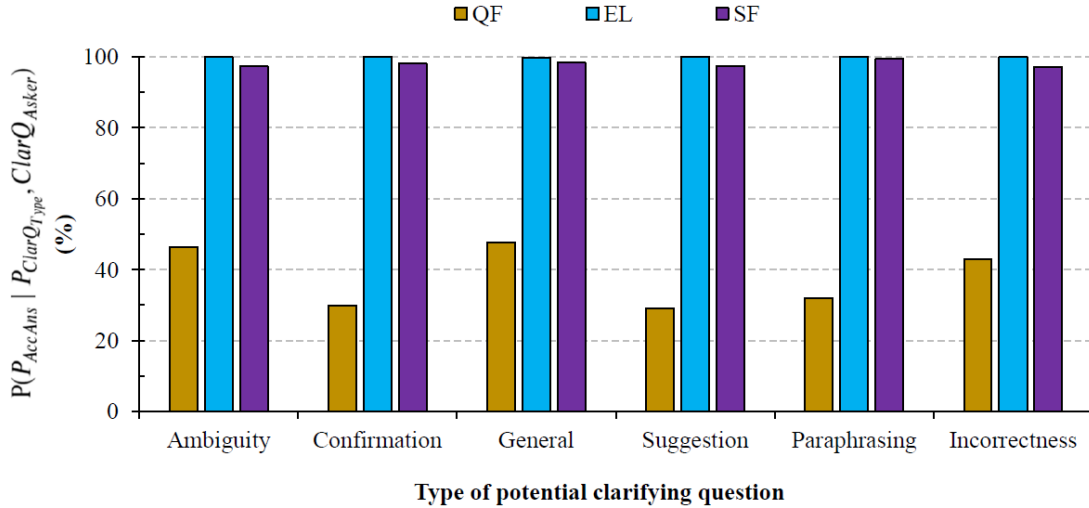


Figure 3.19: The probability of a post having an accepted answer, given the presence of a clarification question of a specific type.

In the final stage, we detect the patterns of useful CQs and investigate if there is any relationship between CQs answered by the Asker, posts with an accepted answer, and the type of CQs. We then compare and analyse the highest frequency patterns for the useful and non-useful CQs. We noticed that the patterns exhibited by useful CQs align with the top 20 identified patterns. This observation suggests that we have successfully identified the patterns associated with useful CQs, irrespective of the size of the sampled dataset. We also notice there are 65%, 65%, and 85% similarities in the most popular patterns (top 20 patterns) of the two sites of *QF* and *EL*; the two sites of *QF* and *SF*; and the two sites of *EL* and *SF*, respectively.

We present the most popular patterns (the share and sequence of terms) of useful CQs

with respect to the type of CQs, collectively extracted from three sites, in Figure 3.20. The gaps in this figure are patterns with a low frequency compared to others, which are not shown. The patterns “*Is it (noun)*” and “*Is there a/any*” in the type of *Ambiguity/Incompleteness*, the pattern of “*Do you mean*” in the type of *Confirmation*, the pattern of “*Can you give*” in the type of *General*, the pattern of “*Are you asking/looking*” in the type of *Paraphrasing*, the pattern of “*Are you sure*” in the type of *Incorrectness* and the pattern of “*Do you know*” in the type of *Suggestion* are the patterns with the highest frequency in the useful CQs. We also observe that apart from the pattern of “*Is it (noun)*” which exists in both types of *Ambiguity/Incompleteness* and *Confirmation*, there are no other similar patterns in different types of useful CQs. Our pattern analysis also shows that more than 80% useful CQs can be generated with 25 patterns. This finding suggests that the identified patterns can be used in asking/generating useful CQs in information-seeking systems.

After studying the characteristics, types and patterns of useful CQs, we investigate the differences between the pattern of useful and non-useful clarifications. We compare popularity distributions of the patterns of the useful ($P(x)$) and non-useful CQs ($Q(x)$) by computing point-wise Kullback-Leibler Divergence (D_{KL}). The popularity of distribution is based on the frequency of patterns. Figure 3.21 shows the top ten and bottom ten patterns by D_{KL} score. Positive-scored patterns are used more in useful CQs, while negative scores are patterns used more in non-useful clarifications. Comparing Figures 3.20 and 3.21 shows that there are some popular patterns that are shared in both useful and non-useful CQs. However, some of them are more common in one group. For example, the patterns *Do you mean* and *What is/are (noun)* are in popular patterns of useful CQs presented in Figure 3.20. However, the pattern of *Do you mean* is more likely to be asked in useful clarifications. In contrast, the pattern *What is/are (noun)* is common with non-useful clarifications.

3.4 Discussion

We understood from Table 3.3 that there are notably fewer Responders answering in the *QF* site compared to the other sites. We speculate that the nature of a site’s topic influences the behaviour of users in engaging with CQs. In domains where expert knowledge is required, e.g., *QF*, the questions are more specific, where the Asker is the only person who can answer.

We also observed similar trends in three domains in terms of the number of asked potential CQs per post with an accepted answer versus the percentage of potential CQs that was answered by the Asker in each post. This observation contrasts with the findings of Zou et al. [135], who investigated clarification question-based systems and showed that Askers normally answer 11-21 CQs on average. We conjecture that underlying differences between the analysed platforms account for this divergence: the architecture and the nature of Zou et al. ’s system was different from a typical QA forum: while we investigate natural language questions on QA forums, Zou et al. [135] performed their experiments in the domain of the *Amazon* retail platform, where a user answered questions generated by the system with a “Yes”, a “No” or a “Not Sure”, to find a target product to buy.

The usefulness of clarifications in this study was investigated using manual annotation, and to enhance the accuracy of the analysis, we sampled 557 potential CQs with their answers for labelling, which was more than double the sample size of the previous study conducted by Braslavski et al. [14]. Our annotation showed that those CQs answered by the Asker are more valuable in terms of adding value to the post and the accepted answer. The findings led to the proposing of new definitions to distinguish useful and non-useful CQs.

As the first step towards understanding how to generate useful CQs in information-seeking conversations, we attempted to find the type and patterns of such clarifications. When we detected the type and patterns of the useful and non-useful CQs, we applied them to all potential CQs in three investigated sites and found similar trends. While the top 20 patterns and related types of useful CQs extracted from manual annotation and useful potential CQs extracted by the automatic classifier were found to be all the same, in type recognition analysis presented in Figures 3.15 to 3.19 we only focused on potential CQs. However, comparing the results with manually labelled CQs, we found that the results are consistent.

Our type classification was also more detailed compared to previous studies [54, 14]. As mentioned in Chapter 2, the type of *Reason* proposed by Kato et al. [54] questioned a general aspect of the post (*General* type) or an unknown aspect (*Ambiguity* type). We decomposed it into the *General* and *Ambiguity* types. In addition, the types of *Selection* and *Check* had noticeable similarities, and we merged them into the type *Confirmation*.

The types *More Info* and *Experience* in previous studies were the same as our types of *Ambiguity/Incompleteness* and *Suggestion*, respectively. In this work, we also introduced two new types: *Paraphrasing* and *Incorrectness*.

We also noticed that there are some potential CQs in each type that are not actual questions. For example, in the type of *Suggestion*, the Responder normally did not look for any answer (e.g. “*Monte Carlo simulation in financial engineering*” by Glasserman help you?). This case was found to be dominant in the type of *Suggestion*. Nevertheless, even in our study, types may overlap in some cases, which means a CQ may belong to more than one type. For example, the CQ “*Have you considered using the Fama/French market factor as a reference?*” can fall into either the *Confirmation* or *Suggestion* types; in such cases, it was assigned to both types.

Comparing our results with the previous study carried out by Braslavski et al. [14] showed us the most frequent patterns are different when we investigate all CQs (regardless of their answers) compared with those useful CQs that are answered by the Asker and help the post to get an accepted answer. Even some of the most frequent patterns of general CQs suggested by Braslavski et al. [14] fall into non-useful CQs (Figure 3.21). This highlights the importance of distinguishing useful CQs from non-useful ones. The same thing was observed in the types of CQs. The type of *Ambiguity/Incompleteness* was found to be well above others in terms of frequency in the useful CQs, which is different from the previous finding. It is worth noting that three completely different sites of *Quantitative Finance* (from Business domain), *English Language and Usage* (from Culture/Recreation domain) and *Science Fiction and Fantasy* (from Life/Arts domain), were chosen for this study. While the nature of these domains is very different, we observed similar trends in terms of user interaction and type and patterns of CQs. We also mentioned earlier that there are 65%, 65%, and 85% similarities in the most popular patterns (top 20 patterns) of the two sites of *QF* and *EL*; the two sites of *QF* and *SF*; and the two sites of *EL* and *SF*, respectively. This suggests that the findings can be applicable to cross-domain.

3.5 Summary

As the first step toward the understanding of useful and engaging CQs in information-seeking systems, in this chapter, we explored conversations extracted from multiple *Stack Exchange* sites, studied human-generated CQs based on their answers, and provided insights into useful clarifications—those whose answers help reach accepted answers of user information needs. We also investigated their characteristics, including their types and patterns, through a fine-tuned manual annotation. We examined and answered the following research questions:

- What clarification questions are more useful? (in terms of helping the Asker to get a correct answer)?
 - We found that many CQs are left with no answers. Furthermore, some of them do not add value to the post, i.e., answering the user information needs is independent of those clarifications. This shows that it is important to identify and characterise useful clarifications. Therefore, we discerned useful CQs from non-useful ones based upon the results.
- What are the characteristics of useful clarification questions?
 - We classified useful CQs into different types based on user intents and extracted their patterns. Our analysis showed that the type of *Ambiguity/Incompleteness* is the most frequent compared to the other types, regardless of the topic. Moreover, we showed that three other types, *Confirmation*, *General* and *Suggestion*, are also useful for the post to obtain an accepted answer as they can lead to about 41.27% more successful resolution of information needs. Investigation of useful and non-useful CQs showed that there are specific patterns, which are not only the most common for useful CQs but also are less asked for non-useful CQs. Such patterns can be employed by information-seeking systems for generating CQs that are more likely to result in user engagement with a conversational system.

In the next two chapters, we will look at model-generated CQs in a search engine to provide a broad insight into what makes a CQ engaging from a user's point of view.

Chapter 4

Asking Engaging Clarification Question in Search Engines: Task Formulation and Limitations

In the previous chapter, we explored the role of human-generated clarification questions (CQs) in community question-answering forums. We showed that a useful CQ has an informative answer and is valuable for the post to obtain an accepted answer. In this chapter, we shift our focus to investigating model-generated CQs in search engines.

4.1 Introduction

As the reliance on search engines continues to grow, the importance of effective information retrieval has become increasingly paramount. While the Search Engine Results Page (SERP) features play a crucial role in enhancing various tasks in the field of information retrieval, including improving the user's search experience, the relevance of search results, and providing valuable insights into specific entities or concepts. They cannot address the user's information need when the query is complex or ambiguous. In such cases, one key aspect of the search system can be the ability to ask CQs. Although clarification-generating models are able to generate multiple CQs for a given query, they still face the challenge of recommending the most effective CQ that is likely to engage the user from among the various options generated. The ultimate goal of clarification generating and selecting

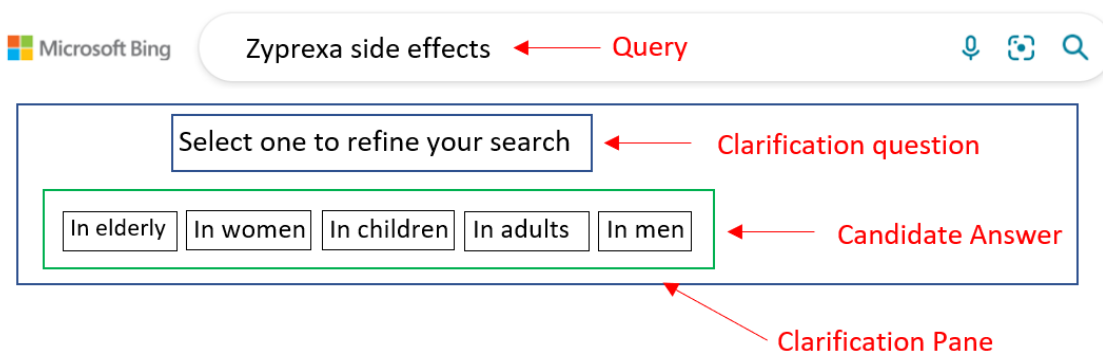


Figure 4.1: Example of query and clarification pane (i.e., A clarification question plus candidate answers).

models is to increase the probability of a user getting engaged with the CQ when the retrieved documents are not satisfactory. In the absence of user interaction signals (online evaluation) or human annotation (offline evaluation) for every generated clarification pane (CP) (i.e., a clarification question and its candidate answers, see Figure 4.1) for a given query in a real scenario, we aim to understand whether SERP features can help us identify the most engaging clarification question from a user’s perspective or not.

Similar to other tasks in IR, any new clarification generating and selecting model needs to be evaluated using both online and offline evaluations. Offline evaluations involve testing a model using pre-collected datasets and predefined evaluation metrics, while online evaluations involve testing a model using real-time data and user interactions.

The reason why offline evaluations are preferred over online evaluations is that they provide a controlled environment for testing and fine-tuning a model. With offline evaluations, researchers can test a wide range of model variations and evaluate their performance without the influence of external factors such as user behaviour or changes in the data distribution. This allows researchers to quickly and efficiently evaluate and compare different models, which can save time and resources. However, while offline evaluations are useful for initial testing and fine-tuning, they are not always a reliable indicator of a model’s performance in a real-world scenario. This is where online evaluations come in. By testing a model in a live environment, researchers can assess its performance in real time and collect data on user behaviour and interactions. This data can be used to further improve the model and identify areas where it may be lacking. However, it is important to have a clear understanding of how online and offline evaluations correspond with each other. Otherwise, any conclusion might be misleading. In this chapter, we aim to address two research questions as follows:

- Can the SERP feature help us identify the most engaging clarification question from a user’s Perspective?
- Is there any relationship between online and offline evaluations in search clarification

using the *MIMICS* dataset (the only search clarification dataset)?

To answer the first research question, we focus on the Microsoft *Bing* search engine that presents a clarification pane to the user after submitting a query. The user then is free to engage with the CP if the landing results (retrieved documents) are not satisfactory to narrow down the search results. We formulate the task of identifying the most engaging clarification pane (MECP) for a given query as a learning-to-rank (LTR) problem using SERP features. We train and test the LTR models on the *MIMICS* dataset, which is the only available search clarification dataset containing both online and offline evaluations for clarification panes. The *MIMICS* dataset also provides us with the opportunity to investigate the relationship between online and offline evaluations in search clarification to answer the second research question by training the LTR model using the online dataset and testing on the offline dataset and vice versa.

4.2 Methodology

In this chapter, our aim is to propose a model to perform the task of ranking the CPs and identifying the top one for a query using the *MIMICS* dataset [125]. The *MIMICS* dataset is a collection of clarification datasets sampled from the *Bing* query logs. The evaluation metrics are overall user engagement level and graded quality labels of clarification questions.

The *MIMICS* dataset provides a set of information for each query-clarification pair including query context, retrieved document title, retrieved document snippet, related search context, video title, and video description. As described in Chapter 2, this dataset consists of three datasets as below:

MIMICS-Click: Includes over 400,000 unique queries, their associated CPs, the corresponding aggregated user interaction signals including an Impression Level (low, medium, or high), an *Engagement Level* (i.e., an integer between 0 to 10 presenting the level of total engagement received by users in terms of click-through rate), and the conditional click probability for each individual candidate answer. This dataset contains queries with only one CP.

MIMICS-ClickExplore: Includes over 60,000 unique queries. Similar to *MIMICS-Click*, it contains queries' associated CPs, the corresponding aggregated user interaction signals and the conditional click probability for each individual candidate answer. However, contrary to *MIMICS-Click*, this dataset contains queries with multiple CPs.

Each query is associated with multiple CPs in addition to the user interaction signals, similar to *MIMICS-Click*.

MIMICS-Manual: Includes over 2,000 unique search queries with multiple CPs, landing results pages, and manually annotated quality labels (i.e., the label is assigned based on how

fluent and grammatically correct the CP is and how accurately the CP addresses different intents of the query. Each CP is given a label 2 (Good), 1 (Fair), or 0 (Bad).

Since the aim of this study is to rank clarification questions for each query, we only focus on *MIMICS-ClickExplore* and *MIMICS-Manual* containing queries with multiple CPs, contrary to *MIMICS-Click*, which contains queries with only one CP. *MIMICS-ClickExplore* can be seen as a dataset for online evaluation because it contains signals based on user engagement with *Bing* CPs. However, *MIMICS-Manual* contains manual expert annotations for clarification quality, and it follows the traditional approach for offline evaluation.

A close look at the *MIMICS-ClickExplore* dataset unveils that there are several queries that have two identical CPs but with different *Impression* or *Engagement* levels. One CP has to be eliminated from every two identical CPs for a given query in the *MIMICS-ClickExplore* dataset, as it is not possible for a unique CP receives two different impression or engagement levels. This cleaning process leaves 708 queries with only one CP that cannot be used in the task of ranking search clarification and has to be removed from the dataset. Therefore, of 64,007 initial queries, only 63,299 queries can be used in this study. We also notice there are several queries that their associated CPs have received the same *Engagement* levels. These queries are also not suitable for the task of identifying the MECP as ranking the list of CPs based on the *Engagement* has to be random, making them inappropriate to be considered as ideal ranked lists. We also removed these queries, leaving the dataset with only 61,222 queries for this study.

There is no repetitive query-clarification pair in *MIMICS-Manual*; however, the majority of queries in this dataset have only one CP or all their CPs receive the same quality label. Consequently, this dataset only contains 66 queries useful for this study, a major limitation of *MIMICS-Manual*. Table 4.1 shows the statistics of the datasets.

4.2.1 Task Formulation

To answer the research question, first, we study any relationship between SERP elements and the *Engagement Level* (provided in *MIMICS-ClickExplore*) and *Overall Quality* label (provide in *MIMICS-Manual*) of CPs. In other words, we investigate if SERP elements can help us rank the CPs similar to the ideal CP ranked lists created based on our ground truths, *Engagement Level* and *Overall Quality*. The main inputs to LTR models are TF-IDF with Cosine Similarity [85], BM25 [90] and Overall Match algorithms [40] and query length. An extensive set of features and their combinations are explored – 110 features in total, as shown in Table 4.2.

We also generate two Non-SERP features of “Number of options for each clarification question” and “Is a clarification question a question or a statement?” to distinguish two

Table 4.1: Statistics of the datasets.

| Property | MIMICS-ClickExplore | MIMICS-Manual |
|---|---------------------|-------------------------------------|
| No. of unique queries | 64,007 | 2,464 |
| No. of CPs per query | 2.64 ± 1.11 | 1.15 ± 0.36 |
| Min. no. of CPs per query | 2 | 1 |
| Max. no. of CPs per query | 89 | 3 |
| No. of deleted queries with at least two identical CPs | 708 (1.11%) | 0 |
| No. of unique queries after removing repeated queries | 63,229 | 2,464 |
| No. of unique queries with more than two CPs | 63,299 | 367 (2097 queries with only one CP) |
| No. of unique queries with more than two CPs and at least one CP with a different <i>Engagement Level</i> | 61,222 | 66 |
| No. of queries without the related search | 7,585 | 14 |
| No. of queries without video | 7,544 | 52 |
| No. of queries without related search and video | 1,142 | 3 |

clarification panes generated for a query and also to consider the format of the clarification question itself. The features are linearly normalised based on their min/max values. The code for extracting these features and their descriptions are available on GitHub¹.

Several LTR algorithms, including MART, RankNet, RankBoost, Coordinate Ascent, LambdaMart, and Random Forests, are employed in this study to perform the task of ranking CPs. A brief description of each algorithm is provided below:

- *MART* [37]: Mart is a Multiple Additive Regression Tree, which is a boosted prediction model formed from an ensemble of decision trees.
- *RankNet* [16]: RankNet uses a neural net as its function class, and feature vectors are computed for each query/clarification pair.
- *RankBoost* [36]: This algorithm searches for a weak learner that maximises the pairwise ranking accuracy. The accuracy is defined as the number of document pairs that receive the correct ranking.
- *Coordinate Ascent* [107]: Coordinate Ascent is a gradient-based listwise method that directly optimises mean average precision (MAP).

¹<https://github.com/Leila-Ta/Clarification-LTR-Features>

Table 4.2: LTR Features and Feeding Inputs.

| Type | Feature | Input 1 | Input 2 | # of Features |
|----------|--|---|-------------------------------|---------------|
| SERP | - TF-IDF using Cosine Similarity - BM25 Similarity - Overall Term-Matching | Query Document Title | CP | 18 |
| | | Related search Video title Video description Snippet | CQ + Candidate answer # | 18@5: 90 |
| Non-SERP | Number of candidate answers for each clarification question | Not applicable | Not applicable | 1 |
| | Is a clarification question a question or a statement? | Question: 1 Statement: 0 | Not applicable | 1 |

- *LambdaMart* [116]: This algorithm employs gradient boosting or MART. The gradient function is the loss function and is called the lambda function.
- *RandomForests* [15]: Random Forests are a collection of tree-based regressors where each of them casts a prediction for a given input instance. The final prediction is the arithmetic mean of the scores, which is produced by the regressors in the forest.

We train and tune the employed LTR models in a cross-validation manner. We use a 10-fold split of the training set into training and development sets. The models are trained on one dataset and then tested on the same dataset. To normalise each feature, we use linear normalisation based on its min/max values.

4.2.2 Experimental Results

In this section, we answer two research questions we discussed earlier.

Using SERP Feature to Identify the Most Engaging Clarification Pane from a User’s Perspective

We first rank clarification panes in the *MIMICS-ClickExplore* dataset using the *Engagement Level* and in the *MIMICS-Manual* dataset using the *Overall Quality*. Next, the CPs for each query are ranked using the LTR models and the generated features. Then, we compare two ranked lists (i.e., clarification ranked lists created by LTR models and ideal clarification ranked lists created based on either *Engagement Level* or *Quality Label* for the queries in *MIMICS-ClickExplore* and *MIMICS-Manual* datasets, respectively).

It is essential to have appropriate evaluation metrics that can accurately measure the performance of the LTR models in creating ranked lists. In information retrieval, precision and mean reciprocal rank (MRR) are two commonly used metrics to evaluate the effectiveness of ranked lists. Precision measures the fraction of relevant items in the top k results

Table 4.3: Performance of LTR models based on P@1, trained and tested on *MIMICS-ClickExplore*. The ground truth is the ranked lists based on the *Engagement Level*.

| LTR Models | Features | All Features | SERP Features | Non-SERP Features |
|-------------------|----------|--------------------|--------------------|--------------------|
| | | | | |
| MART | | 0.431 [†] | 0.448 [†] | 0.477 [†] |
| RankNet | | 0.439 [†] | 0.476 [†] | 0.490 [†] |
| RankBoost | | 0.474 [†] | 0.497 [†] | 0.572 [†] |
| Coordinate Ascent | | 0.454 [†] | 0.472 [†] | 0.474 [†] |
| LambdaMART | | 0.437 [†] | 0.469 [†] | 0.456 [†] |
| RandomForests | | 0.433 [†] | 0.449 [†] | 0.481 [†] |
| Mean | | 0.440 | 0.469 | 0.492 |
| Random Ranker | | 0.325 | 0.346 | 0.377 |

[†] Significantly different from the Random Ranker baseline (Student's t-test, $p < 0.05$).

of a ranked list, where k is a fixed threshold. MRR is calculated by the position of the top-rated document, here clarification pane, according to the *Engagement Level* or *Quality Label*, depending on the dataset being used for evaluation. These metrics provide valuable insights into the relevance and accuracy of the ranked lists, allowing us to determine which list performs better for a given task.

Tables 4.3 and 4.4 present the values of precision for each LTR model on both datasets considering different types of features. We can see from this Table that the LTR models except *LambdaMART* model performed better when using Non-SERP features as the input features for the models, although their better performances were marginal. Our first research question was to investigate whether the SERP features improve the task of identifying the MECPs, and this experiment showed that it seems SERP features have no positive impact on the task. However, We also randomly ranked the CPs using a Random Ranker, and we observed that LTR models performed better than a Random Ranker. Table 4.4 shows the performance of LTR models on the *MIMICS-Manual* dataset. The performance of LTR models considering SERP features started to show some improvements for some LTR models. However, the improvements were not significantly different.

We showed in Table 4.2 that we used *Related search* and *Video title or description* as two SERP-based inputs for LTR models. However, an investigation of *MIMICS-ClickExplore* shows that about 22.12% of queries have no related search or video (perhaps this is a limitation of the dataset), which may impact the performance of the LTR models. Therefore, in the second step, we repeat the experiment without considering the related search- and video-based features. Tables 4.5 and 4.6 show the results of the experiments on the *MIMICS-ClickExplore* and *MIMICS-Manual* datasets, respectively. Comparing Tables 4.5 and 4.6 with Tables 4.3 and 4.4 does not show any significant changes in the performance of LTR models. The findings indicate that using SERP features as inputs for LTR models cannot enhance the task of ranking CPs and identifying the MECPs.

Table 4.4: Performance of LTR models based on P@1, trained and tested on *MIMICS-Manual*. The ground truth is the ranked lists based on the *Overall Quality* label.

| LTR Models | Features | | |
|-------------------|--------------------|--------------------|--------------------|
| | All Features | SERP Features | Non-SERP Features |
| MART | 0.489 [†] | 0.497 | 0.499 [†] |
| RankNet | 0.530 [†] | 0.529 [†] | 0.527 [†] |
| RankBoost | 0.473 [†] | 0.518 [†] | 0.565 [†] |
| Coordinate Ascent | 0.448 [†] | 0.499 | 0.491 [†] |
| LambdaMART | 0.464 [†] | 0.518 [†] | 0.507 [†] |
| RandomForests | 0.525 [†] | 0.486 | 0.418 |
| Mean | 0.488 | 0.508 | 0.501 |
| Random Ranker | 0.365 | 0.446 | 0.439 |

[†] Significantly different from the Random Ranker baseline (Student's t-test, $p < 0.05$).

Table 4.5: Performance of LTR models based on P@1, trained and tested on *MIMICS-ClickExplore*, without considering the related search- and video-based features. The ground truth is the ranked lists created using the *Engagement Level*.

| LTR Models | Features | | |
|-------------------|--------------------|--------------------|--------------------|
| | All Features | SERP Features | Non-SERP Features |
| MART | 0.415 [†] | 0.445 [†] | 0.468 [†] |
| RankNet | 0.417 [†] | 0.477 [†] | 0.486 [†] |
| RankBoost | 0.444 [†] | 0.493 [†] | 0.565 [†] |
| Coordinate Ascent | 0.444 [†] | 0.458 [†] | 0.467 [†] |
| LambdaMART | 0.424 [†] | 0.457 [†] | 0.449 |
| RandomForests | 0.417 [†] | 0.445 | 0.470 [†] |
| Mean | 0.427 | 0.463 | 0.484 |
| Random Ranker | 0.355 | 0.340 | 0.361 |

[†] Significantly different from the Random Ranker baseline (Student's t-test, $p < 0.05$).

Comparing *MIMICS-ClickExplore* with *MIMICS-Manual* shows that while the engagement level varied between 0 to 10, only the values of 0, 1 and 2 could be assigned to the *Overall Quality* label. These different scoring may impact the performance of LTR models on the *MIMICS-ClickExplore* with *MIMICS-Manual* datasets. Therefore, in the third round, we repeat the experiments on the *MIMICS-ClickExplore* dataset using all features as the inputs for the LTR models but with a new arrangement for the *Engagement Level*. The original *Engagement Levels* are mapped as below:

Engagement Level = [0] \rightarrow [0]

Engagement Level = [1, 2, 3, 4, 5] \rightarrow [1]

Engagement Level = [6, 7, 8, 9, 10] \rightarrow [2]

Figures 4.2 and 4.3 present the changes in the performance of the LTR models after

Table 4.6: Performance of LTR models based on P@1, trained and tested on *MIMICS-Manual*, without considering the related search- and video-based features. The ground truth is the ranked lists created using the *Overall Quality* label.

| LTR Models | Features | | |
|-------------------|--------------------|--------------------|--------------------|
| | All Features | SERP Features | Non-SERP Features |
| MART | 0.425 [†] | 0.502 [†] | 0.513 [†] |
| RankNet | 0.404 [†] | 0.548 [†] | 0.532 [†] |
| RankBoost | 0.425 [†] | 0.524 [†] | 0.565 [†] |
| Coordinate Ascent | 0.420 [†] | 0.510 [†] | 0.505 [†] |
| LambdaMART | 0.411 [†] | 0.502 [†] | 0.540 [†] |
| RandomForests | 0.417 [†] | 0.513 [†] | 0.518 [†] |
| Mean | 0.417 | 0.517 | 0.529 |
| Random Ranker | 0.461 | 0.474 | 0.406 |

[†] Significantly different from the Random Ranker baseline (Student's t-test, $p < 0.05$).

the *Engagement Level* mapping in the *MIMICS-ClickExplore* dataset for three scenarios of considering all features, only SERP features and only Non-SERP features with and without using related search and video-based features. It is evident that the performance of the LTR models on the *MIMICS-ClickExplore* dataset enhanced to some extent but yet well below their performances on the *MIMICS-Manual* dataset. The improvement was found to be significantly different according to Student's t-test, $p < 0.05$. In conclusion, the average performance of LTR models in terms of P@1 showed that the SERP features cannot support the task of identifying the MECP for a given query.

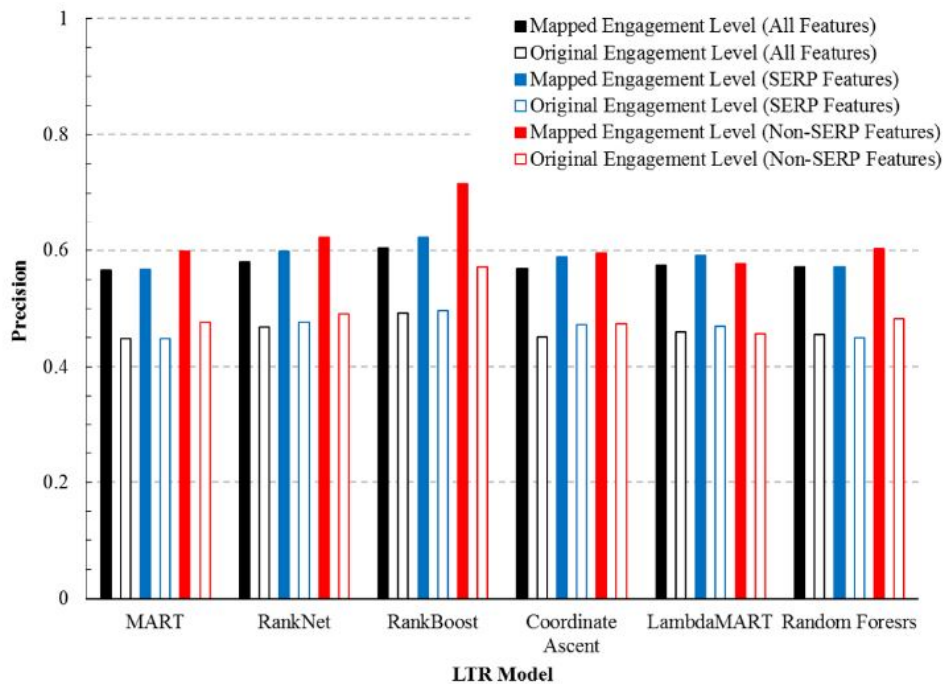


Figure 4.2: Performance of LTR models after mapping the *Engagement Level* considering the related search- and video-based features.

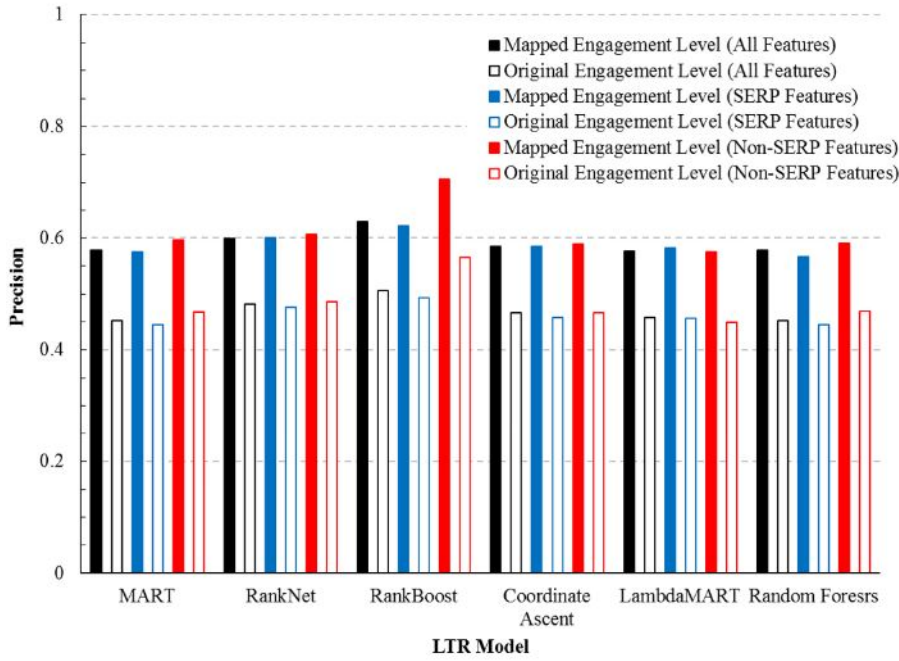


Figure 4.3: Performance of LTR models after mapping the *Engagement Level* without considering the related search- and video-based features.

Relationship Between Online and Offline Evaluations in Search Clarification

In all experiments so far, the LTR models were trained on one dataset and tested on the same dataset to focus solely on the impact of the SERP features on the task of ranking CPs and selecting the MECP. Here, we aim to investigate the relationship between online and offline evaluations in search clarification to answer the second research.

We now train LTR models on *MIMICS-ClickExplore* and evaluate them on both *MIMICS-ClickExplore* and *MIMICS-Manual* and compare the performance of the models when we consider all features as the models' input. We repeat this experiment contrariwise and train LTR models on *MIMICS-Manual* and evaluate them again on both *MIMICS-Manual* and *MIMICS-ClickExplore*. When the test dataset is *MIMICS-ClickExplore*, our ground truth is the *Engagement Level*, and when the test dataset is *MIMICS-Manual*, our ground truth is *Overall Quality* label.

At this stage, we also measure the MRR metric as the second evaluation criterion. Tables 4.7 and 4.8 show the results of the experiments. According to the results shown in 4.7, the performance of the different models, when trained on *MIMICS-ClickExplore* and tested on the *MIMICS-Manual* dataset, varies substantially across models (in terms of both P@1 and MRR), while this is not the case for the models when trained and tested on *MIMICS-ClickExplore*. On *MIMICS-ClickExplore*, *RankBoost* and *Coordinate Ascent*

perform similarly, while there is a substantial difference between their performance on *MIMICS-Manual*. On the other hand, *RandomForests* performs better than *Coordinate Ascent* on *MIMICS-Manual*, which is not the case on *MIMICS-ClickExplore*. We carried out a t-test between the model effectiveness scores on the *ClickExplore* and *Manual* collections, respectively, with a threshold of $p < 0.05$ to determine significance. For *ClickExplore*, 11 of 15 pairwise tests show significant differences for P@1, and 8 of 15 pairwise tests show significant differences for MRR.

We observed the same situation when we trained the models on *MIMICS-Manual* and then evaluated them on the *MIMICS-ClickExplore* and *MIMICS-Manual*. For example, in Table 4.8, we can see *RankNet* performs relatively well when the test dataset is *MIMICS-Manual*, while it shows the poorest performance on *MIMICS-ClickExplore*. The number of significant pairwise differences also shows variations, with 2 of 6 for P@1 on *MIMICS-ClickExplore* and *MIMICS-Manual* and 2 versus 1 for MRR on *MIMICS-ClickExplore* and *MIMICS-Manual*, respectively, when the training dataset changed for a model. We can also observe that when we removed the queries that have the same *Overall Quality* label, the performance of the LTR models, when they were trained and tested on the *MIMICS-Manual*, dropped dramatically (comparing Table 4.8 with Table 4.4), confirming the hypothesis we had. Comparing Tables 4.7 and 4.8 also shows that the values of P@1 and MRR are in the same order, indicating the performance of the LTR models is independent of the training dataset.

Differences between using the online and offline datasets in this study are further highlighted when examining the weight of each feature learned by an LTR model from each dataset. We focus on *RankBoost* according to Table 4.7, which has produced the best overall performance. Similar observations hold for the top features selected by other models as well. Table 4.9 shows the top 10 features with the highest weights according to *RankBoost* when it was trained on the *MIMICS-ClickExplore* and *MIMICS-Manual*

Table 4.7: Performance of LTR models trained on *MIMICS-ClickExplore*. (Significance test results are explained in the text.)

| Model | P@1 | | MRR | |
|-------------------|-----------|--------|-----------|--------|
| | ClickExp. | Manual | ClickExp. | Manual |
| MART | 0.415 | 0.394 | 0.667 | 0.697 |
| RankNet | 0.417 | 0.409 | 0.668 | 0.705 |
| RankBoost | 0.444 | 0.606 | 0.683 | 0.803 |
| Coordinate Ascent | 0.444 | 0.424 | 0.683 | 0.712 |
| LambdaMART | 0.424 | 0.561 | 0.671 | 0.780 |
| RandomForests | 0.417 | 0.455 | 0.667 | 0.727 |
| Mean | 0.427 | 0.475 | 0.673 | 0.737 |
| Random Ranker | 0.301 | 0.330 | 0.540 | 0.600 |

Table 4.8: Performance of LTR models trained on *MIMICS-Manual*. (Significance test results are explained in the text.)

| Model | P@1 | | MRR | |
|-------------------|-----------|--------|-----------|--------|
| | ClickExp. | Manual | ClickExp. | Manual |
| MART | 0.425 | 0.485 | 0.673 | 0.742 |
| RankNet | 0.404 | 0.530 | 0.660 | 0.765 |
| RankBoost | 0.425 | 0.439 | 0.672 | 0.720 |
| Coordinate Ascent | 0.420 | 0.424 | 0.669 | 0.712 |
| LambdaMART | 0.411 | 0.439 | 0.664 | 0.720 |
| RandomForests | 0.417 | 0.561 | 0.667 | 0.780 |
| Mean | 0.417 | 0.480 | 0.667 | 0.740 |
| Random Ranker | 0.300 | 0.346 | 0.541 | 0.660 |

Table 4.9: The top 10 features with the highest weight learned by *RankBoost* from *MIMICS-ClickExplore* and *MIMICS-Manual*.

| Training on MIMICS-ClickExplore (online) | Training on MIMICS-Manual (offline) |
|---|---|
| BM25 (Query, Clarification Question + Option2) | CosinSimilarity_TF-IDF (Query, Clarification Question + Option1) |
| BM25 (Query, Clarification Question + Option1) | BM25 (Query, Clarification Question + Option4) |
| CosinSimilarity_TF-IDF (Query, Clarification Question + Option2) | BM25 (Query, Clarification Question + Option5) |
| CosinSimilarity_TF-IDF (Query, Clarification Question + Option1) | BM25 (Related Search, Clarification Question + Option5) |
| The clarification is a Statement or a Question | BM25 (Clarification Pane, Query) |
| BM25 (Clarification Pane, Query) | BM25 (Query, Clarification Question + Option3) |
| CosineSimilarity_TF-IDF (Clarification Pane, Query) | BM25 (Video Title, Clarification Question + Option5) |
| Overall Matching Terms (Clarification Pane, Query) | CosineSimilarity_TF-IDF (Video Title, Clarification Question + Option4) |
| Overall Matching Terms (Snippet, Clarification Question + Option1) | Overall Matching Terms (Related Search, Clarification Question + Option5) |
| BM25 (Video Description, Clarification Question + Option2) | BM25 (Video Title, Clarification Question + Option1) |

datasets, respectively. It is striking that only two out of ten features are shared across the datasets, suggesting that *RankBoost* learned substantially different ranking functions from *MIMICS-ClickExplore* and *MIMICS-Manual*, even though the available features were the same.

4.2.3 Discussion: Limitations of Existing Resources

Exploring the *MIMICS* dataset [125], the largest search clarification dataset that was collected from Bing log data, shows that less than 30% of CPs receive positive engagement, indicating the current practice of asking clarification questions from users to narrow down the search result is not yet successful. We showed in Chapter 2 that Sekulić et al. [95] aimed to predict the *Engagement Level* for the CP in *MIMICS-Click* using a BERT-based model and investigated the impact of SERP elements on the model performance. They formulated the task as a supervised regression. While their model outperformed some baselines in terms of predicting CP user engagement, it could not achieve any Coefficient of Determination (R^2)

better than 0.1124. They also observed that their ELBERT model, trained with different combinations of SERP elements, performed better when only query or query-pane settings were used. Later, Lotze et al. [71] followed the same approach to predict the *Engagement Level* and then coupled the model with *RankNet* to rank generated CPs for a given query in *MIMICS-ClickExplore*. The *RankNet* model they tested on *MIMICS-ClickExplore* achieved an MRR value of 0.620. However, our study using the *RankNet* model with all defined input features obtained an improved MRR value of 0.668 (Table 4.7). Comparing these two studies, we performed relatively better in ranking CPs for a given query. However, if we consider $P@1$ as an evaluation metric for identifying the MECP for a given query (putting the MECP in the first place), the LTR models with SERP information as input were unsuccessful, with low $P@1$ values around 0.420-0.500. This was observed regardless of training and testing on *MIMICS-ClickExplore* or *MIMICS-Manual* datasets. This highlights the need to gain more insights into the characteristics of the MECP to successfully identify it, which is crucial for enhancing search engine performance in meeting user information needs.

The different performances of the models when we changed the training dataset from *MIMICS-ClickExplore* to *MIMICS-Manual* unveil the impact of the evaluation strategy for the task of generating and selecting CPs. While the *MIMICS-ClickExplore* dataset is an online dataset (i.e., containing CTR information, an explicit signal of user engagement with a CP), *MIMICS-Manual* is an offline dataset as it contains the human annotation on the quality of the CPs. The different performances of the same LTR models with the same input features on online and offline datasets raise a red flag that the evaluation of clarification generation and ranking models needs to be carried out with caution.

To the best of our knowledge, *MIMICS* is the only data collection that provides both online and offline signals for evaluating search clarification; however, its limitations are a barrier to drawing any robust conclusion on the relationship between online and offline evaluations or advancing other investigations in search clarification. We ideally require queries that exist in both online and offline datasets. However, a close look at both datasets shows there were 204 query-clarification pairs collectively in the *MIMICS-Click* and *MIMICS-ClickExplore* datasets, including queries with single clarification panes, which we could find the same ones in the *MIMICS-Manual* dataset. Even in this small dataset, there are many queries that associated clarification panes had the same *Engagement Level*, which cannot be used for such comprehensive analyses. In total, there are only 106 query-clarification pairs shared between *MIMICS-Manual* and *MIMICS-ClickExplore*. One cannot draw conclusions based on an analysis of 106 query-clarification pairs with the limitation mentioned above, and thus, tasks related to search clarification suffer from a lack of available resources for both online and offline evaluations.

4.3 Summary

In this chapter, we focused on the task of identifying the most engaging clarification pane for a given query. We formulated this task as a learning-to-rank problem using several LTR models. We defined a series of features based on the query, SERP information and the clarification pane to feed the models. We trained and tested the models on two available clarification datasets of *MIMICS-ClickExplore* and *MIMICS-Manual* and evaluated the performance of the models in terms of how accurately the models can create the clarification ranked lists similar to the ideal ranked lists created based on the *Engagement Level*, a click-through rate signal, and the quality of CPs, labelled through manual annotation.

We showed that although *MIMICS* datasets, as the only available search clarification dataset to date, opened new opportunities to investigate the task of ranking multiple CPs for a given query in a search engine, the absence of an enriched dataset that is a collection of user interaction with the CPs in addition to the information about the characteristics of CPs have impeded further progress. Advancing the state of the art in generating, selecting, and presenting CQs is tightly coupled with developing effective evaluation methodologies and resources for their quantitative assessment. Examining existing public resources for search clarification demonstrated that they are not sufficient for a multi-dimensional evaluation of search clarification methods. The outcome of the experiments and the limitations of the available search clarification datasets motivate us to develop a new dataset that not only provides valuable information about the characteristics of the most engaging clarification panes from a user's point of view it can also be used for training and testing generating and selecting models in addition to study of online and offline evaluations in search clarification. This dataset is called *MIMICS-Duo* and will be introduced in the next chapter.

Chapter 5

Introducing MIMICS-Duo: A Dataset for Online and Offline Evaluation of Search Clarification

In Chapter 4, we showed that resources for training and evaluating search clarification methods are not yet sufficient. This motivated us to develop a new multi-dimensional search clarification dataset. We introduce *MIMICS-Duo*¹, a new freely available dataset of 306 search queries with multiple clarifications (a total of 1,034 query-clarification pairs) that are sampled from *MIMICS ClickExplore* [125]. *MIMICS-Duo* contains online signals, such as user engagement based on click-through rate (CTR) and fine-grained annotations on clarification questions (CQs) and their candidate answers that enhance the existing *MIMICS* datasets by enabling multi-dimensional evaluation of search clarification methods.

MIMICS-Duo can be used for training and evaluating many search clarification tasks: generating CQs; ranking clarification panes (CPs); re-ranking candidate answers; unbiased click models and user engagement prediction for clarification; and analyzing user interaction with search clarification. This newly introduced dataset also helps us establish the relationships between different aspects of clarification panes, which can be used for further improvement of generating and asking clarification models.

¹MIMICS-Duo is available at <https://github.com/Leila-Ta/MIMICS-Duo>

5.1 Data Sampling from MIMICS-ClickExplore

We use the *MIMICS-ClickExplore* dataset that contains the corresponding aggregated user interaction signals (i.e., *Impression Level*, *Engagement Level*, and conditional click probability) for queries with multiple CPs. As the first selection criterion, we discard the queries that had two CPs, as they are not good candidates for ranking CPs and are not helpful for establishing any relationship between online and offline evaluations in search clarification. The query length (number of words in each query) in this dataset varied between 1 and 9.

To create a new diverse search clarification dataset, we divide the queries and related CPs into nine sub-classes based on the query length. Next, we subdivide all queries in each bin of query length based on the highest *Engagement Level* obtained by one of the associated CPs. After the data pre-processing described in Chapter 4, every query has one CP that has the highest *Engagement Level* compared to other panes in the set, and this highest level varies between 1 to 10 (e.g., if the highest *Engagement Level* of a CP is one, then the *Engagement Level* of others for a given query is zero).

Finally, we create the *MIMICS-Duo* dataset that contains almost 11% from each query length bin and 10% from each *Engagement Level* bin, depending on availability. Also, wherever possible, we select query-clarification pairs that have different *Impression Levels*. This process leads to a collection of 306 queries with at least three CPs (1,034 query-clarification pairs) that have diversity in query length and *Engagement Level*. This dataset has the same format as the *MIMICS* dataset for simplicity in any analyses and comparisons in the future. The statistics of *MIMICS-Duo* dataset are presented in Table 5.1. In order to have a representative dataset, we attempt to select queries with the highest diversity in terms of *Engagement Level*, *Impression Level*, options and the number of options in their CPs.

Table 5.1: Statistics of MIMICS-Duo dataset.

| | |
|---------------------------------------|-----------|
| Number of queries | 306 |
| Number of query-clarification pair | 1,034 |
| Number of clarifications per query | 3.38±0.68 |
| Min & max clarifications per query | 3 & 8 |
| Number of candidate answers | 3.59±1.2 |
| Min & max number of candidate answers | 2 & 5 |

5.2 Task Design

To create *MIMICS-Duo* that overcomes the shortcomings of the current search clarification datasets, we conduct online experiments² through Human Intelligence Tasks (HIT) on Amazon Mechanical Turk³ (AMT) and Qualtrics⁴ to gather labels.

We designed three tasks to collect judgements from AMT workers on CPs related to queries and search engine results pages. We then analyse the correlation between collected labels and the *Engagement Level* of CQs and the click-through rate of candidate answers. The tasks are designed to capture overall CP preference and their quality and characteristics. Figure 5.1 shows an overview of the three tasks in this study.

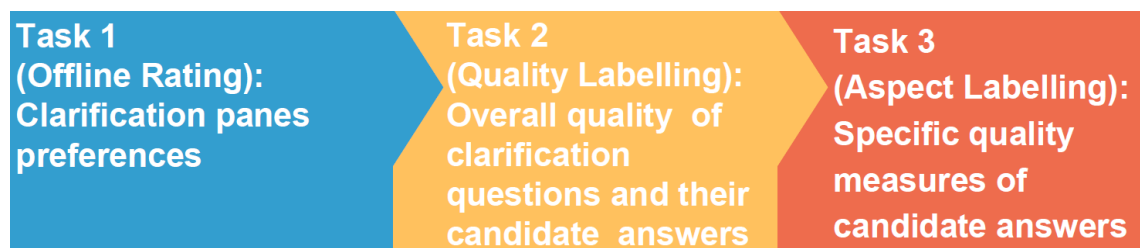


Figure 5.1: An overview of the three steps of the data collection.

Since the entire process is conducted online, it is necessary to prepare the instruction for each task in plain English, which is fully digestible for any worker with any level of education, and avoid academic and, in particular, information retrieval terms. We provide the required information about the survey’s aim, steps that need to be taken, and the number of questions.

AMT workers are redirected to Qualtrics to complete the tasks. This is to ensure we create a professional and user-friendly interface for the tasks. Each task has five components (i) the informed consent, including the IRB approval number and participant information sheet, (ii) the instruction, (iii) the survey body (i.e., the task itself), (iv) a feedback page, and (v) a completion code generator on the last page.

Task 1 (Offline Rating): Clarification panes preferences.

We provide workers with a query and its top eight retrieved document summaries provided in *MIMICS-ClickExplore*. Workers are then presented with multiple (varied between 3 to 8 depending on the query) CPs for that query and asked to rate them using a 5-star rating.

²Reviewed and approved according to Anonymous University IRB procedures for research involving human subjects. The IRB approval number is 66-19/22334.

³<https://www.mturk.com>

⁴<https://www.qualtrics.com>

We aim to simulate online user clicks in our task by showing all generated CPs for a given query at once. So the workers could rate them based on their preferences.

This Task also includes an attention check: before the workers are asked to rate the CPs, we show them all CPs and ask them to write down the number of CPs that have been generated for the given query. If a worker gives an invalid answer, the HIT is rejected, and the worker is blocked from completing further HITs. In total, 306 queries with multiple CPs (306 HITs) were launched on AMT.

Task 2 (Quality Labelling): Overall quality of clarification pane (i.e., CQs and their candidate answers).

Workers are shown a query and its top eight retrieved document summaries provided in the *MIMICS-ClickExplore* dataset. A single CP is shown to workers, and they are asked to rate the *overall quality* of that CP, as well as its individual candidate answers. This task is analogous to the ratings made in the *MIMICS-Manual* dataset; however, recall that the overlap of those queries with *MIMICS-ClickExplore* was insufficient to enable meaningful exploration of the relationship between user engagement and the quality of CPs. This task is designed to overcome the limitations of the *MIMICS* dataset. The overlap between *MIMICS-Manual* with *MIMICS-Click* and *MIMICS-ClickExplore* datasets was small (only 201 query-clarification pairs, including queries with tied CPs), not large and diverse enough to establish any knowledge about the relationship between the user engagement and the quality of CPs. The manual annotation is similar to *MIMICS-Manual* to collect the quality labels for CPs, but a CQ or a candidate answer is assessed on a 5-level rating scale (1 (very bad), 2 (bad), 3 (fair), 4 (good), and 5 (very good)) compared to a 3-level rating in the *MIMICS* dataset. A wider range of labelling helps get more accurate labels for CPs.

Similar to Task 1, this task also has two attention check questions to ensure a high level of quality for each HIT. In total, 1,034 query-clarification pairs (1,034 HITs) are launched on AMT.

Task 3 (Aspect Labelling): Specific quality measures of clarification panes.

Workers are again provided with a query and its top eight retrieved document summaries, together with a single CP and asked to judge the *Coverage*, *Diversity*, *Understandably*, and *Candidate Answer Order* of the CP. While the *MIMICS-ClickExplore* dataset showed that all CPs generated for a given query did not receive the same user *Engagement Level*, it doesn't provide information to explore the characteristics that may lead to these differences. To address this critical question, *Aspect Labelling* is carried out in this research. Since

the *Bing* search engine generated CPs in a multi-choice question format and not in full sentences, the aim is to investigate whether the CP is understandable and whether the presented candidate answers are diverse enough and cover all possible query intents or not. Also, the importance of the order of the candidate answers is evaluated. The overall goal is to gather data about key characteristics of CPs, support research into their relationship with *Engagement Level*, and be able to generate more engaging CPs. The findings can also support the re-ranking of CPs. Workers rate each aspect on a 5-level scale: strongly disagree, somewhat disagree, neither agree nor disagree, somewhat agree and strongly agree to answer the following questions, based on seeing the query and the top eight retrieved documents:

1. Does the clarification pane have a high coverage for the given query?
2. Does the clarification pane have a high diversity for the given query?
3. Is the clarification pane understandable for the given query?
4. Does the clarification pane have the correct order for the given query?

To ensure that the workers understand the definition of each aspect, descriptive examples for each question, showing CPs with high and low coverage, high and low diversity, understandable and non-understandable CPs, and with and without correct orders⁵ are presented to the workers. This task also has two gold questions to keep the quality of the collected data as high as possible. In total, 1,034 query-clarification pairs (1,034 HITs) are launched on AMT.

5.3 Pilot Tasks

We launched two series of AMT pilot surveys, containing nine HITs for Task *Offline Rating*, 32 HITs for Tasks *Quality Labelling*, and 32 HITs for Task *Aspect Labelling*. These pilots enabled us to analyse the flow of the tasks, estimate the required time to finish each task, collect the workers' feedback, check the quality of collected data, and revise the tasks if needed. For instance, we optimised the layout, task examples, and attention check questions (gold questions) with the aim of high validity throughout the tasks, which led to a high success rate of 89%, 91%, and 100% for *Offline Rating*, *Quality Labelling* and *Aspect Labelling*, respectively, at the end of the second pilot survey. The feedback obtained from workers during both pilot runs and the main surveys confirmed that the task and examples were clear enough to make the justification easy for them.

⁵The full instructions and examples presented to participants are available at <https://github.com/Leila-Ta/MIMICS-Duo>

5.4 Quality Assurance and Attention Measures

Four quality assurance and attention measures are embedded in each task. First, to ensure workers pay attention to the different aspects of the query and the document summaries, we show eight relevant summaries and one irrelevant summary. Workers are then asked to identify the irrelevant document summary, which is placed in a random rank position for each task. This first check provides both an attention measure (i.e., workers are forced to inspect all summaries) and a gold question (i.e., a question with a pre-defined answer). Second, we randomly insert a second gold question from a pool of 15 pre-defined questions (e.g., *What is 2+2? Please choose five from the choices below.*). Third, we incorporate a robot detection step (CAPTCHA⁶) in each task. Lastly, workers are provided with a randomly generated code at the end of the task, which they are asked to submit to AMT as a final quality check. The answers of workers who do not pass these gold questions are removed and are not included in the final dataset. Furthermore, workers who fail the checks are also blocked from completing further tasks.

We perform regular quality checks throughout the data collection process to ensure high-quality data, and after collecting the data, we manually check 10% of submitted HITs per task as a final quality assurance check. If we observe any invalid submissions, we remove those submissions, prevent the workers from completing subsequent tasks, and open the HITs to the different workers. The rigorous task design and continuous quality checks of submitted HITs help us collect high-quality labels.

5.5 Crowd-sourcing

This study was carried out using the AMT crowd-sourcing platform between 27 January 2022 and 16 February 2022. Workers with the following qualifications were able to participate in the study:

- Only participants located in Australia, Canada, Ireland, New Zealand, the United Kingdom and the United States with a HIT approval rate of 95% or higher and a minimum of 5,000 previously approved HITs were allowed to participate in order to maximise the survey success rate and the likelihood that users were native English speakers or had a high level of English.

⁶CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a type of security measure known as challenge-response authentication.

- Users could only participate once in each task.
- All three tasks were launched at different times and days to maximise the diversity of the participants.
- Based on experience from the pilot tasks, the hits conducted by participants who took less than 90 seconds to complete the full task were labelled as low quality and removed from the dataset, and the workers were not eligible for future tasks.

Each HIT was assigned to at least three different AMT workers. Depending on the task, the workers were paid 0.45, 0.72 and 0.95 USD per HIT for *Offline Rating*, *Quality Labelling* and *Aspect Labelling*, respectively. The collection of this dataset cost 9,880 USD. For each labelling task, we used majority voting to aggregate the annotation. In case of disagreements, the HIT was opened again to more workers until a final majority vote label could be assigned. The mean agreement was 73.44%, 74.36% and 76.63% for *Offline Rating*, *Quality Labelling* and *Aspect Labelling*.

5.6 AMT Workers' Feedback

We collected the workers' feedback as mentioned in Section 5.2 to improve the quality of each task for the next rounds and also to understand if they had any comments in general about the query and CPs we showed them. To do so, at the end of each survey, before they submit their hits, we asked them to provide their general feedback about the hit and also if they have any comments about the query, related documents and CPs.

We understand from workers' feedback that they were mainly satisfied with the hits. They found the hits well-designed, understandable, engaging, educating and very clear in terms of the instruction, examples, and aim with a fair pay rate yet challenging. However, every time we collected the data, there was some feedback to improve the quality of the tasks. For example, some workers asked to add more details on the instruction page, such as the time the survey will take or adding a progress bar to the surveys.

Apart from their general feedback, many workers provided their opinions about the concept of asking CQs in search engines. They believed that it could be a helpful feature, although some workers thought it was only useful for certain queries. There were also some workers who thought the CP generated for the given query did not cover all potential intents, even with three to five candidate answers and didn't provide a broad enough selection of responses. They asked for more options! For example, there are many brands for a printer, and when the CQ asks the user to specify the brand she is looking for, there is no way to cover all brands in only five candidate answers. One way to tackle this challenge is to

show more than one CP to the user, or in some cases where it is appropriate, one candidate answer could show “other,” which sends the users to other brands that are not included in that CP. Another feedback was about the specificity of the CQ. The CP should not lead to answers that are too general to be useful. We also need to generate a smart CP. For instance, when a user asks for information about a 60-inch TV, the user most probably wants to compare brands before looking at a specific brand; otherwise, she can simply add the brand she had in mind to the query. We also noticed that the order of candidate answers is very subjective, and sometimes they are equally important, and the workers faced challenges to justify whether the candidate answers have a correct order or not.

5.6.1 MIMICS-Duo Dataset Analysis

Quality Labelling

The distribution of the quality labels for clarification panes (overall quality of clarification questions and their answers) and the quality labels of the individual candidate answers are shown in Table 5.2. The number label assigned to each candidate answer is an index of its position within the clarification pane, counting from left to right. The results show that around 77% of clarification panes had *Good* or *Very Good* ratings. This means the majority of generated clarification panes for the given queries were relevant and satisfactory. We can also understand that although the quality of the majority of candidate answers was *Good* or *Very Good*, the mean quality rating of the candidate answers decreases from left to right across the clarification panes (i.e., with an increase in the position index of candidate answers).

To investigate the impact of the quality of candidate answers on the overall quality of clarification panes, we calculated the mean value of quality labels given to candidate answers of a clarification pane by the workers for every clarification pane. We found out that there was a strong correlation between the quality of candidate answers and the overall quality of clarification panes regardless of the number of candidate answers ($r=0.708$).

The distribution of the overall quality of clarification panes is shown for every engagement level bin (0 to 10) in Figure 5.2. We can see that regardless of the engagement level, almost 75% of clarification panes had *Good* or *Very Good* overall quality, and more than 96% of clarification panes had *Fair* or a higher quality label. This is a signal that a simple CTR as an indicator of user interaction with the clarification pane is not a strong metric to evaluate the performance of generating or asking clarification questions in search engines. This figure also indicates that generating high-quality clarification panes does not necessarily lead to more user engagement. Users showed that they could sometimes be reluctant to get engaged with high-quality clarifications, and they may also be engaged

Table 5.2: Distribution of the quality label of clarification panes and their candidate answers.

| Criterion | Statistics | | Labels ¹ (%) | | | | |
|--------------------|------------|------------|-------------------------|------|-------|-------|-------|
| | μ | σ^2 | 1 | 2 | 3 | 4 | 5 |
| Clarification Pane | 3.95 | 0.58 | 0.39 | 3.19 | 19.44 | 54.55 | 22.44 |
| Candidate Ans. #1 | 4.12 | 0.83 | 1.16 | 3.48 | 19.05 | 35.11 | 41.20 |
| Candidate Ans. #2 | 4.01 | 0.81 | 0.77 | 5.13 | 19.92 | 40.33 | 33.85 |
| Candidate Ans. #3 | 3.93 | 0.84 | 0.78 | 5.09 | 25.59 | 37.60 | 30.94 |
| Candidate Ans. #4 | 3.88 | 0.9 | 1.33 | 4.75 | 29.47 | 33.46 | 30.99 |
| Candidate Ans. #5 | 3.89 | 0.94 | 1.15 | 7.45 | 24.07 | 36.39 | 30.95 |

¹ Label meaning: 1 (Very Bad), 2 (Bad), 3 (Fair), 4 (Good), 5 (Very Good).

with poor-quality ones. Therefore, it appears that click-through information can be noisy and biased and does not necessarily reflect the user's perception of information quality and, therefore, needs to be used carefully alongside other evaluation methods.

We also compared the quality labels of candidate answers with the click-through rate probability of candidate answers. While no correlation was found between offline answer labelling and online interaction ($\rho=0.032$), if we ranked the candidate answers based on their quality labels and click-through rate probability (ideal ranking), P@1 and MRR were calculated at 0.338 and 0.597, respectively.

Aspect Labelling

To support the investigation of the relationship between the characteristics of clarification panes and engagement level, overall quality and offline rating of clarification panes, we carried out the *Aspect Labelling* task. Four aspects – *Coverage*, *Diversity*, *Understandability* and *Candidate Answer Order* – were evaluated. Table 5.3 shows the distribution of characteristic labels of clarification panes. It is evident that apart from the *Candidate Answer Order*, the majority of clarification panes had high *Coverage*, *Diversity* and *Understandability*, with the trend being strongest for *Understandability*. More than 40 per cent of AMT workers chose the “neither agree nor disagree” response with respect to the candidate answer order aspect: here, they were asked to rate whether the candidate answers for a given query were in the correct order or not (i.e. in importance order, from left to right). It appears that workers were mostly undecided regarding this aspect.

In another analysis, we classified the clarification panes into five categories based on their overall quality labels (Very Bad, Bad, Fair, Good and Very Good) and investigated the contribution of each aspect to the overall quality by calculating the mean value for each aspect in each category, shown in Figure 5.3. We can see the more a clarification pane had higher *Coverage* and was more *Understandable*, the higher overall quality was achieved. We can also see that *Diversity* had second place as the influential factor and *Candidate Answer Order*, as mentioned, had no clear impact.

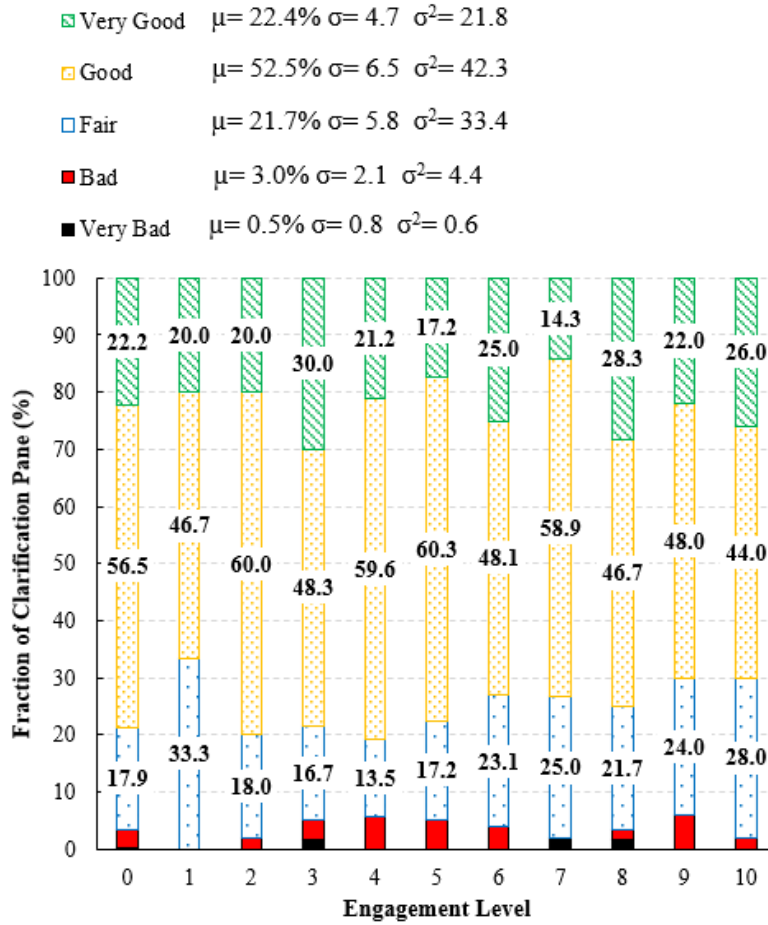


Figure 5.2: Quality Label vs. Engagement Level.

The correlations between all online and offline annotations are shown in Table 5.4. It is evident that the engagement level collected in MIMICS-ClickExplore (online evaluation) had no correlation with any offline measure, while different correlations can be easily found between offline measures. For example, there is a medium correlation between coverage and diversity, as expected, and overall quality or offline ranking has a higher correlation with *Coverage* compared to *Diversity* and *Understandability*. While the correlation between candidate answer order and other offline measures is very weak, the number of candidate answers also has a higher correlation with coverage and diversity compared to no correlation with understandability. This is expected as a multi-choice clarification pane can only get high coverage or diversity when the number of candidate answers is high.

5.7 Research Enabled by MIMICS-Duo

In this section, we introduce the research problems in search clarification that can be addressed using the new MIMICS-Duo dataset:

Table 5.3: Distribution of the characteristics label of clarification panes.

| Criterion | Statistics | | Labels ¹ (%) | | | | |
|-----------------|------------|------------|-------------------------|-------|-------|-------|-------|
| | μ | σ^2 | 1 | 2 | 3 | 4 | 5 |
| Coverage | 3.78 | 1.18 | 3.00 | 14.02 | 12.19 | 43.23 | 27.56 |
| Diversity | 3.74 | 1.15 | 1.45 | 16.73 | 15.09 | 40.14 | 26.60 |
| Understand. | 4.61 | 0.53 | 0.39 | 2.13 | 6.09 | 18.67 | 72.73 |
| Can. Ans. Order | 3.43 | 0.87 | 1.55 | 12.86 | 40.23 | 31.62 | 13.73 |

¹ Label meaning: 1 (Strongly disagree), 2 (Somewhat disagree), 3 (Neither agree nor disagree), 4 (Somewhat agree), 5 (Strongly agree).

**Figure 5.3:** Mean values of different aspect labels for clarification panes with various overall quality.

Offline and Online Evaluation: A key research task in search clarification is generating and asking CQs in information-seeking problems. Since MIMICS-Duo has a large query overlap with MIMICS-ClickExplore, it enables researchers and practitioners to conduct a detailed analysis of clarification selection and generation models from both online (real users) and offline (annotators) perspectives. Therefore, MIMICS-Duo complements the existing datasets for search clarification and will significantly impact the progress in this area of research.

User Engagement and Clarification Quality: The manual labelling of CPs includes information about the *coverage*, *diversity*, *understandability* of CPs and the *importance order* of candidate answers. This information helps the researchers to study the characteristics of CPs that impact user engagement. Moreover, Since the user *Engagement Level* is also

Table 5.4: Correlations between online and offline measures. (*Cov, Div, Und, IO, OQ, EL, OR* and *#Ans.* stand for *Coverage, Diversity, Understandability, Importance Order, Engagement Level, Offline Rating* and *Number of Candidate Answers*, respectively.)

| | Cov | Div | Und | IO | OQ | EL | OR | #Ans. |
|-----------|-----|-------|-------|-------|-------|--------|--------|--------|
| Cov | NA | 0.421 | 0.313 | 0.178 | 0.227 | -0.061 | 0.273 | 0.306 |
| Div | | NA | 0.260 | 0.117 | 0.176 | -0.029 | 0.245 | 0.269 |
| Und | | | NA | 0.159 | 0.226 | 0.034 | 0.227 | 0.055 |
| IO | | | | NA | 0.064 | 0.003 | 0.044 | -0.178 |
| OQ | | | | | NA | -0.032 | 0.225 | 0.165 |
| EL | | | | | | NA | -0.001 | -0.079 |
| OR | | | | | | | NA | 0.262 |
| # of Ans. | | | | | | | | NA |

available in the dataset, the analyses can be expanded to more CP characteristics such as coherency, comprehensiveness, relevance and usefulness, which can add more value to this dataset and, in general in the field of search clarification.

Clarification Click Models: MIMICS-Duo contains several query-clarification pairs for a given query whose only differences are in the order of candidate answers. This information, in addition to manual annotation about the importance order of candidate answers, enables further study on training and evaluating click models for answer ranking in search clarification.

5.8 Summary

Comparing online and offline evaluation is an understudied area, including in the context of search clarification. However, available search clarification datasets are either created using online user interaction signals (click-through rate) or manual annotation of quality, and there is no dataset that covers both sides. This motivated us to create the *MIMICS-Duo* dataset to help bridge the gap between available search clarification datasets.

MIMICS-Duo is a search clarification data collection containing both online and offline evaluations that are designed to work with the existing *MIMICS-ClickExplore* dataset. It contains 306 unique queries with multiple CPs (1,034 query-clarification pairs) with interactions of real users, collected from the *Bing* search logs and graded quality labels including multiple CPs rating, overall quality labelling for CPs and their individual candidate answers and labels for different aspects of CPs. This dataset was created through fine-tuned crowd-sourcing, and extensive quality assurance and attention measures were considered to ensure the accuracy of the collected labels.

This dataset enables us to analyse the relationship between the online and offline eval-

uations in search clarification, which will be discussed in the next chapter. Although *MIMICS-Duo* does not compare offline evaluation with live online experimentation (e.g., real-time A/B testing with users engaging with a live system), it provides a unique opportunity for researchers to evaluate any search clarification task using offline evaluations and compare it with online signals, which was not possible before.

In Chapter 6, we will show how this dataset overcomes the shortcomings of the *MIMICS* dataset. We will further analyse the relationships between online and offline evaluations in search clarifications and provide some recommendations for generating and selecting clarification questions in search engines.

Chapter 6

Online and Offline Evaluation in Search Clarification

In the previous chapter, we introduced a new search clarification dataset that overcomes the limitations of existing datasets. In this chapter, we delve deeper into the evaluation approaches of search clarification in search engines, exploring the relationship between online and offline evaluation methods. Specifically, we investigate how our collected clarification pane quality labels in *MIMICS-Duo* help us to perform better in ranking clarification panes and identifying the most engaging clarification pane (MECP) from a user’s perspective. Our research unveils that when it comes to search clarification and identifying the most engaging clarification panes, online and offline evaluations harmoniously converge although the alignment between online and offline evaluations in retrieval quality has long been debated. We show that the query length does not influence the relationship between online and offline evaluations, and reducing uncertainty in online evaluation strengthens this relationship. We illustrate that an engaging clarification needs to excel from multiple perspectives, and SERP quality and characteristics of the clarification are equally important. We also investigate if human labels can enhance the performance of Large Language Models (LLMs) and Learning-to-Rank (LTR) models in identifying the most engaging clarification questions from the user’s perspective by incorporating offline evaluations as input features. Our results indicate that Learning-to-Rank models do not perform better than individual offline labels. However, GPT (generative pre-trained transformer) emerges as the standout performer, surpassing all Learning-to-Rank models and offline labels.

6.1 Introduction: Current Practice and Knowledge Gap

In a practical situation, when a user submits a search query on a search engine like *Bing*, in addition to the results page, the search engine often presents a multi-choice clarification pane. This pane can be helpful when the retrieved documents do not fully address the user's information needs. Although multiple clarification panes can be generated for a single query, only one clarification pane is typically presented to the user at a time. Despite the advancements in generating clarification questions in search systems, the success rate of users engaging with those clarification questions remains low [125]. An analysis of the largest search clarification dataset, *MIMICS* [125], demonstrates that users tend to engage more with certain clarification panes than others for a particular query. Moreover, a high percentage of clarifications are left unengaged, regardless of how many times they are presented to users. This indicates that users are not easily engaged with clarification panes, and clarifications are not equally engaging from users' perspectives. These observations raise questions about the overall impact of the models. To improve the performance of clarification models in engaging users in search systems, proper evaluation approaches must be used to consider user behaviour and the characteristics of engaging clarification questions. We aim to understand what makes a clarification question engaging to enhance the clarification model performance and improve user engagement. This can be done by investigating the relationship between user engagement (online evaluation) and the characteristics of clarifications that are manually evaluated (offline evaluation). Effectively evaluating the quality of clarification panes is essential for several reasons, including:

- **Enhancing search accuracy:** Effective evaluation of clarification questions ensures that the generated prompts or queries effectively address user needs, leading to more precise and tailored search results.
- **Improving user satisfaction:** Well-designed clarification questions facilitate a more interactive and personalised search experience, enabling users to express their information needs more precisely and receive more relevant results.
- **Reducing search iteration:** Effective evaluation of clarification questions helps in reducing the need for multiple search iterations. By proactively suggesting relevant clarification prompts or refining user queries, search systems can improve the efficiency of information retrieval, saving users' time and effort.
- **Supporting diverse user intents:** By analysing and understanding user queries and their corresponding clarification prompts, search systems can adapt and provide

tailored responses based on the specific needs of each user, thus accommodating a wide range of user intents and preferences.

- Advancing conversational search: In conversational search scenarios, where users interact with search systems through natural language, the effectiveness of clarification questions directly impacts the quality of the dialogue and the system’s ability to understand user intent accurately.

The typical evaluation process in deploying new models in search engines involves (1) *offline evaluation* with labelled test collections and (2) *online evaluation* through user interactions, often using A/B testing. Having a dependable offline evaluation dataset facilitates iterative research and the development of models and features. Researchers typically conduct online experiments based on the results obtained from the offline evaluation. However, the relationship between offline and online evaluations in search clarifications is relatively unexplored. To address this knowledge gap, this chapter investigates a fundamental question on *how insights gained from offline evaluation align with the real-world performance of models during online experiments*. A comprehensive examination of the correlation between offline and online evaluations will enable researchers to make informed decisions about the effectiveness and reliability of their search engine models, ultimately leading to improved search experiences for users.

We focus on clarification panes, each consisting of a clarification question and up to five candidate answers. We explore the relationship between online and offline evaluation in search clarification by studying the following three primary research questions:

- What are the best overall practices in designing offline evaluation methodologies for search clarification that correspond with online evaluation?

We initially focus on evaluating the effectiveness of an oracle¹ clarification selection model. This model has access to every offline label and is compared to the online label to assess its performance. We move beyond assuming the independence of offline labels and delve into their combination, leveraging techniques such as LTR models to determine if these combinations align with online evaluation. Additionally, we use an LLM to predict online user engagement with the clarification, considering the provided offline labels. Motivated by Zamani et al. [127], who showed user behaviour is different in short queries (often keyword queries) and long queries (often natural language questions), we extend our study to ask:

¹In the context of machine learning, an Oracle typically refers to an idealised entity or concept that provides perfect information or answers to a given problem. It is often used as a theoretical reference point to establish performance bounds or to measure the efficiency and effectiveness of an algorithm.

- Does query length impact the relationship between online and offline evaluations in search clarification?

Here, we split our dataset into short and long queries. We observe differences when investigating the relationship between online and offline evaluations for short and long queries. We also explore the impact of uncertainty in the collected online evaluation on how well online and offline evaluations correspond to each other by studying the following question:

- Does uncertainty in the online evaluation impact the relationship between online and offline evaluation?

Here, we control uncertainty in the online labelling based on the number of times a clarification question is presented to users in A/B testing, often called *impression count*. The higher the impression count, the more reliable (thus less uncertain) online labels based on click-through rate are. We observe that reducing uncertainty in online evaluation strengthens the relationship between online and offline evaluations.

In contrast to the widely held notion that online and offline evaluations do not always coincide regarding retrieval quality [27, 38, 33, 38, 94], our study shows that offline evaluations align with online evaluations in search clarification. However, certain essential factors should be considered. This study also enhances our comprehension of the performance of LLMs in predicting online user engagement with clarifications when offline labels are employed as input for the models. The insights gained from our investigation will aid in refining the evaluation methodology for search clarification, leading to more efficient decision-making in deploying clarification models.

6.2 Methodology

6.2.1 Dataset

We compare online and offline evaluation in search clarification using the *MIMICS-Duo* dataset that we introduced in Chapter 5. It is a search clarification data collection containing online and offline evaluations for 306 queries with multiple clarification panes (1,034 query-clarification pairs). Here again, we briefly explain online and offline labels that are used in this chapter.

Online Labels

According to the definition provided with the *MIMICS* dataset and stated in Section 4.2, the *Engagement Level* is constructed based on the click-through rate of real user interactions

with clarification panes in Bing [125]. An equal-depth method was used for *Engagement Level*, dividing all the positive click-through rates into ten bins. Hence, the *Engagement Level* is an integer between 1 to 10 presenting the level of total engagement received by users in terms of click-through rate. Moreover, an *Engagement Level* of 0 was assigned to clarification panes with no clicks). The *MIMICS* dataset also provides an *Impression Level* computed on the number of times the given query-clarification pair was presented to users. Every query-clarification pair in the dataset was shown at least ten times to search engine users. The *Impression Level* has three quality values (low, medium, and high) and correlates with the query frequency. This online label is used in this study to group the clarification panes for the experiments (Subsection 6.3.3).

Offline Labels

Offline labels in the *MIMICS-Duo* dataset include a series of crowd-sourcing labels consisting of (1) Offline Rating, (2) Quality Labelling, and (3) Aspect Labelling.

The *Offline Rating* was collected based on crowd-sourced worker preferences. Workers were simultaneously shown all generated clarification panes (varied between three to eight depending on the query) for a given query. They were asked to rate the clarification panes using a 5-point rating (five means highest preference, and one means lowest preference). The nature of this label is different from other labels. For this label, all clarification panes for a given query were relatively rated with respect to each other, while for the rest of the labelling, workers were shown one clarification pane and asked to annotate only one characteristic of the clarification pane in isolation.

The *Quality Labelling* consists of two quality measures, the *Overall Quality* of the full clarification panes and *Option Quality*, that is, the quality of individual options (clarification pane candidate answers). Crowd-source workers rated the clarification panes and the quality of their options with a 5-point rating (five means very good quality, and one means very bad quality). *Aspect Labelling* consists of four sub-labels, that is, *Coverage* (i.e., the extent to which the clarification pane covers every potential aspect of the query), *Diversity* (i.e., the extent to which the clarification pane does not contain redundant information), *Understandability* (i.e., the extent to which the clarification pane is digestible and meaningful), and *Importance Order* (i.e., the extent to which the most relevant and important candidate answers are positioned from left to right). Workers were asked to label a clarification pane for these aspects through a 5-point rating (e.g., five means the worker strongly agreed that the clarification pane had high coverage, and one means the worker strongly disagreed that the clarification pane had a high coverage).

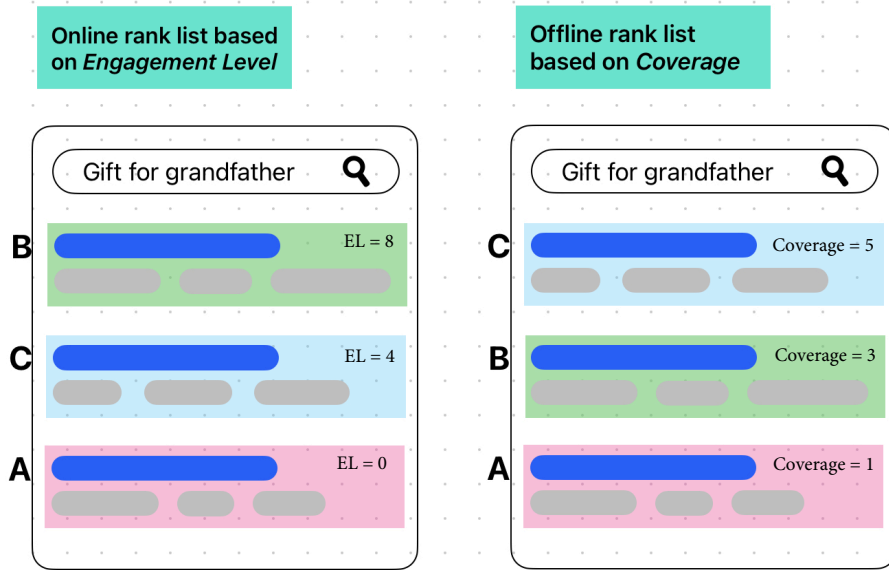


Figure 6.1: Two examples of clarification pane rank lists based on the *Engagement Level* and *Coverage* labels for a query.

6.2.2 Experimental Design

We showed that each clarification pane has two types of labels, online and offline. We use one online label (i.e., *Engagement Level*) and five offline labels (i.e., *Offline Rating*, *Overall Quality*, *Coverage*, *Diversity*, and *Importance Order*).

In the *MIMICS-Duo* dataset, *Overall Quality* and *Option Quality* labels have a very high correlation. This is understandable as the clarification question in more than 95% of the clarification panes in the dataset is the general question of “*Select one to refine your search*”. Therefore, the overall quality of a clarification pane is mainly based on the quality of its options. Hence, this study only focuses on *Overall Quality*. We also do not investigate the *Understandability* label in this study. The mean value of *Understandability* across the *MIMICS-Duo* dataset is 4.6 (out of 5), showing that more than 90% of the workers agreed that the clarification panes were highly understandable. Therefore, this characteristic has a minor impact on our evaluations.

Figure 6.1 shows an example of ranking three clarification panes [A, B, C] for a given query if the corresponding *Engagement Levels* and the *Coverage* labels are [0, 8, 4] and [1, 3, 5], respectively. We can see from this example that the ranked list generated based on the offline label was not completely successful in replicating the ideal ranked list, except for the clarification pane A.

6.2.3 Data Analysis

We designed the experiments in this study to answer the research questions as follows:

Overall Practices in Online and Offline Evaluations for Search Clarification

The main aim of this research is to compare the clarification ranked lists created using offline labels with the ideal clarification ranked lists created using the *Engagement Level*, in general, and to compare the MECPs, in particular. First, we investigate the relationship between online and offline labels on all 306 queries in the *MIMICS-Duo* dataset without applying any filtering or grouping on the dataset. In the next step, we investigate if collected offline labels can be used as input features in learning-to-rank (LTR) models to create ranked lists more similar to ideal ranked lists created based on the *Engagement Level*. We use four offline labels of *Overall Quality*, *Coverage*, *Diversity*, and *Importance Order*, as well as the number of candidate answers in each clarification pane as input features in the models. The features are linearly normalised based on their min/max values. We do not use the *Offline Rating* label considering its different nature. While other labels offer insights into various aspects of clarification panes, this label is based on the relative rating of all clarification panes of a given query. We employ four LTR models, including *Mart*, *RandomForests*, *RankBoost*, *AdaRank* that are implemented in *RankLib* [28]. We also utilise *SVM-rank* [50]² with a linear kernel. We use 5-fold cross-validation to evaluate our models. In each fold, the dataset is split into training and testing sets by the ratio of 4:1.

Finally, we exploit the capabilities of an LLM (GPT-3.5), an advanced language model, to forecast online user engagement with the clarification panes. We use *GPT-3.5-turbo* model.³ The task assigned to GPT-3.5 is to predict the *Engagement Level* within a range of 0 to 10. The prompt that we use to feed the GPT model contains (1) a *query*, (2) a *clarification pane* that includes *Clarification Question* and associated *Options* (*Candidate Answers*) and (3) four *offline labels* similar to LTR models.

In our initial experiments, we explored various prompts that focused on the same task. We observed that when the prompt lacked sufficient detail, there were instances where GPT-3.5 either failed to provide the *Engagement Level* or provided it in a quantitative format instead of a range from 0 to 10. The most successful prompt template utilised in this study is shown in Figure 6.2.⁴ We prompt the model to generate an *Engagement Level* for 1,034 query-clarification pair. We conduct experiments using various temperature settings, specifically, $temp = \{0.0, 0.5, 1.0\}$. The temperature parameter regulates the degree of randomness in the generated text. During text generation, the model generates a probability distribution over the next word or token, and the temperature parameter influences the shape

²https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

³Last accessed on the 29th of May 2023.

⁴The prompt template used in this study, along with other versions of prompts, is publicly accessible at https://github.com/Leila-Ta/On_Off-Eval-Search-Clarification

of this distribution. A higher temperature value, such as 1.0, results in a more uniform distribution and increases the randomness in the generated output. This can lead to a wider range of diverse and creative responses but may also introduce more errors or nonsensical text. On the other hand, a lower temperature value, such as 0.2, sharpens the distribution, making it narrower and less random. This tends to produce more focused and deterministic responses. Choosing the appropriate temperature value depends on the desired balance between randomness and coherence in the generated text. By experimenting with different *temp* values, we aim to identify the optimal setting for aligning online and offline evaluations in search clarification. Subsequently, we rank the clarification panes for each query based on the predicted *Engagement Level* by GPT-3.5 and compare these rankings against the ideal ranked lists created using actual *Engagement Level*.

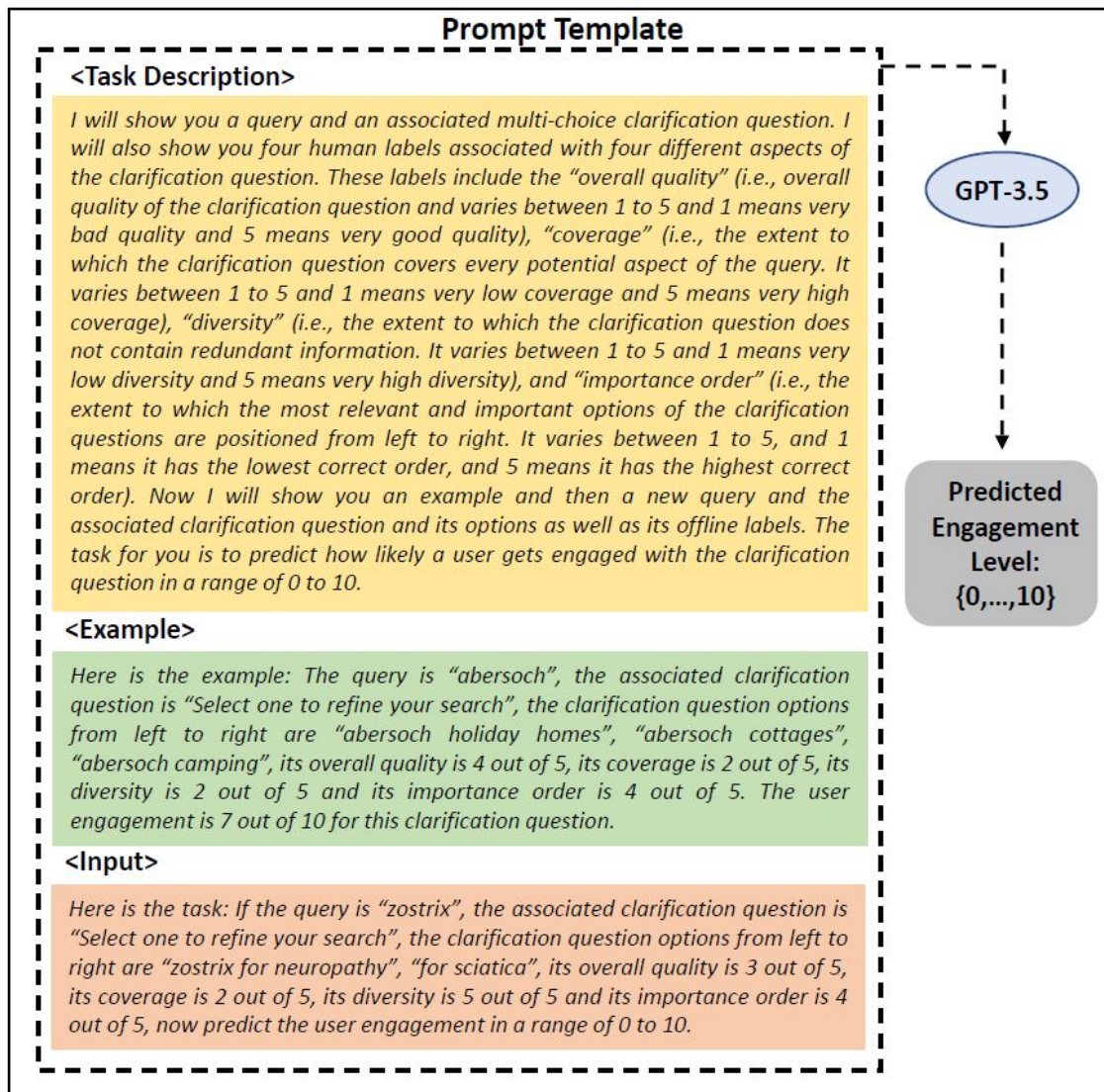


Figure 6.2: The prompt template used to feed the GPT model.

Impact of Query Length on the Relationship Between Online and Offline Evaluation

Here, we investigate the impact of query length on the relationships between online and offline evaluations in search clarification. While there is no universal definition of what constitutes a short or long query, some researchers have used a threshold of 3–5 words for short queries and 6 or more words for long queries. For example, Bendersky and Croft [12] defined short queries as those containing up to four words and long queries as those containing five or more words. In another study, Huston and Croft [45] used thresholds of 2, 4, and 5 words to distinguish between very short, short, and long queries. The *MIMICS-Duo* contains queries with a length of 1 to 9 words. However, the number of queries in the dataset for each query length varies. For instance, there are 45 queries with one word, while only 7 queries with 9 words. To investigate the impact of the query length and keep a balance between the groups in terms of the number of queries and query-clarification pair, we assume a query is short if the length is between 1–4 words (126 queries with 415 query-clarification pairs) and it is long if the length is between 5–9 words (180 queries with 619 query-clarification pairs).

Impact of Uncertainty in Online Labelling on Corresponding with Offline Evaluations

Here, we group the clarification panes based on the *Impression Level* and discard any query-clarification pair with a low *Impression Level*. As mentioned in Section 6.2.1, there is a three-step *Impression Level* per query-clarification pair (i.e., low, medium, high). The *MIMICS Impression Level* was computed based on the number of times the given query-clarification pair was shown to users. Hence, the *Impression Level* correlates with the query frequency. This highlights the fact that for the query-clarification pairs with low *Impression Level*, the *Engagement Level* obtained by the query-clarification pairs does not necessarily reflect how engaging it was. This part of the study helps us to focus on more reliable data. Removing the query-clarification pairs with the low *Impression Level* leaves the dataset with 212 queries and 703 query-clarification pairs with medium and high *Impression Level* and with one further step of filtering by removing the query-clarification pairs with medium *Impression Level*, 70 queries with 287 query-clarification pairs remain.

6.2.4 Evaluation Metrics

The ground truth in this study is the ranked list of clarification panes for a given query generated based on the *Engagement Level*. We want to know (1) how similar the offline

labels can produce the ranked lists of clarification panes, and (2) if they are able to place the MECP at the first position in the ranked list. An ideal ranked list of clarification panes is a list that has the MECP at the first position, and the rest of the clarification panes are sorted based on the *Engagement Level* in descending order.

If we assume the user behaviour is different for various reasons, such as the type of query they submit, users may get engaged with some clarification more than others. Since our aim is to show the MECP to the users, it does not matter whether the clarification pane with the *Engagement Level* of 10 is the top-rated or with the *Engagement Level* of 4. Hence, metrics such as precision at position one ($P@1$) or mean reciprocal rank (MRR) are appropriate for evaluating the position of the MECP in the ranked list, without taking into account the specific *Engagement Level*. We define $P@1$ as:

$$P@1 = \frac{TP}{TP + FP}$$

where true positive (TP) and false positive (FP) are the total numbers of clarification panes that are correctly and incorrectly top-rated, respectively, for all queries.

To measure MRR, we calculate the reciprocal rank at which the MECP is retrieved in a ranked list of clarification panes and calculate the mean value across all queries.

We also measure normalised discounted cumulative gain at position one (NDCG@1) that considers the relevance factor (here, the *Engagement Level*) when evaluating the top-rated clarification pane.

If our objective is to assess how well the search system ranks clarification panes, it becomes crucial to consider the overall quality of the ranked list in terms of arranging the clarification panes based on user engagement. To evaluate the ranked lists comprehensively, we utilise NDCG@3. The choice of a cutoff at 3 is based on the observation that approximately 70% of queries consist of only three clarification panes. Furthermore, for queries with four or more clarification panes, around 50% of those panes receive no user engagement. Hence, NDCG@3 ensures a fair evaluation of all clarification panes at a consistent depth.

We also calculate rank-biased precision (RBP) [76] and ranked-biased overlap (RBO) [114] that consider a binary relevance factor in the evaluation of the top-rated clarification pane in the list. RBP measures the utility rate that is gained by a user at a given degree of persistence (p), representing an aspect of user behaviour. Moffat and Zobel [76] assumed that a user inspects the first document and proceeds from the i th document to the $i+1$ th with fixed conditional probability p .

For instance, if $p=0.5$, the user obtains a high average per document utility, which means there is a relevant document in the first one or two rank positions. The RBP equation is

proposed below:

$$RBP = (1 - P) \sum_{i=1}^d r_i \cdot p^{i-1}$$

where r_i indicates the binary relevance of the i th ranked document scored as either 0 (not relevant) or 1 (relevant).

The RBP metric was introduced to measure the effectiveness of a ranked list retrieved for a query and varies between 0 and 1. It cannot be used directly in this study as only one clarification pane is shown to a user, not a list of clarification panes. To employ RBP in this study, we assume: (1) regardless of the value of *Engagement Level*, if there is a positive *Engagement Level* for a given clarification pane, $r_i=1$ and if not, $r_i=0$, and (2) since only one pane is shown to a user, we assume $p=0.05$, which means the probability of a user checking the second clarification pane (if it exists) is roughly 5%. We also calculate RBP for p values of 0.5 and 0.7 to investigate the clarification pane ranked lists at deeper depths. We calculate the RBP for every ranked list generated by each offline label and report the average RBP for each label.

The second rank-biased metric is RBO, which was developed by Webber et al. [114] and is a similarity measure to compare two ranked lists, quantifying how far the observed ranking deviates from the ideal ranking. It has the same assumptions as RBP and can be calculated using the equation below:

$$RBO = (1 - P) \sum_{k=1}^{\infty} p^{k-1} \frac{|A_{1:k} \cap B_{1:k}|}{k}$$

where A and B are two ranked lists, k is the depth of comparison, $|A_{1:k} \cap B_{1:k}|$ is the size of intersection between two lists at depth k .

RBO varies between 0 and 1; 1 means both ranked lists are identical, and 0 means they are completely disjoint. It is evident that RBO investigates the overlap and ordering between two ranked lists (the number of identical documents shared between two ranked lists). The current RBO definition cannot be used in this study as the clarification panes for a given query in the ranked lists generated by any two labels are always the same. Therefore, RBO in the current definition is always 1. To adopt RBO in this study, we define the size of the intersection of two ranked lists based on the number of panes that have the same positions in both lists. We calculate RBO between the ideal ranked list generated by *Engagement Level* and ranked lists generated by offline labels.

6.3 Results

We present the results of experiments on online-offline evaluations in search clarification.

Table 6.1: Relationships between the ranked lists of clarification panes created by the *Engagement Level* and created by offline labels.

| Engagement Level vs. | | Metric | | | | | |
|-----------------------|-------------------------|--------------|--------------|---------------------------|---------------------------|--------------|--------------|
| | | NDCG@1 | NDCG@3 | P@1 | MRR | RBP | RBO |
| <i>Offline Rating</i> | | 0.459 | 0.729 | 0.559 [†] | 0.749 [†] | 0.520 | 0.339 |
| Aspect | <i>Overall Quality</i> | 0.433 | 0.724 | 0.562 [†] | 0.760 [†] | 0.503 | 0.301 |
| | <i>Coverage</i> | 0.448 | 0.725 | 0.569 [†] | 0.747 [†] | 0.510 | 0.329 |
| | <i>Diversity</i> | 0.454 | 0.731 | 0.523 [†] | 0.726 [†] | 0.515 | 0.323 |
| | <i>Importance Order</i> | 0.412 | 0.706 | 0.484 [†] | 0.710 [†] | 0.455 | 0.275 |
| | <i>Mean</i> | 0.438 | 0.723 | 0.535 | 0.736 | 0.496 | 0.307 |
| <i>Random Ranker</i> | | 0.403 | 0.706 | 0.307 | 0.561 | 0.469 | 0.285 |

[†] Significantly different from the Random Ranker baseline (Tukey HSD test, $p < 0.05$).

6.3.1 Overall Practice in Designing Online and Offline Evaluations in Search Clarification

This analysis involves examining the position of the MECP within the offline ranked lists and assessing the similarity between the two lists (i.e., clarification ranked lists created based on the offline labels and the ideal clarification ranked lists created based on the *Engagement Level*). The results are shown in Table 6.1. For the sake of reproducibility, our results and codes are publicly available.⁵ To assess the performance of the offline labels in comparison to a baseline, we additionally rank the clarification panes for each query using a Random Ranker.⁶ The evaluation metrics of P@1, MRR, NDCG@1, and RBO and RBP at $p=0.05$ are used to evaluate the performance of the offline labels and models in identifying the MECP for a given query, and NDCG@3 is used to evaluate the similarity of the ranked lists created by the offline labels and models with the ideal ranked lists based on the *Engagement Level*. We perform Tukey honestly significant difference (HSD) [108] to find the means that are significantly different from each other for each column in the table. The Tukey HSD test is a post hoc test used when there are equal numbers of subjects in each group for which pairwise comparisons of the data are made [99]. The highest-performing label is highlighted in bold within each column.

Offline Labels. In the first round of the experiments, we aimed to understand how well offline labels correspond with user engagement (online evaluation) in ranking clarification panes and whether they can help us identify the MECP for a given query. Table 6.1 shows that (1) the MECPs are more likely to have the highest *Overall Quality* and *Coverage* compared to other clarification panes; (2) all offline labels perform noticeably better than

⁵https://github.com/Leila-Ta/On_Off-Eval-Search_Clarification

⁶Random Ranker is repeated 1000 times, and the mean values are reported.

a Random Ranker (e.g., *Coverage* shows 85% improvement over a Random Ranker in presenting the MECP for a given query at the top of the ranked list). However, *Importance Order* evaluation methodology showed the poorest performance among all offline methods. These findings are derived from the analysis of P@1 and MRR metrics, revealing statistically significant differences between them. The slight improvements over a Random Ranker shown by other metrics (i.e., NDCG@1, NDCG@3, RBP, and RBO) are not significant. This indicates that the metrics used to compare online and offline evaluations in search clarification have noticeable influences on the result justifications. For instance, P@1 and MRR are unconcerned about the user *Engagement Level*, and they only check the rank of the MECP. While for NDCG@1, if an engaging clarification that is not the MECP is ranked top, it still receives a score. Such an evaluation increases the chance of a Random Ranker showing a better performance than when the evaluation is only based on the position of the MECP.

As indicated in Section 6.2.4, we also calculated RBP and RBO for two higher p values (i.e., 0.5 and 0.7) in addition to 0.05 that is shown in Table 6.1 to investigate the similarity in the ranked lists at deeper depths. We observed that the performance of offline labels merged toward a Random Ranker by increasing the p value. We also considered the Kendall (τ) [56], and Spearman (r_s) [115] rank correlations between online and offline ranked lists but did not observe correlations. The majority (70%) of the ranked lists only had three clarification panes, and such a correlation analysis may not be accurate enough to draw conclusions.

During the second phase of the experiments, our objective was to assess the abilities of LL and LTR models in prioritising clarification panes. This involved incorporating offline labels as input features for the models. The performances of the LTR and GPT-3.5 models in ranking the clarification panes are shown in Table 6.2.

LTR Models. *SVM-rank* exhibits better performance compared to other LTR models. However, its superior performance is not significantly different from the other LTR models. When evaluating the effectiveness of LTR models using P@1 and MRR and comparing them to the *Overall Quality* or *Coverage* labels in Table 6.1 (two outperforming offline labels based on the same metrics), it becomes apparent that LTR models that incorporate the offline labels as input features do not outperform the individual offline labels in accurately ranking the MECPs at the highest positions in the lists. However, the performances of *SVM-rank* and *AdaRank* are significantly better than the Random Ranker, presented in Table 6.1. It seems the complexity of the LTR models may not be adequate to capture the underlying patterns present in the data. Furthermore, the characteristics and size of the training data can also impact the performance of LTR models, posing a challenge for the models to effectively learn robust patterns and generalise effectively.

Large Language Model. Table 6.2 also indicates the performance of GPT-3.5. We ex-

Table 6.2: Evaluation of three GPT-3.5 configurations across varying temperature settings and five LTR models, utilising offline labels to generate ranked lists of clarifications.

| Engagement Level vs. | Metric | | | | | |
|-----------------------------|--------------|--------------|------------------------------|------------------------------|--------------|--------------|
| | NDCG @1 | NDCG @3 | P@1 | MRR | RBP | RBO |
| <i>RandomForests</i> | 0.473 | 0.739 | 0.357 ^{†‡£} | 0.611 ^{†‡£} | 0.507 | 0.358 |
| <i>AdaRank</i> | 0.472 | 0.736 | 0.426 ^{†‡£§} | 0.673 ^{†‡£§} | 0.498 | 0.340 |
| <i>MART</i> | 0.468 | 0.733 | 0.341 ^{†‡£} | 0.609 ^{†‡£} | 0.508 | 0.342 |
| <i>RankBoost</i> | 0.459 | 0.733 | 0.364 ^{†‡£} | 0.639 ^{†‡£} | 0.486 | 0.345 |
| <i>SVM-rank</i> | 0.456 | 0.741 | 0.427 ^{†‡£§} | 0.698 ^{†‡£§} | 0.495 | 0.346 |
| <i>GPT-3.5 (temp = 0.0)</i> | 0.460 | 0.734 | 0.663 ^{†§*} | 0.830 ^{†§\$} | 0.525 | 0.382 |
| <i>GPT-3.5 (temp = 0.5)</i> | 0.439 | 0.718 | 0.588 [§] | 0.778 [§] | 0.487 | 0.363 |
| <i>GPT-3.5 (temp = 1.0)</i> | 0.468 | 0.732 | 0.539 [§] | 0.751 [§] | 0.523 | 0.386 |

^{†, ‡, £} Significantly different from GPT-3.5 with *temp* = 1.0, *temp* = 0.5, and *temp* = 0.0, respectively.

[§] Significantly different from the Random Ranker baseline (Table 6.1).

^{*} Significantly different from *Coverage*, the best performing label in terms of P@1, Table 6.1.

^{\$} Significantly different from *Overall Quality*, the best performing label in terms of MRR, Table 6.1.

amined GPT-3.5 using three different temperature settings: 0.0, 0.5, and 1.0. Comparing Table 6.1 and 6.2 reveals that GPT-3.5 outperforms LTR models in terms of P@1 and MRR when a temperature of 0.0, 0.5 and 1.0 are utilized. Moreover, GPT-3.5 shows significantly better performance compared to the individual offline labels of *Overall Quality* and *Coverage* when a temperature of 0.0 is used. Obtaining the best results with a temperature value of 0 suggests that GPT-3.5 has achieved optimal performance by using a deterministic approach. This deterministic behaviour is advantageous when we want to prioritise consistency and precise predictions. However, it is important to note that using a temperature of 0 may lead to overly rigid and repetitive outputs, as the lack of randomness can result in a lack of diversity. Depending on your specific application and requirements, you may need to strike a balance between precision and diversity by exploring other temperature values. When the temperature value is set to 0, it means that the output generated by GPT-3.5 is determined solely by the model’s confidence scores. In other words, the model selects the most probable word or token at each step without any randomness or variation. This finding emphasises the efficacy of GPT-3.5 in predicting online user engagement and, hence, accurately identifying the MECPs when incorporating offline labels as input features. However, similar to LTR models and offline labels, GPT-3.5 falls short of significantly surpassing the performance of the Random Ranker in ranking multiple clarification panes for given queries (no significant differences were observed between the performances of

GPT-3.5 and the Random Ranker in terms of NDCG@3.

We also observed that when GPT-3.5 is provided with high-quality human-annotated labels of clarification characteristics, it can show better performance compared to the *Offline Rating* labelling approach conducted by crowd-source workers. In the crowd-sourcing task, all the generated clarification panes for a given query were presented to workers simultaneously, and the workers were asked to rate all the panes based on their preferences (without having access to the *Aspect* labels). Although GPT-3.5 did not have the ability to predict the relative *Engagement Level* among the panes and evaluated each pane independently, its user engagement prediction resulted in more successful identification of the MECPs compared to the *Offline Rating* labelling method.

6.3.2 Impact of Query Length on the Relationship Between Online and Offline Evaluations

Table 6.3 shows the calculated metrics for short (1–4 words) and long (5–9 words) queries. We see that if a query is short, the *Offline Rating* evaluation performs better than other offline labels in placing the MECP at rank one (i.e., obtaining the highest P@1, MRR and RBO). However, if the query is long, selecting the MECP from a pool of clarification panes generated for a query can be carried out using *Overall Quality* and *Coverage* evaluations. Similar to the previous table, no conclusion can be drawn about the impact of the query length on the similarity of the ranked lists, as they do not show any significant improvement over a Random Ranker (no significant differences were measured in NDCG@3 between offline labels and the Random Ranker).

The results in Tables 6.1 and 6.3 show that the rankings of P@1 and MRR are consistent. To further analyse, we performed a Tukey HSD test on the calculated P@1 and MRR values for short, long, and all queries. The results indicate that there are no significant differences, suggesting that the length of the query does not have an impact on the relationship between offline evaluations and online evaluations in the context of search clarification.

6.3.3 Impact of Uncertainty on the Relationship Between Online and Offline Evaluations

In the third experiment, we separate the clarification panes based on the *Impression Level*. We learned from Zamani et al. [125] that a clarification pane with high *Impression Level* was shown to the users more than a clarification pane with low *Impression Level*. Therefore, the obtained *Engagement Level* by such a clarification pane is likely to be more reliable. In other words, the uncertainty in the collected online data is less. Table 6.4 shows the calculated

Table 6.3: Impact of the query length on relationships between the ranked lists of clarifications created by the *Engagement Level* and created by offline labels. (Short Query: 126 queries with 415 query-clarification pairs; Long Query: 180 queries with 619 query-clarification pairs.)

| Engagement Level vs. | | | Metric | | | | | |
|----------------------|----------------|------------------|------------|------------|--------------------|--------------------|-------|--------------------|
| | | | NDCG @1 | NDCG @3 | P@1 | MRR | RBP | RBO |
| Short Query (1-4) | Offline Rating | | 0.461 | 0.721 | 0.561 [†] | 0.751 [†] | 0.495 | 0.368 [†] |
| | Aspect | Overall Quality | 0.408 | 0.707 | 0.539 [†] | 0.748 [†] | 0.495 | 0.280 |
| | | Coverage | 0.412 | 0.702 | 0.539 [†] | 0.737 [†] | 0.473 | 0.317 |
| | | Diversity | 0.455 | 0.725 | 0.533 [†] | 0.737 [†] | 0.511 | 0.362 [†] |
| | | Importance Order | 0.371 | 0.680 | 0.478 [†] | 0.710 [†] | 0.422 | 0.269 |
| | | Mean | 0.412 | 0.704 | 0.522 | 0.733 | 0.475 | 0.307 |
| | Random Ranker | 0.376 | 0.684 | 0.289 | 0.550 | 0.422 | 0.259 | |
| Long Query (5-9) | Offline Rating | | 0.458 | 0.740 | 0.556 [†] | 0.745 [†] | 0.549 | 0.300 |
| | Aspect | Overall Quality | 0.469 | 0.748 | 0.595 [†] | 0.777 [†] | 0.490 | 0.325 |
| | | Coverage | 0.498 | 0.758 | 0.611 [†] | 0.762 [†] | 0.554 | 0.348 |
| | | Diversity | 0.452 | 0.741 | 0.508 [†] | 0.712 [†] | 0.512 | 0.270 |
| | | Importance Order | 0.472 | 0.743 | 0.492 [†] | 0.710 [†] | 0.503 | 0.293 |
| | | Mean | 0.473 | 0.748 | 0.552 | 0.740 | 0.515 | 0.309 |
| | Random Ranker | 0.441 | 0.739 | 0.333 | 0.578 | 0.516 | 0.302 | |

[†] Significantly different from the Random Ranker baseline (Tukey HSD test, $p < 0.05$).

metrics for all offline labels for the query-clarification pairs with high *Impression Level* and with medium and high *Impression Levels*. Table 6.4 indicates that when query-clarification pairs with low *Impression Level* were removed from the dataset (i.e., eliminating uncertainty from online evaluation), the clarification panes with the highest *Overall Quality* are likely to be the MECPs (obtaining high values of P@1 and MRR). However, no significant differences over a Random Ranker were observed for NDCG@3, showing that the offline labels are unable to produce clarification ranked lists better than a Random Ranker.

Table 6.4: Impact of the *Impression Level* on relationships between the ranked lists of clarifications created by the *Engagement Level* and created by offline labels.

| Engagement Level vs. | | | Metric | | | | | |
|----------------------|----------------|------------------|--------------|--------------|----------------------------|----------------------------|--------------|--------------|
| | | | NDCG @1 | NDCG @3 | P@1 | MRR | RBP | RBO |
| High | Offline Rating | | 0.617 | 0.837 | 0.614 [†] | 0.781 [†] | 0.701 | 0.417 |
| | Aspect | Overall Quality | 0.667 | 0.860 | 0.729 ^{†§} | 0.848 ^{†§} | 0.793 | 0.475 |
| | | Coverage | 0.657 | 0.849 | 0.657 [†] | 0.785 [†] | 0.765 | 0.461 |
| | | Diversity | 0.649 | 0.842 | 0.649 [†] | 0.782 [†] | 0.740 | 0.449 |
| | | Importance Order | 0.577 | 0.818 | 0.614 [†] | 0.764 [†] | 0.714 | 0.305 |
| | | Mean | 0.638 | 0.842 | 0.661 | 0.795 | 0.753 | 0.423 |
| | Random Ranker | 0.626 | 0.841 | 0.429 | 0.644 | 0.751 | 0.360 | |
| Medium-High | Offline Rating | | 0.524 | 0.765 | 0.623 [†] | 0.789 [†] | 0.588 | 0.427 |
| | Aspect | Overall Quality | 0.533 | 0.776 | 0.665 ^{†§} | 0.816 ^{†§} | 0.606 | 0.405 |
| | | Coverage | 0.535 | 0.772 | 0.618 [†] | 0.775 [†] | 0.613 | 0.404 |
| | | Diversity | 0.528 | 0.772 | 0.613 [†] | 0.773 [†] | 0.597 | 0.409 |
| | | Importance Order | 0.446 | 0.734 | 0.519 [†] | 0.731 [†] | 0.499 | 0.303 |
| | | Mean | 0.511 | 0.764 | 0.604 | 0.774 | 0.579 | 0.380 |
| | Random Ranker | 0.473 | 0.744 | 0.401 | 0.634 | 0.553 | 0.357 | |

[†] Significantly different from the Random Ranker baseline.

[§] Significantly different from the same metric calculated on all query-clarification pairs in Table 6.1.

By examining Tables 6.1, 6.3, and 6.4, it becomes evident that the *Importance Order* had the poorest performance compared to other offline labels. This implies that the engagement of users with the clarification pane is not significantly influenced by the order of candidate answers. Moreover, comparing Tables 6.1 and 6.4 shows much higher values for P@1 and MRR when we removed the query-clarification pairs with low *Impression Level* from the dataset. We performed a Tukey HSD test on the calculated P@1 and MRR values for *Overall Quality* between high *Impression Level* query-clarification pairs (top section in Tables 6.4) and all query-clarification pairs (Table 6.1) and between medium and high *Impression Level* query-clarification pairs (bottom section in Tables 6.4) and all query-clarification pairs (Table 6.1). The results indicate a significant difference between the two. This suggests that offline evaluation aligns more closely with online evaluation when the uncertainty in online evaluation is minimal, and the observed differences are unlikely to be random occurrences due to the sample size.

Additionally, we conducted GPT prompts using query-clarification pairs that only had a high *Impression Level*. We then compared the model’s performance in predicting the *Engagement Level* with the results obtained when using all query-clarification pairs. We only measured P@1, MRR, NDCG@1 and NDCG@3 here as the metrics of RBP and RBO did

Table 6.5: Impact of the *Impression Level* on the performance of three GPT-3.5 configurations across varying temperature settings.)

| Impression Level | Engagement Level (EL) vs. | Metric | | | |
|------------------|-----------------------------|--------------------|--------------------|--------------------|--------------------|
| | | NDCG @1 | NDCG @3 | P@1 | MRR |
| High | <i>GPT-3.5 (temp = 0.0)</i> | 0.658 [†] | 0.860 [†] | 0.786 [†] | 0.890 [†] |
| | <i>GPT-3.5 (temp = 0.5)</i> | 0.648 [†] | 0.844 [†] | 0.657 | 0.821 |
| | <i>GPT-3.5 (temp = 1.0)</i> | 0.614 [†] | 0.828 [†] | 0.529 | 0.749 |
| Low–Med.–High | <i>GPT-3.5 (temp = 0.0)</i> | 0.460 | 0.734 | 0.663 | 0.830 |
| | <i>GPT-3.5 (temp = 0.5)</i> | 0.439 | 0.718 | 0.588 | 0.778 |
| | <i>GPT-3.5 (temp = 1.0)</i> | 0.468 | 0.732 | 0.539 | 0.751 |

[†] Significantly different from GPT-3.5 with the same *temp* when using all query-clarification pairs.

not show the required capabilities for such comparisons. The results indicated a significant improvement in GPT performance, particularly when using *temp* = 0.0, compared to when using all query-clarification pairs. According to the findings presented in Table 6.5, when there is reduced uncertainty in the online evaluation and a stronger correlation between offline and online assessments, the performance of GPT-3.5 in predicting online user engagement improves when the GPT prompt includes offline labels.

6.3.4 The Most vs. the Least Engaging Panes

To enhance our understanding of how the offline labels correspond with the online label in MECPs, we compared the most engaging clarification panes with the least engaging clarification panes (LECPs) for queries that their clarification panes had high *Impression Level*. High *Impression Level* query-clarification pairs were chosen to ensure that the uncertainty in the low *Engagement Level* obtained by the LECPs is minimal. We observed that the *Overall Quality* of MECPs was higher than of the LECPs for more than 51% of the MECPs, and it agrees with our observations in Table 6.4 (see Figure 6.3). Although the percentage of the MECPs with higher *Coverage*, *Diversity* and the number of candidate answers were also higher than the LECPs, the observed higher percentages were not significantly different according to Student’s t-test. This indicates the *Overall Quality* of a clarification pane contributes to making it engaging from a user’s perspective.

We showed in Chapter 5 that the *Overall Quality* label has positive correlations with *Coverage*, *Diversity* and the number of candidate answers in a clarification pane. This implies that a clarification pane that solely focuses on having high *Coverage*, *Diversity* or a higher number of options does not guarantee user engagement. The findings indicate that an engaging clarification pane typically possesses a high overall quality, which means it needs to excel from multiple perspectives.

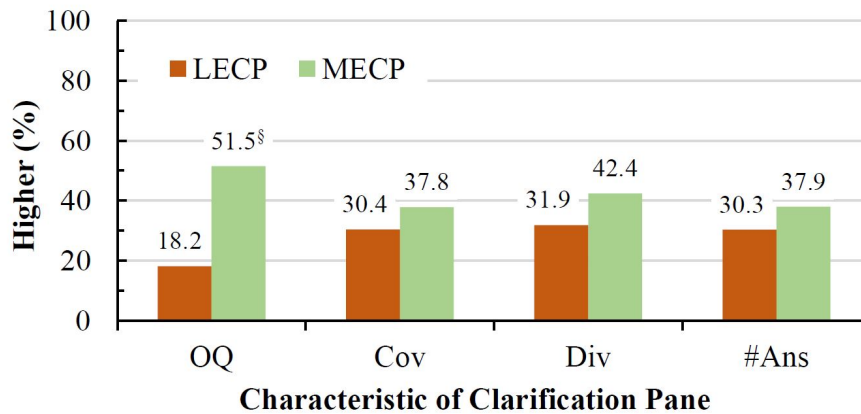


Figure 6.3: Variations of *Overall Quality* (OQ), *Coverage* (Cov), *Diversity* (Div) and the number of candidate answers (# Ans) in the MECPs when compared to the LECPs. (§: means the percentage of the MECPs that have higher *Overall Quality* than the LECPs is significantly different, Student’s t-test, $p < 0.05$).

6.3.5 Manual Clarification Pane Inspection

To explore the scenarios where a clarification pane with low quality might engage users more than a high-quality pane, we conducted a manual inspection of two queries. For these queries, the online and offline labels did not align well with their MECPs and LECPs.

In the case of the first query, “yucca”, the term can potentially refer to either a shrub or Yucca Mountain in Nevada, USA. The MECP is associated with the mountain, whereas the LECP is related to the plant. Upon analysing the clarification options for the MECP, we observed that they predominantly focused on a single intent and exhibited limited diversity. Specifically, terms such as “mountain”, “valley”, and “canyon” represented similar aspects of Yucca Mountain. Conversely, the clarification options for the LECP encompassed aspects of the yucca plant, indicating a greater diversity in the coverage of relevant information (see Tables 6.6).

In the data collection process, the workers were initially presented with the query and eight associated retrieved documents before annotating a label. Each retrieved document included a title and snippet. The workers were instructed to review these documents to understand various aspects related to the query before proceeding with the labelling task. In the case of the “yucca” query, we noticed that all the retrieved documents shown to the workers focused on the shrub, with no documents about the mountain. It is speculated that the workers inferred the query’s intent based on the content they reviewed in these documents and performed the labelling task with that intent in mind. However, the users recorded in the online data got more engaged with a different clarification pane, which covered the query’s intent not reflected in the retrieved documents (see Table 6.7). This suggests that as long as a clarification pane addresses an aspect of the query that is absent in the retrieved documents, users are likely to engage with it, irrespective of its quality.

Table 6.6: Examples queries and their most and least engaging clarification panes.

| Query | Pane | Clarification Options | | | | |
|---------------------------|------|-----------------------|--------------------------------|--------------|----------------|--------------|
| | | Option 1 | Option 2 | Option 3 | Option 4 | Option 5 |
| yucca | MECP | yucca valley | yucca mountain | yucca desert | yucca lake | yucca canyon |
| | LECP | yucca benefits | yucca nutrition facts | yucca powder | yucca for sale | <i>null</i> |
| why is my printer offline | MECP | hp | why is my printer offline dell | <i>null</i> | <i>null</i> | <i>null</i> |
| | LECP | in windows 10 | windows 8 | windows 7 | windows xp | <i>null</i> |

Table 6.7: Examples queries with online and offline labels.

| Query | Pane | Engagement Level | Overall Quality | Coverage | Diversity |
|---------------------------|------|------------------|-----------------|----------|-----------|
| yucca | MECP | 3 | 4 | 4 | 3 |
| | LECP | 1 | 5 | 3 | 4 |
| why is my printer offline | MECP | 8 | 3 | 2 | 2 |
| | LECP | 0 | 5 | 2 | 3 |

For the second query, “*why is my printer offline*”, the MECP asked for the printer brand, while the LECP requested clarity from a software point of view. The coverage and diversity labels for both clarification panes were shallow and correctly rated by the human annotators. However, the annotators believed that the LECP had higher quality than the MECP as it perhaps provided more options than the MECP, with only two options. Upon reviewing the retrieved documents, it becomes evident again that all of them are focused on printer issues occurring on various versions of Windows. None of the documents provide information specifically related to the brand of the printer.

Examining these two examples underscores the significance of soliciting clarification questions from users when the quality of retrieved documents is subpar. Moreover, it reveals that the accuracy of offline labelling is greatly influenced by the information provided to the workers before the labelling process in some instances.

6.4 Discussion

As highlighted in Section 6.1 of our discussion, the successful deployment of a clarification model necessitates thorough examination through both online and offline evaluation methods. Our research revealed a significant knowledge gap concerning the relationship between online and offline evaluations in the context of search clarification. This gap is considered one of the primary reasons for the low success rates of clarification models in effectively engaging users in real-world scenarios. Existing models were typically evaluated using either online or offline evaluations individually, neglecting the crucial understanding of how these two evaluation approaches correlate. To address this gap, we conducted a comprehensive investigation, as discussed in Subsection 2.4, to examine the relationship

between online and offline evaluations in the field of Information Retrieval. Our findings demonstrated substantial disparities between offline and online evaluations. The extent to which these disparities exist in search clarification remained unknown, motivating our development of the *MIMICS-Duo* dataset and conducting a detailed exploration in this chapter. Without a clear understanding of how online and offline evaluations align in the context of search clarification, the likelihood of proposing a clarification model that proactively engages users remains limited.

The results of this study diverge from the previous study conducted on search clarification [125]. Zamani et al. [125] examined the MIMICS dataset and investigated correlations between online and offline evaluations using a single offline label. They concluded that no correlation was observed between the two evaluation methods. In contrast, the current study analysed the *MIMICS-Duo* dataset focusing on identifying the MECP for a given query among multiple generated clarification panes, the ultimate goal of every clarification model. This approach allowed for the utilisation of various offline labels and reduced uncertainty in the dataset. As a result, this analysis revealed a relationship between online and offline evaluations in the context of search clarification. This is important as only one clarification pane is shown to a user at a time. Hence, we observed that the offline labels can assist us with identifying the most engaging clarification pane among multiple generated panes for a given query. Nevertheless, it is important to consider the following factors in any offline evaluations in search clarification.

We demonstrated that assessing the retrieval quality using online and offline evaluations frequently leads to divergent outcomes [27, 33, 38, 94]. We did observe a distinct connection between online and offline assessments, especially when the online data is relatively certain. However, our research supports previous studies by revealing a discrepancy between online and offline evaluations when it comes to ranking clarification panes for a given query.

We manually examined various panes to understand why users might show more engagement with lower-quality clarification panes. We observed that while the human annotation was carried out accurately based on the available information, it does not always guarantee that the annotators can accurately capture the user’s intent. This finding helps to explain the contradictions observed between online and offline evaluations.

In attempting to explain these discrepancies, we consider two explanations proposed by Teevan et al. [105] and Liu et al. [70]. Teevan et al. [105] suggested that different users who issue the same textual query may have distinct information needs or intentions, leading to varying evaluations. This implies that users’ subjective preferences and expectations play a significant role in assessing the quality of clarification panes. Liu et al. [70], on the other hand, proposed that there may be notable disparities between assessors’ judgements and users’ assessments due to differences between satisfaction prediction and document

relevancy prediction. Satisfaction, to some extent, is subjective, as different users may have varying opinions on what constitutes a satisfying experience.

Apart from the reasons mentioned here, it is essential to acknowledge that the information provided to annotators can impact the correlation between online and offline evaluations. When determining the MECPs, it is essential to assess the quality of the SERP and the clarification pane simultaneously and in relation to each other. Evaluating either component independently may lead to misleading conclusions in certain scenarios.

This study demonstrates the value of using collected offline labels for predicting online user behaviour and identifying the MECP within generated options for a query, particularly when employing Language Models for task formulation. Despite having identical input features, we observed different performances between the GPT and LTR models. The observations can be attributed to several factors:

- **Model Architecture:** GPT and LTR models have different architectures and underlying principles. GPT is a transformer-based language model that excels at capturing semantic and contextual information in text. On the other hand, LTR models, such as AdaRank or RankBoost, are specifically designed for learning to rank tasks and may have different assumptions and optimisations.
- **Learning Approaches:** GPT utilises unsupervised learning through language modelling objectives, which allows it to capture a wide range of language patterns and context. In contrast, LTR models often rely on supervised learning techniques with explicit relevance labels or features specific to ranking tasks.
- **Representation Power:** GPT, with its deep transformer architecture, can capture intricate patterns and dependencies within the input text. It can learn from vast amounts of data, which can provide it with a more comprehensive understanding of the relationships between queries, documents, and relevance labels.
- **Contextual Understanding:** GPT models excel in contextual understanding, as they consider the entire input sequence and can generate coherent and contextually appropriate responses. This contextual comprehension may enable GPT to better capture the nuances and relevance of user engagement, leading to superior performance in predicting MECPs.

6.5 Summary

How well online and offline evaluations correspond to each other in search clarification is the knowledge gap that was addressed in this study by answering the research questions

below:

- What are the best overall practices in designing offline evaluation methodologies for search clarification that correspond with online evaluation?

Offline evaluations can complement online evaluations in identifying the most engaging clarification pane for a given query. This suggests that offline evaluation methodologies can be useful for assessing the effectiveness of search clarification models in terms of user engagement (the main purpose of search clarification models). We have demonstrated that clarification panes must excel in multiple aspects to be considered engaging from a user's perspective. Merely having high *Coverage* or *Diversity* does not guarantee engagement. However, when it comes to ranking multiple clarification panes for a given query, offline evaluations do not outperform a Random Ranker. This implies that current offline evaluation methodologies may not be well-suited for evaluating the ranking performance of search clarification models. We also showed that some offline labels, in particular, *Overall Quality* and *Coverage* perform better than others in corresponding with the online label.

To identify the MECP from a user's perspective for a given query, we automated the ranking of clarification panes. We employed GPT and LTR models and utilised the offline labels as the input features for the models. The LTR models did not demonstrate advantages over individual offline labels. On the other hand, GPT surpassed both the LTR models and offline labels by successfully placing the MECP in the top position for a given query, showcasing its superior performance in this task.

- Does query length impact the relationship between online and offline evaluations in search clarification?

The impact of query length on the relationship between online and offline evaluations in search clarification is minimal. The evaluation metrics obtained from offline evaluations remain in the same order regardless of query length. However, the highest-performing offline label differs between short and long queries, indicating that different evaluation criteria may be more relevant depending on query length.

- Does uncertainty in the online evaluation impact the relationship between online and offline evaluation?

The reliability of online evaluation data influences the strength of the relationship between online and offline evaluation. When online data is more reliable, a stronger correspondence with offline evaluation is expected. This suggests that ensuring the quality of online evaluation data is crucial for obtaining meaningful insights.

Furthermore, we employed six distinct evaluation metrics and found that the specific choice of metrics can influence the relationship between online and offline evaluations in search clarification. If the goal is to examine both online and offline evaluations in

the context of identifying the most engaging clarification for a given query, we suggest focusing on the Precision at Rank 1 (P@1) and Mean Reciprocal Rank (MRR) metrics as top priorities. Metrics such as RBO and RBP that consider binary relevance are not appropriate for comparing online and offline evaluations in search clarification.

Despite the valuable insights provided by this study, there are certain limitations that should be acknowledged. The limitations include:

- While our research primarily examined five offline approaches, it was demonstrated that offline evaluations may not always align fully with online evaluations in certain instances. Enhancing the information given to annotators can improve the consistency between online and offline assessments.
- The study primarily focused on five specific offline evaluation approaches. While these approaches provided valuable insights, other potential methodologies or variations of existing approaches may exist that were not explored in this study.
- The study's findings were based on specific datasets and evaluation metrics. The generalisability of the results to other domains or search clarification scenarios may require additional investigation.
- User engagement is a subjective aspect, and different users may have varying preferences. While the study considered multiple aspects of user engagement, individual preferences and subjective interpretations of engagement may not be fully captured.

In our study, while acknowledging the potential influence of dataset size, the statistically significant differences we observed in our analysis provide a solid basis for drawing trustworthy conclusions. We have employed rigorous statistical methods to ensure the reliability of our findings, and the observed effects are unlikely to have occurred by chance alone.

In the upcoming chapter, we examine how the clarification modality influences user preferences. Additionally, we explore the effectiveness of text-to-image models in generating relevant images for the clarification panes.

Chapter 7

Understanding Modality Preferences in Clarification Questions

In Chapter 6, we investigated computer-generated clarification questions in a search engine and studied what makes a clarification engaging from a user’s perspective. As the final stage in this research, we explore the use of a novel clarification question (CQ) form, namely *multi-modal clarification*, that has been, to our knowledge, largely unstudied. *Multi-modal* clarification entails using multiple media types, such as text and image, to refine and enhance search results.

We create a *multi-modal* search clarification dataset called *MIMICS-MM* that contains a collection of multi-choice clarification questions drawn from the *MIMICS-Manual* dataset with associated images that are handpicked and collected by an expert annotator.¹ Through crowd-sourcing efforts, we analyse user preference regarding different clarification modes, scrutinising the impact of image and text quality, clarity, and relevance on user preference. Building upon these observations, our study further explores the automatic generation of images, effectively comparing the quality and relevance of these model-generated images with human-collected ones. Additionally, we investigate user preferences concerning model-generated images versus human-generated images.

¹*MIMICS-MM* is publicly available at <https://github.com/Leila-Ta/MIMICS-MM>

7.1 Introduction

Effective communication between users and intelligent systems is crucial to finding a user’s information needs. One common obstacle encountered by Information Retrieval systems is the inherent ambiguity present in human language. Clarification questions can play a pivotal role in search interactions, allowing users to refine their queries and obtain more precise search results. Traditionally, clarification questions have been presented in a textual format, allowing users to respond with further textual input. While clarification has become an important component of many conversational and interactive information-seeking systems [128], previous research has shown that even though clarification questions receive positive engagement, users are not easily engaged with them [126, 124].

Recent advancements in technology have introduced new modalities, such as visual prompts or multi-modal, which is a combination of text and visuals. As emphasised in the most recent Alexa Prize TaskBot Challenge [1], there are instances in which *multi-modal* interactions (e.g., text and image) impact the user experience in conversational information-seeking systems [31].

Although the incorporation of visual elements allows users to provide more context and improve query accuracy, the extent to which different modalities (such as text or image) can enhance user interaction in search engines is still uncertain. Previous studies have primarily focused on text-based clarification, neglecting the potential benefits of *multi-modal* approaches. By exploring user preferences over various modalities, we can shed light on which modalities are perceived as more effective and intuitive for optimising both user experience and system performance. We study *multi-modal* clarification questions from the user behaviour perspective and explore user preference on clarification question modalities, specifically focusing on *text-only*, *visual-only*, and *multi-modal* approaches. By systematically analysing user feedback, we can gain valuable insights into the advantages and limitations of each modality and the influential parameters.

A clarification pane typically consists of a multi-choice clarification question and a list of candidate answers [124]. A *multi-modal* clarification pane contains both visual and textual content for each candidate answer (see Figure 7.1). We aim to understand if adding a visual presentation to *text-only* clarification panes enhances the user experience.

We explore user preferences over three modalities for clarification panes: (i) textual, (ii) visual, and (iii) *multi-modal* (i.e., a combination of the two). We randomly sample 100 query-clarification pairs from the *MIMICS* dataset [124]. Then, we create the visual and *multi-modal* clarification panes for the sampled query-clarification pairs through a controlled manual expert annotation process. Pairwise user preferences are collected for different modalities following a post-task questionnaire to answer the following research

question:

- Do users prefer *multi-modal* clarification panes over *uni-modal* (i.e., textual or visual)?

In this study, we investigate the impact of the image quality, image/text clarity, and relevance of the text and image, in addition to various image aspects, on user preference. Finally, we explore whether generating corresponding images to the clarification panes can be automated using text-to-image generation models. The quality and the relevance of generated images, in addition to user preferences over human-collected and model-generated images, are investigated through manual annotation. Our experiments reveal that:

- In the majority of cases (70-80%), users prefer *multi-modal* clarification panes over *visual-only* and *text-only* clarification panes. They also prefer *visual-only* clarification over *text-only* clarification in 54% of cases.
- Crowd-source workers prefer *multi-modal* clarification panes as they are easier to understand, which helps users make better and faster decisions.
- Image quality, clarity, and relevance, in addition to text clarity, have a direct impact on self-reported user perceptions.
- Text-to-image generation models, such as Stable Diffusion [91], are capable of automating image generation for creating *multi-modal* clarification panes.

Our contributions in this chapter include:

- Gaining a better understanding of user preferences when it comes to different clarification modalities.
- Evaluating the influence of image and text properties on user preference. By investigating how different factors related to images and text affect user choices, we gain insights into the impact of these properties on search clarification.
- Exploring the capabilities of text-to-image generation models in the context of search clarification. By studying the effectiveness of these models in generating relevant images based on textual queries, we investigate their potential use in enhancing the search clarification process.

Overall, our findings provide valuable insights into how to engage the user better with clarifications in information-seeking systems. By understanding user preferences and leveraging *multi-modal* approaches, we can create more effective systems that cater to the needs of users in search clarification scenarios.

7.2 Experimental Design

We now describe the methodology and structure of the data collection, including the experiments.

Query and clarification panes sampling. We use the *MIMICS-Manual* dataset to select textual CPs. We randomly select 100 queries and their corresponding multi-choice CPs to create the *MIMICS-MM* dataset. The number of candidate answers in the CP varies between two and five.

Clarification image collection. To assign an image to each candidate answer of CPs, an expert annotator searches the online website for corresponding images to those candidate answers using the Google images search engine.² In total, 314 images are matched with 314 textual candidate answers. The annotator re-evaluates the quality of the images and, if needed, replaces them with images of greater quality.

Experimental design. Online experiments are conducted on Amazon Mechanical Turk (AMT) to gather user preference labels through Human Intelligence Tasks (HITs). We designed three tasks to collect judgements from AMT workers on user preferences over different modalities in search clarification. We run pairwise comparisons as follows:

- Task I: *text-only* (T) vs. *visual-only* (V)
- Task II: *text-only* (T) vs. *multi-modal* (MM)
- Task III: *visual-only* (V) vs. *multi-modal* (MM)

A query and two modalities are shown in Figure 7.1. At the end of this data collection process, three different subsets are created.

Post-task questionnaire. After showing a query and two clarification question options, workers are presented with a post-task questionnaire assessing their presentation style preference and feedback (Figure 7.2). Thus, after inspecting the query, CQ, and candidate answers, workers indicate which presentation they prefer (*Q1*). Workers are also asked to justify their preference with four questions (*Q2–5*). The second question (*Q2*) contain checkboxes with options about the text and images' clarity, quality, and relevance. Workers are asked three more questions to obtain the motivation behind their choice of which modality is easier to understand (*Q3*), which helps them make better (*Q4*) and faster decisions (*Q5*) on a 5-point slider (e.g., in Task 2, labels 1 and 2 means *text-only* modality is preferred, label 0 means they have no preference, and labels 4, and 5 mean multi-modality is preferred).

²We watermark the images for copyright compliance.



| Query: spectre Clarification question: Select one to refine your search | |
|---|--|
| Option 1: Textual answers | Option 2: Multi-modal answers |
| Spectre comics |  Spectre comics |
| Spectre organization |  Spectre organization |

Figure 7.1: An example of Task II (T vs. MM).

Quality assurance. We include two quality assurance checks. For example, each task contains a gold question (i.e., a question with a known answer) with the aim of high validity throughout the task. Workers who fail to answer the gold question are prohibited from completing other tasks, and their answers are removed. We also manually check 10% of submitted HITs per task as a final quality assurance check. Invalid submissions are removed, and the workers are denied from completing subsequent tasks. We then open those HITs to other workers.

AMT pilot tasks were carried out³ to analyse the flow, acquire users' feedback, check the quality of collected data, and estimate the required time to finish each task and a fair pay rate.

Workers. Only workers based in Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States, with a minimum HIT approval rate of 98% and a minimum of 5,000 accepted HITs, are allowed to participate in the study, maximising the collected data quality and the likelihood that workers are either native English speakers or have a high level of English. Each HIT is assigned to at least three different AMT workers, enabling us to use an agreement analysis measure on their modality preferences. In case of disagreements, we administer the HIT again to more workers until we achieve a final majority vote. Each worker is allowed to perform 25 tasks (a portion used for each launch). Workers have a five-minute time limit to finish the task and are compensated with 0.74

³Pilot study was conducted in February 2022.

1) Which presentation do you prefer?

☐ Option 1 (textual) ☐ About the same ☐ Option 2 (multimodal)

2) Why did you pick the previous option considering the query?

☐ The text is clearer and more understandable compared to multimodal

☐ The multimodal is clearer and understandable compared to only text

☐ The images have good quality

☐ The images have poor quality

☐ The texts is not relevant compared to the query

☐ The images is not relevant compared to the query

☐ Other

3) What is Iron Man?

☐ Superhero ☐ Dancer

4) Which presentation did you find easier to understand?

Option 1 (textual) Option 2 (multimodal)

5) Which presentation helped you make a good decision?

Option 1 (textual) Option 2 (multimodal)

6) Which presentation helped you make the decision faster?

Option 1 (textual) Option 2 (multimodal)

Figure 7.2: Questionnaire template.

USD per HIT.

Image generation for *multi-modal* clarification. Following a crowd-sourcing approach, we utilise two text-to-image generation models, namely Stable Diffusion [91] and Dall·E 2 [86]. These models are employed to produce images related to candidate answers, with the aim of exploring their potential in generating multi-modal clarifications.

Stable Diffusion⁴ is a neural text-to-image model that uses a diffusion model variant called the latent diffusion model. It is capable of generating photo-realistic images given text input. Dall·E 2, created by OpenAI, generates synthetic images corresponding to an input text. Our input to generate a corresponding image to a candidate answer of a CP is the concatenation of the query and the candidate answer text. This input is used to generate all corresponding images for all candidate answers (two images are generated by two employed models per candidate answer).

Comparing human-collected versus computer-generated images. First, we evaluate and compare the generated images' visual aspects with manually collected images. We extract the visual aspects of the images using *OpenIMAJ* [42], a tool for multimedia content analysis.

⁴The Diffusers library available at <https://github.com/huggingface/diffusers> is used for this study.

The nine visual aspects investigated are *brightness*, *colourfulness*, *naturalness*, *contrast*, *RGB contrast*, *sharpness*, *sharpness variation*, *saturation*, and *saturation variation* [106]. We conduct a manual annotation to investigate the generated images’ relevance to the text, compare the images’ quality, and assess the user preference over generated and collected images. Three annotators, two men and a woman with proficient English and a higher degree, complete the labelling. Each annotator labels 314 generated images. We collect all annotations, aggregate them, and in case of any disagreements, majority voting is used for the final label.

We show the concatenation of the query and the candidate answer in the text and the corresponding generated image to the annotators (similar to *Q1*, Figure 7.2). We ask annotators if the image is relevant to the text or not on a binary scale (i.e., label 1 means relevant, and label 0 means irrelevant). This label is similar to the label collected for the human-collected images during crowd-sourcing. Then, we show the collected image for the same text from the crowd-sourcing part and ask the annotators to compare the quality of generated and collected images regardless of the presented text on a 3-point scale (i.e., the quality of the computer-generated image is higher (2), are the same (1), or the human-collected image has a higher quality (0)). Finally, the annotators are asked to indicate their preferred image between two images on a 3-point scale (i.e., annotators prefer the computer-generated image (2), have no preference (1), or prefer the human-collected image (0)).

7.3 Results

In this section, we investigate the impact of various clarification modality characteristics and visual aspects of the images on user preference. Furthermore, we explore whether the CPs’ visual modality can be automated.⁵

User preference and clarification modality. We first investigated user preferences over the clarification modality in each pairwise comparison (i.e., *text-only* vs. *visual-only*, *text-only* vs. *multi-modal*, and *visual-only* vs. *multi-modal*). To understand whether a preferred modality in each pairwise comparison is significantly different from the other two options, we performed the Tukey honestly significant difference (HSD) test [108]. This statistical significance test helped us determine, for instance, if the number of users who preferred *multi-modal* over *text-only* was significantly higher or not. Table 7.1 indicates the percentage of user preference in each pairwise clarification modality comparison. In Task 1, where the workers indicated their preferences between the *text-only* and *visual-only* clarifications,

⁵For the sake of reproducibility, our results and codes are publicly available. <https://github.com/Leila-Ta/MIMICS-MM>

Table 7.1: Pairwise preference for clarification modality (%)

| Task | Prefer Text | Prefer Visual | Prefer Multi-Modal | No preference |
|-------------------------------|-----------------|------------------|-----------------------|------------------|
| Text vs. Visual | 39 [†] | 54 [†] | NA | 7 [†] |
| Text vs. Multi-Modal | 17 [†] | NA | 79 [†] | 4 [†] |
| Visual vs. Multi-Modal | NA | 17 | 71 [†] | 12 |

[†] Significantly different from the other two preferences (Tukey HSD test, $p < 0.05$).

we observed that 54% of the workers preferred *visual-only* over *text-only* CPs. In Tasks 2 and 3, where the workers indicated their preferences between *uni-modal* and *multi-modal* CPs, the workers strongly preferred *multi-modal* CPs, no matter whether the uni-modal CP is *text-only* or *visual-only*. The workers' preferences were significantly different from other options, indicating that in 70-80% of the cases, a *multi-modal* clarification was preferred.

Post-task questionnaire analysis. We asked the workers to explain if the text/image clarity relevance, and image quality impacted their preferences. We calculated the Pearson correlations between the workers' preferences and the characteristics of the clarification modalities in each Task. In Task 1, we observed a positive correlation ($\rho=0.476$) between user preference (i.e., preferring *visual-only* clarifications over *text-only* ones) and image quality. There was also a strong positive correlation ($\rho=0.677$) between user preference and image clarity, and user preference had a strong negative correlation ($\rho=-0.686$) with text clarity. The same correlation trends and orders were observed for the user preference (i.e., preferring *multi-modal* clarifications over *text-only* ones) with image quality ($\rho=0.458$), image clarity ($\rho=0.626$) and text clarity ($\rho=-0.627$). However, in Task 3, the user preference (i.e., preferring *multi-modal* clarifications over *visual-only*) had correlations only with the text clarity ($\rho=0.505$) and image clarity ($\rho=-0.301$). A closer look at the collected feedback from workers showed that the text and the image in more than 95% of CPs were relevant. This explained low to zero correlations between user preference and the relevance of the text and the image. We calculated the Tukey HSD test and observed the calculated correlations were significantly different from each other.

In the pairwise comparison between *multi-modal* and *visual-only* CPs, although the collected images for the CPs were the same, the workers preferred *multi-modal* CPs over the *visual-only* ones when the images were not clear and the text helped them understand the candidate answers to the CPs. The users preferred *visual-only* clarifications in more than 54% of cases when the text clarity was low, and the image quality and clarity were high, although the text and image were relevant in most cases.

We investigated the users' motivation for their preferences in the post-task questionnaire. We asked users whether the preferred modality was easier to understand and helped them make better and faster decisions. Table 7.2 shows the user preferences in each pairwise

Table 7.2: Motivations behind user preference (%).

| Motivation | T vs. V | | T vs. MM | | V vs. MM | |
|----------------------|-------------|---------------|-------------|--------------------|---------------|--------------------|
| | Prefer Text | Prefer Visual | Prefer Text | Prefer Multi-Modal | Prefer Visual | Prefer Multi-Modal |
| Easier to understand | 25 | 31 | 7 | 61 | 6 | 67 |
| Better decision | 22 | 36 | 6 | 68 | 3 | 67 |
| Faster decision | 27 | 36 | 10 | 62 | 6 | 66 |
| None of the above | 8 | 12 | 4 | 9 | 9 | 5 |

modality. We see when users preferred *visual-only* CPs over *text-only* ones, 31% of users believed that the *visual-only* CPs were easier to understand, and the *visual-only* modality helped 36% of users make better and faster decisions. When comparing *multi-modal* CPs with *text-only* and *visual-only* CPs, between 60 to 70% of users believed that *multi-modal* CPs were easier to understand and helped them make better and faster decisions. Table 7.2 shows that there were small groups of users whose motivations behind their preferences were not listed in our questions.

User preference and impact of visual aspects. In the next step, we investigated the impact of visual aspects of the collected images on user preference over the clarification modality. We calculated the point-biserial correlation⁶ [102] between the visual aspects of images and user preferences, the image quality and the image clarity. The average value of each aspect was calculated across all candidate answers for each CP. Therefore, one value was obtained per visual aspect for every CP. There was a low correlation between the image’s visual aspects and user preference, including the image quality and clarity that the workers judged. To further explore the impact of visual aspects of images on user preference, we developed a feature-level attribution explanation to rate the image’s visual characteristics based on their user preference. We utilised the Gini importance of the random forest with visual aspects as the input and target label user preference (i.e., 0 means Text preferred over Multi-Modal and 1 means Multi-Modal preferred over Text). The Gini importance is a metric that determines the relative significance of features in a random forest model. In this case, the visual aspects of the data were considered when calculating the Gini importance. By incorporating visual aspects into the Gini importance calculation, it is likely that the model was able to capture and evaluate the relevance of visual features in the dataset. This can be particularly useful in scenarios where visual information plays a significant role or provides valuable insights for the given problem or task. We performed this analysis for Task 2, and the results indicate that *brightness*, *naturalness*, *RGB contrast*, *sharpness variation*, and *saturation variation*, among other studied aspects, accounted for more than

⁶The point-biserial correlation measures the relationship between a binary (i.e., user preference, image quality, and clarity) and a continuous variable (i.e., image aspects).

65% of the differences in user preferences. In particular, *brightness* and *naturalness* were the two most important visual features.

Automatic image generation for clarification panes. Finally, we investigated if generating the corresponding images to the candidate answers can be automated. First, we compared the visual aspects (e.g., *brightness*, *colourfulness*, *naturalness*, ...) of the generated images with the collected ones. We observed that the generated images had relatively the same visual aspects as the collected ones. However, the Stable Diffusion model generated images that had similar sharpness to the human-collected images.

Second, we compared computer-generated images with human-collected ones regarding image relevance, quality, and user preference. Table 7.3 shows that 87% of Stable Diffusion generated images were relevant to the text. Even though only 20.7% of the generated images had a higher quality compared to human-collected ones, more than 57% of images had higher or equal qualities compared to collected ones. Only 12.7% of the generated images were preferred over the human-collected images. However, as can be seen from Table 7.3, 39.8% of the users either preferred the generated images or had no preferences over the generated and collected images (same preference). A slight improvement in the model performance was observed when we removed the irrelevant generated images from the collection (i.e., the percentage of generated images that had higher quality than the collected images rose from 20.7% to 21.2%, and the percentage of generated images that were preferred over collected images rose from 12.7% to 14.6%). The annotators preferred the collected images over ~60% of computer-generated images. This observation was expected as the collected images were gathered through online searching to select the most suitable images, while a text-to-image model generated an image from only text. However, the Stable Diffusion model could generate relevant and high-quality images. As, in ~80% of cases, users preferred a *multi-modal* CP over a *text-only* one; such a text-to-image model can ease and fasten the task of generating *multi-modal* CPs.

Table 7.3: Comparison of human-collected and computer-generated search clarification question images.

| Collection Method | Relevance | Image Quality ¹ | Image Preference ² |
|----------------------------------|-----------|----------------------------|-------------------------------|
| Human-Collected | 96% | 42.7% | 60.2% |
| Stable Diffusion model-Generated | 87% | 20.7% | 12.7% |

¹ 36.6% of users indicated the quality of the generated and collected images were the same.

² 27.1% of users indicated that they had no preferences over the generated and collected images.

7.4 Discussion and Summary

In this thesis, we aimed to understand how user engagement with clarification questions in information-seeking systems can be improved, as the literature review showed that users are often reluctant to get engage with the clarifications. Apart from the characteristics of engaging clarification questions, we understood that the presentation and modality of the clarification question play an important role in enhancing user interaction with the clarification question. *Multi-modal* clarification questions, which involve both text and other forms of media (such as images or videos), offer several benefits over text-only clarification in search engines. Here are some advantages:

- **Enhanced understanding:** Including multiple modalities helps users better understand and articulate their information needs. Sometimes, it is challenging to describe complex visual concepts or specific details accurately through text alone. By incorporating visual elements, users can provide more precise context and refine their queries effectively.
- **Richer context:** Images, videos, or other media can provide valuable context that complements textual information. For instance, when searching for a specific object or location, a multi-modal clarification question can include an image to convey visual details that may be difficult to describe adequately using words alone. This additional context can lead to more accurate search results.
- **Improved relevance:** By incorporating multiple modalities, search engines can leverage the power of both textual and visual cues to deliver more relevant search results. The inclusion of visual content in clarification questions enables search engines to understand user intent better and provide more targeted responses.
- **Expanded search scope:** Text-only clarification questions may limit the scope of search results to primarily textual content. However, multi-modal queries can extend the search to include a broader range of media-rich resources, such as images, videos, audio clips, or interactive content. This expands the possibilities for finding relevant information and discovering diverse content types.
- **Accessibility and inclusivity:** Multi-modal clarification questions can be particularly beneficial for individuals with visual impairments or those who may struggle with expressing their information needs through text alone. By accommodating different modes of communication, search engines can provide more accessible and inclusive experiences for a wider range of users.

- Efficient search experience: Incorporating multi-modal elements into clarification questions can lead to more accurate search results from the initial query, reducing the need for iterative refinement. Users can convey their intent more precisely, saving time and effort in the search process.

In certain situations, clarification models that exclusively produce textual clarification questions may lack clarity when dealing with queries that have multiple intentions. In such instances, incorporating visual content alongside the text could prove beneficial for users in fulfilling their informational requirements. Our study in this chapter showed that users generally preferred *multi-modal* clarification panes over *text-only* and *visual-only* ones. This suggests that incorporating both text and visual content enhances understanding and decision-making processes for users. Users found it easier to understand the information presented in *multi-modal* panes, which consequently helped them make better and faster decisions. This indicates that combining textual and visual elements can improve the effectiveness of clarification panes in assisting users. The study identified that image quality and clarity directly influenced users' preferences. When images were clear and of high quality, users favoured *multi-modal* panes. Therefore, it is crucial to ensure that the visual content provided in clarification panes is of good quality and easily comprehensible. We also showed that in situations where the images were unclear and of low quality, users preferred *text-only* clarification panes, even if the images were relevant. This suggests that when visual content is inadequate, relying solely on text can be more effective in conveying the necessary information. Finally, the research explored the task of automatically generating *multi-modal* clarification panes. The results indicated that text-to-image generation models, such as Stable Diffusion, are capable of producing relevant visual content of high quality. This discovery indicates that automated generation techniques can be utilised to produce *multi-modal* panes for search clarifications, which has the potential to reduce manual work and enhance efficiency. Nonetheless, it is crucial to note that these methods have not yet achieved the ability to completely replicate human annotation when it comes to gathering relevant images for *text-only* clarification panes. Users still exhibit a strong preference for images collected by humans rather than those generated by models.

In our research, we recognise the potential impact of the dataset size. However, the statistically significant differences observed in our analysis form a reliable foundation for drawing valid conclusions. We have utilised robust statistical techniques to ensure the credibility of our findings, and it is unlikely that the observed effects are solely due to random chance.

Chapter 8

Conclusions and Future Work

In the vast digital landscape of search engines, the significance of clarification questions cannot be overstated. Clarification questions often hold the key to unlocking precise and relevant information, yet users often hesitate to engage with them. Understanding the characteristics of engaging clarification questions becomes crucial in enhancing the performance of information retrieval systems. With a deeper comprehension of the importance of clarification questions, we can pave the way for more efficient and satisfying online searches, empowering users to extract the knowledge they seek with ease.

In this thesis, we discussed the progress and challenges of information-seeking systems in assisting users with finding answers to their questions and highlighted the need for clarification questions to handle complex and ambiguous queries effectively.

We described a series of experiments conducted to analyse user behaviour when interacting with clarification questions on various information-seeking platforms. We analysed human-generated clarification questions, categorised them using a novel taxonomy, and investigated their contribution to the original post and accepted answers.

We expanded our study to clarification questions in search engines, using the *MIMICS* dataset and examined the task of selecting the most engaging clarification question from multiple options for a given query and framed it as a learning-to-rank problem. Different models were explored and evaluated using various metrics, including a comparison with a large language model. Limitations of the *MIMICS* dataset were identified, which led to the introduction of a new dataset called *MIMICS-Duo*, which facilitates multi-dimensional evaluation of search clarification. Using this dataset, we further explored the task of

identifying engaging clarification questions and investigating the relationship between online and offline evaluations, which was relatively unexplored in the existing literature.

Finally, we discussed *multi-modal* clarifications in search engines. User preferences for *multi-modal* clarifications were explored, and we demonstrated the use of text-to-image generation systems for generating *multi-modal* clarification questions.

Overall, we analysed clarification questions in information-seeking systems from various angles to understand what makes a clarification question engaging from a user's perspective.

8.1 Thesis Contributions

We now provide a summary of the thesis contributions by chapter.

Chapter 1 – Introduction: We described the reasons behind our motivation for undertaking this thesis and the scope of our research. Additionally, we provided an overview of the challenges within the field of search clarification and highlighted our contributions to addressing these challenges.

Chapter 2 – Background: In this thesis, we furnished the contextual groundwork by examining various aspects related to search clarification. This involved investigating the use of both human- and model-generated clarification questions in community question-answering forums and search engines. We delved into the models used for selecting and generating clarifications, exploring the advancements made in this area. Furthermore, we explored the existing datasets available for clarification tasks, examining their respective advantages and disadvantages. Lastly, we discussed the significance of online and offline evaluations in the field of information retrieval, emphasising the limited understanding and lack of established approaches for evaluating search clarification.

Chapter 3 – Useful Clarification Questions in Community Question Answering Forum: We aimed to answer below research questions in this chapter:

- What clarification questions are more useful (in terms of helping the Asker to get a correct answer)?
- What are the characteristics of useful clarification questions?

We introduced novel taxonomies to investigate the clarification questions and classify Useful clarifications into different types based on user intents and their patterns. The investigation of useful and non-useful clarifications also identified specific patterns that were common in useful clarifications but less prevalent in non-useful ones. Such patterns can be utilised by information-seeking systems to generate and ask clarification questions that are more effective and contextually relevant, enhancing the user experience in information-seeking systems.

Chapter 4 – Asking Engaging Clarification Question in Search Engines: Task Formulation and Limitations: We focused on the task of ranking clarification questions and identifying the most engaging ones for a given query in a search engine and addressed the research questions below:

- Can the SERP feature help us identify the most engaging clarification question from a user’s Perspective?
- Is there any relationship between online and offline evaluations in search clarification using the *MIMICS* dataset (the only search clarification dataset)?

The task was formulated as a learning-to-rank problem, and several SERP and non-SERP features were defined and used as input features for the models. We showed SERP features cannot help identify the most engaging clarification question among multiple clarifications generated for a given query. We also observed that the existing datasets are limited in their ability to provide comprehensive insights into the characteristics of engaging clarifications and user interactions with them. They also cannot facilitate the study of online and offline evaluations in search clarification.

Chapter 5 – Introducing MIMICS-Duo: A Dataset for Online and Offline Evaluation of Search Clarification: To overcome the limitations of available clarification datasets, we introduced *MIMICS-Duo* dataset using the crowd-sourcing approach. This dataset contains both online and offline evaluations for 1,034 query-clarification pairs. Fine-tune crowd-sourcing protocols and experiments were designed and carried out, and high-quality labels were collected. This dataset can be used for various search clarification research, such as exploring the relationship between online and offline evaluations, characteristics of engaging clarifications and clarification generation and selection models.

Chapter 6 – Overall Practices in Online and Offline Evaluations for Search Clarification: We aimed to address the knowledge gap in evaluation methodologies in search clarification. The research questions below were addressed in this chapter:

- What are the best overall practices in designing offline evaluation methodologies for search clarification that correspond with online evaluation?
- Does query length impact the relationship between online and offline evaluations in search clarification?
- Does uncertainty in the online evaluation impact the relationship between online and offline evaluation

We showed that online and offline evaluations support each other in identifying the most engaging clarification for a query. We observed that the offline labels can be utilised

as input features for large language models to predict online user engagement. However, when it comes to ranking multiple clarification questions for a given query, offline evaluations and investigated, models fall short in corresponding with online evaluations. We also investigated the impact of query length and uncertainty in online evaluation on the relationship between online and offline evaluations and discussed the observations.

Chapter 7 – Understanding Modality Preferences in Clarification Questions: We examined how different communication modalities impact user preference; in particular, we aimed to understand if adding a visual presentation to *text-only* clarification panes enhances the user experience. We found that users often favoured *multi-modal* clarification prompts over *text-only* and *visual-only* ones. We introduced *MIMICS-MM* dataset, a *multi-modal* search clarification dataset that was created through fine-tuned crowd-sourcing. This dataset contains both text and images for every candidate answer in a clarification pane for 100 search queries. We emphasised that *multi-modal* clarification panes were easier to comprehend and facilitated better and quicker decision-making for users. The impact of the quality and clarity of the image and the text on users’ preferences was also evaluated. We conducted research on automatically generating *multi-modal* CPs and demonstrated that text-to-image generation models like Stable Diffusion can generate visually relevant modalities of high quality for search clarifications.

8.2 Discussion and Summary

This research has several significant contributions to the field of information retrieval. Here is a short discussion of the overall implications of the main findings:

- **Enhancing User Experience through Effective Clarification Questions:** The introduction of novel taxonomies and identification of patterns in useful clarifications provides a valuable framework for information-seeking systems. This enables them to generate more effective and contextually relevant clarification questions. This has far-reaching implications for user experience, as it can significantly improve the quality of search results and reduce user frustration.
- **Limitations of SERP Features in Identifying Engaging Clarifications:** The observation that SERP features are not effective in identifying the most engaging clarification question highlights a gap in current search practices. This finding calls for a reevaluation of the reliance on SERP features for optimisation and suggests the need for more complex approaches to determining the best clarification.
- **Critical Role of Datasets in Understanding Clarification Characteristics:** The development of the *MIMICS-Duo* dataset is a pivotal contribution, addressing the limitations

of existing datasets. This dataset not only enables a more comprehensive study of clarification characteristics and user interactions but also opens up avenues for research in areas like clarification generation and selection models.

- **Online and Offline Evaluations as Complementary Tools:** The demonstration that online and offline evaluations complement each other in identifying engaging clarifications is a significant insight. It emphasises the need for a holistic approach in evaluating search clarification. The utilisation of offline labels as input features for large language models further expands the potential applications in this domain.
- **Preference for Multi-Modal Clarification Prompts:** The preference for multi-modal clarification prompts over text-only or visual-only prompts highlights the importance of catering to diverse user preferences and cognitive styles. This finding has substantial implications for the design of search interfaces, encouraging the integration of multiple modalities for a more inclusive and effective user experience.

In light of these findings, here are some recommendations for commercial search platforms:

- **Incorporate Taxonomies for Clarification Questions:** Implement the novel taxonomies introduced in this thesis to enhance the generation of clarification questions. This will lead to more accurate and contextually relevant results for users.
- **Diversify Evaluation Metrics for Clarifications:** Recognise the limitations of existing datasets and consider adopting a multi-faceted evaluation approach that combines online and offline metrics. This will provide a more comprehensive understanding of user engagement with clarifications.
- **Leverage Multi-Modal Clarification Prompts:** Integrate multi-modal clarification prompts into search interfaces to cater to diverse user preferences. This can lead to improved comprehension and quicker decision-making for users.
- **Explore Automatic Generation of Multi-Modal Clarification Panes:** Invest in research and development of models, such as Stable Diffusion, for automatically generating high-quality multi-modal clarification panes. This can streamline the process of providing users with visually relevant information.

These recommendations, based on the thesis findings, can significantly advance the effectiveness and user-friendliness of commercial search platforms.

8.3 Future Directions

Although this thesis makes valuable contributions in advancing our understanding of search clarification and the characteristics of engaging clarification questions, there are still intriguing research avenues to explore in order to develop more efficient and effective clarification models for Information Retrieval systems.

Investigating the clarification questions in community question-answering forums showed that generating context-aware clarification can be one further step in this field. Investigating the incorporation of contextual information, such as user profiles, previous interactions, and session history, to generate more context-aware clarification questions can help personalise the clarification process and improve the relevance of the generated questions. The research in this field can be also extended to different domains and information-seeking platforms to assess the generalisability of the findings in terms of the type and patterns of useful clarification questions.

Within this dissertation, we addressed the issue of inadequate search clarification datasets. Although the introduction of the *MIMICS-Duo* and *MIMICS-MM* datasets partially filled this gap, it is important to note that they still possess limitations in terms of their size and scope. Future work should involve the creation of a larger and more diverse dataset. This dataset should cover various aspects of query-clarification pairs, including query intent, topic, clarity, and difficulty. It should also incorporate online user feedback, such as dwell time, clicked documents, mouse hovering, and feedback on usefulness and relevance. A comprehensive dataset would enable more robust evaluations and facilitate the development of effective search clarification models. Future work could involve collecting and analysing more extensive datasets from diverse sources to ensure the generalisability of findings. This could include incorporating data from different domains or platforms to capture a broader range of user interactions and information needs. Based on the conclusions drawn from this study, here are some potential directions for future work:

- Integrating machine learning techniques: Machine learning techniques can be leveraged to improve the accuracy and effectiveness of offline evaluation methodologies. Exploring the use of advanced algorithms, such as large language or deep learning models, for offline evaluation could lead to more robust and reliable evaluation results.
- Expand and refine the evaluation methodologies: The study in Chapter 6 focused on five offline evaluation methodologies, but there is room for exploring additional approaches. Future work could involve the development and testing of new evaluation metrics or the adaptation of existing metrics from related fields. This would help

in obtaining a more comprehensive understanding of search clarification models. Investigating the impact of different types of information provided to annotators and finding ways to improve the correspondence between online and offline evaluations would be valuable.

- Investigate other factors: While the study addressed the impact of query length on the relationship between online and offline evaluation, there are other factors worth exploring. Future research could investigate how query intent, topic, or clarity/difficulty influence the relationship between online and offline evaluations. Understanding these factors would provide deeper insights into the effectiveness of search clarification models.
- Apply the Wizard of Oz approach: Conducting experiments using the Wizard of Oz approach [41], where clarification questions are directly asked from users, can provide valuable insights into what factors contribute to making a clarification engaging. This approach involves simulating the functionality of search clarification models through human operators. By studying user interactions and preferences in this setup, researchers can gain a better understanding of the key elements that make clarifications effective and engaging.
- Improve annotation guidelines: We also mentioned in Chapter 6 that providing more information to annotators can enhance the correspondence between online and offline evaluations. Future work should focus on developing improved annotation guidelines that provide clearer instructions and examples to annotators. Well-defined guidelines would help ensure consistent and reliable annotations, leading to more accurate offline evaluations.
- Explore other user engagement metrics: We focused on evaluating the effectiveness of search clarification models based on user engagement; future research could explore additional metrics. For instance, sentiment analysis could be used to assess user satisfaction or frustration levels. Integrating such metrics into the evaluation framework would provide a more comprehensive understanding of the impact of search clarification on user experience.
- Investigating modality selection: Conducting research to develop advanced *multi-modal* language models that can accurately predict user preferences for different clarification question response modalities. This could involve exploring various factors such as user demographics, task complexity, and content characteristics to determine the most effective modality in different scenarios.

- Improving image generation for clarification modality: Enhancing text-to-image generation models like Stable Diffusion to generate visually relevant modalities with even higher quality. This could involve exploring novel techniques such as conditional adversarial networks or attention mechanisms to produce more realistic and contextually appropriate images.
- Incorporating user feedback: Collecting and leveraging user feedback on generated *multi-modal* clarification questions to refine and optimise the quality and relevance of the generated responses. Use this feedback to iteratively train machine learning models to better align with user preferences and improve the overall user experience.
- Exploring alternative modalities: Considering additional modalities beyond text and images, such as audio or interactive elements, to further enhance the effectiveness and user experience of clarification question responses. This could involve studying the impact of these modalities on user preference, comprehension, and decision-making.
- Conducting longitudinal studies: Performing longitudinal studies to observe how user preferences for clarification question response modalities evolve over time. This could provide insights into any shifts or changes in user behaviour and preferences, allowing for the development of more adaptive and personalised search clarification systems.
- Evaluating real-world deployment: Assessing the performance and user satisfaction of *multi-modal* search clarification systems in real-world deployment scenarios. Conduct user studies and evaluations to understand the practical implications, challenges, and benefits of integrating such systems into existing search or communication platforms.

By focusing on these areas of future work, researchers can further advance the understanding and implementation of *multi-modal* search clarification systems, leading to improved user experiences and more effective communication in various domains.

Bibliography

- [1] The alexa prize taskbot challenge, 2021. URL <https://www.amazon.science/alexa-prize/taskbot-challenge>.
- [2] R. Agrawal, A. Halverson, K. Kenthapadi, N. Mishra, and P. Tsaparas. Generating labels from clicks. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 172–181, 2009.
- [3] A. Al-Maskari and M. Sanderson. The effect of user characteristics on search effectiveness in information retrieval. *Information Processing & Management*, 47(5):719–729, 2011.
- [4] 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*, pages 475–484, 2019.
- [5] M. Aliannejadi, J. Kiseleva, A. Chuklin, J. Dalton, and M. Burtsev. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). *arXiv preprint arXiv:2009.11352*, 2020.
- [6] 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 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, 2021.
- [7] M. Altinkaya and A. W. Smeulders. A dynamic, self supervised, large scale audiovi-

- sual dataset for stuttered speech. In *Proceedings of the 1st International Workshop on Multimodal Conversational AI*, pages 9–13, 2020.
- [8] M. Avriel and A. Williams. The value of information and stochastic programming. *Operations Research*, 18(5):947–954, 1970.
- [9] A. Bampoulidis, J. Palotti, M. Lupu, J. Brassey, and A. Hanbury. Does online evaluation correspond to offline evaluation in query auto completion? In *European Conference on Information Retrieval*, pages 713–719. Springer, 2017.
- [10] J. Beel and S. Langer. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In *International conference on theory and practice of digital libraries*, pages 153–168. Springer, 2015.
- [11] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitingner, and A. Nürnberger. Research paper recommender system evaluation: a quantitative literature survey. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, pages 15–22, 2013.
- [12] M. Bendersky and W. B. Croft. Analysis of long queries in a large scale search log. In *Proceedings of the 2009 workshop on Web Search Click Data*, pages 8–14, 2009.
- [13] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 632–642. Association for Computational Linguistics (ACL), 2015.
- [14] P. Braslavski, D. Savenkov, E. Agichtein, and A. Dubatovka. What do you mean exactly? analyzing clarification questions in cqa. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 345–348, 2017.
- [15] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [16] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.
- [17] Y. T. Cao, S. Rao, and H. Daumé III. Controlling the specificity of clarification question generation. In *WNLP@ ACL*, pages 53–56, 2019.

- [18] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*, pages 1–10, 2009.
- [19] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS)*, 30(1):1–41, 2012.
- [20] Y. Chen, K. Zhou, Y. Liu, M. Zhang, and S. Ma. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 15–24, 2017.
- [21] E. Choi and C. Shah. User motivations for asking questions in online q & a services. *Journal of the Association for Information Science and Technology*, 67(5):1182–1197, 2016.
- [22] K. Christakopoulou, F. Radlinski, and K. Hofmann. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 815–824, 2016.
- [23] A. Chuklin, P. Serdyukov, and M. De Rijke. Click model-based information retrieval metrics. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 493–502, 2013.
- [24] C. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. 1966.
- [25] A. Coden, D. Gruhl, N. Lewis, and P. N. Mendes. Did you mean a or b? supporting clarification dialog for entity disambiguation. In *Joint International Workshop on Summarizing and Presenting Entities and Ontologies, and the International Workshop on Human Semantic Web Interfaces*. CEUR-WS, 2015.
- [26] F. G. Conrad and M. F. Schober. Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly*, 64(1):1–28, 2000.
- [27] P. Cremonesi, F. Garzotto, and R. Turrin. Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(2):1–41, 2012.
- [28] V. Dang. The lemur project-wiki-ranklib. *Lemur Project* (2013), 2013. Available at <https://sourceforge.net/p/lemur/wiki/RankLib>.

- [29] M. De Boni and S. Manandhar. An analysis of clarification dialogue for question answering. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–55, 2003.
- [30] R. Deits, S. Tellex, P. Thaker, D. Simeonov, T. Kollar, and N. Roy. Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction*, 1(1):78–95, 2012.
- [31] Y. Deldjoo, J. Trippas, and H. Zamani. Towards multi-modal conversational information seeking. *Proceedings of SIGIR*, July 2021.
- [32] K. D. Dhole. Resolving intent ambiguities by retrieving discriminative clarifying questions. *arXiv preprint arXiv:2008.07559*, 2020.
- [33] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, and J. A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 161–168, 2014.
- [34] R. Ferreira, D. Silva, D. Tavares, F. Vicente, M. Bonito, G. Gonçalves, R. Margarido, P. Figueiredo, H. Rodrigues, D. Semedo, et al. Twiz: The multimodal conversational task wizard. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6997–6999, 2022.
- [35] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*, 23(2):147–168, 2005.
- [36] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969, 2003.
- [37] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [38] F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. Offline and online evaluation of news recommender systems at swissinfo. ch. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 169–176, 2014.
- [39] B. Geng, L. Yang, C. Xu, X.-S. Hua, and S. Li. The role of attractiveness in web image search. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 63–72, 2011.

- [40] F. Giunchiglia, M. Yatskevich, and P. Shvaiko. Semantic matching: Algorithms and implementation. In *Journal on data semantics IX*, pages 1–38. Springer, 2007.
- [41] B. Hanington and B. Martin. *Universal methods of design expanded and revised: 125 Ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport publishers, 2019.
- [42] J. S. Hare, S. Samangooei, and D. P. Dupplaw. Openimaj and imagerterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 691–694, 2011.
- [43] 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*, pages 1131–1140, 2020.
- [44] H. Hashemi, H. Zamani, and W. B. Croft. Learning multiple intent representations for search queries. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 669–679, 2021.
- [45] S. Huston and W. B. Croft. Evaluating verbose query processing techniques. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298, 2010.
- [46] O. A. S. Ibrahim and E. M. Younis. Hybrid online–offline learning to rank using simulated annealing strategy based on dependent click model. *Knowledge and Information Systems*, pages 1–15, 2022.
- [47] A. Ingber, L. Lewin-Eytan, A. Libov, Y. Maarek, and E. Osherovich. Offline vs. online evaluation in voice product search. In *Proc. 1st International Workshop on Generalization in Information Retrieval (GLARE 2018)*. <http://glare2018.dei.unipd.it/paper/glare2018-paper4.pdf>, 2018.
- [48] 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*, pages 57–66, 2015.
- [49] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, 2002.

- [50] T. Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226, 2006.
- [51] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2):7–es, 2007.
- [52] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, volume 51, pages 4–11. Acm New York, NY, USA, 2017.
- [53] T. Joachims et al. Evaluating retrieval performance using clickthrough data., 2003.
- [54] M. P. Kato, R. W. White, J. Teevan, and S. T. Dumais. Clarifications and question specificity in synchronous social q&a. In *CHI’13 Extended Abstracts on Human Factors in Computing Systems*, pages 913–918. 2013.
- [55] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. In *Acm Sigir Forum*, volume 37, pages 18–28. ACM New York, NY, USA, 2003.
- [56] M. G. Kendall. *Rank Correlation Methods*. Charles Griffin, London, England, 1948.
- [57] J.-K. Kim, G. Wang, S. Lee, and Y.-B. Kim. Deciding whether to ask clarifying questions in large-scale spoken language understanding. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 869–876. IEEE, 2021.
- [58] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Comparing client and server dwell time estimates for click-level satisfaction prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 895–898, 2014.
- [59] Y. Kiyota, S. Kurohashi, and F. Kido. Dialog navigator: A question answering system based on large text knowledge base. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.
- [60] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181, 2009.

- [61] V. Kumar and A. W. Black. Clarq: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301, 2020.
- [62] V. Kumar, V. Raunak, and J. Callan. Ranking clarification questions via natural language inference. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2093–2096, 2020.
- [63] L. T. Le and C. Shah. Retrieving people: Identifying potential answerers in community question-answering. *Journal of the association for information science and technology*, 69(10):1246–1258, 2018.
- [64] S. Li, Y. Abbasi-Yadkori, B. Kveton, S. Muthukrishnan, V. Vinay, and Z. Wen. Offline evaluation of ranking policies with click models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1685–1694, 2018.
- [65] D.-R. Liu, Y.-H. Chen, M. Shen, and P.-J. Lu. Complementary qa network analysis for qa retrieval in social question-answering websites. *Journal of the Association for Information Science and Technology*, 66(1):99–116, 2015.
- [66] J. Liu and R. Yu. State-aware meta-evaluation of evaluation metrics in interactive information retrieval. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3258–3262, 2021.
- [67] Q. Liu, E. Agichtein, G. Dror, E. Gabrilovich, Y. Maarek, D. Pelleg, and I. Szpektor. Predicting web searcher satisfaction with existing community-based answers. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 415–424, 2011.
- [68] Q. Liu, E. Agichtein, G. Dror, Y. Maarek, and I. Szpektor. When web search fails, searchers become askers: understanding the transition. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 801–810, 2012.
- [69] Y. Liu, Y. Fu, M. Zhang, S. Ma, and L. Ru. Automatic search engine performance evaluation with click-through data analysis. In *Proceedings of the 16th international conference on World Wide Web*, pages 1133–1134, 2007.
- [70] 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*, pages 493–502, 2015.
- [71] T. Lotze, S. Klut, M. Aliannejadi, and E. Kanoulas. Ranking clarifying questions based on predicted user engagement. *arXiv preprint arXiv:2103.06192*, 2021.
- [72] B. P. Majumder, S. Rao, M. Galley, and J. McAuley. Ask what’s missing and what’s useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312, 2021.
- [73] J. McAuley and A. Yang. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635, 2016.
- [74] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.
- [75] S. Min, J. Michael, H. Hajishirzi, and L. Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- [76] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):1–27, 2008.
- [77] D. Moldovan and S. Harabagiu. Answering complex list and context questions with lcc’s question-answering server. In *Conference of the Association for Computational Linguistics (ACL-2000)*, pages 563–570, 2000.
- [78] W. Ou and Y. Lin. A clarifying question selection system from ntes_along in convai3 challenge. *arXiv preprint arXiv:2010.14202*, 2020.
- [79] M. O’Brien and M. T. Keane. Modeling result-list searching in the world wide web: The role of relevance topologies and trust bias. In *Proceedings of the 28th annual conference of the cognitive science society*, volume 28, pages 1881–1886. Citeseer, 2006.
- [80] G. Pantazopoulos, J. Bruyere, M. Nikandrou, T. Boissier, S. Hemanthage, B. K. Sachish, V. Shah, C. Dondrup, and O. Lemon. Vica: Combining visual, social, and

- task-oriented conversational ai in a healthcare setting. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 71–79, 2021.
- [81] G. Penha, A. Balan, and C. Hauff. Introducing mantis: a novel multi-domain information seeking dialogues dataset. *arXiv preprint arXiv:1912.04639*, 2019.
- [82] 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*, pages 989–992, 2018.
- [83] C. Qu, L. Yang, C. Chen, M. Qiu, W. B. Croft, and M. Iyyer. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 539–548, 2020.
- [84] S. Quarteroni and S. Manandhar. Designing an interactive open-domain question answering system. *Natural Language Engineering*, 15(1):73, 2009.
- [85] F. Rahutomo, T. Kitasuka, and M. Aritsugi. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1, 2012.
- [86] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [87] S. Rao. Are you asking the right questions? teaching machines to ask clarification questions. In *Proceedings of ACL 2017, Student Research Workshop*, pages 30–35, 2017.
- [88] S. Rao and H. Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. *arXiv preprint arXiv:1805.04655*, 2018.
- [89] S. Rao and H. Daumé III. Answer-based adversarial training for generating clarification questions. In *Proceedings of NAACL-HLT*, pages 143–155, 2019.
- [90] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.
- [91] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- [92] C. Rosset, C. Xiong, X. Song, D. Campos, N. Craswell, S. Tiwary, and P. Bennett. Leading conversational search by suggesting useful questions. In *Proceedings of The Web Conference 2020*, pages 1160–1170, 2020.
- [93] M. Rossetti, F. Stella, and M. Zanker. Contrasting offline and online results when evaluating recommendation algorithms. In *Proceedings of the 10th ACM conference on recommender systems*, pages 31–34, 2016.
- [94] A. Said and A. Bellogín. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 129–136, 2014.
- [95] I. Sekulić, 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, March 28–April 1, 2021, Proceedings, Part I 43*, pages 619–633. Springer, 2021.
- [96] C. Shah, S. Oh, and J. S. Oh. Research agenda for social q&a. *Library & Information Science Research*, 31(4):205–209, 2009.
- [97] V. Shwartz, P. West, R. L. Bras, C. Bhagavatula, and Y. Choi. Unsupervised commonsense question answering with self-talk. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- [98] A. Srinivasan and V. Setlur. Snowy: Recommending utterances for conversational visual analysis. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 864–880, 2021.
- [99] A. Stoll. Post hoc tests: Tukey honestly significant difference test. *The SAGE encyclopedia of communication research methods*, pages 1306–1307, 2017.
- [100] S. Stoyanchev, A. Liu, and J. Hirschberg. Towards natural clarification questions in dialogue systems. In *AISB symposium on questions, discourse and dialogue*, volume 20, 2014.
- [101] A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823. PMLR, 2015.

- [102] R. F. Tate. Correlation between a discrete and a continuous variable. point-biserial correlation. *The Annals of mathematical statistics*, 25(3):603–607, 1954.
- [103] L. Tavakoli. Generating clarifying questions in conversational search systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3253–3256, 2020.
- [104] L. Tavakoli, H. Zamani, F. Scholer, W. B. Croft, and M. Sanderson. Analyzing clarification in asynchronous information-seeking conversations. *Journal of the Association for Information Science and Technology*, 2021.
- [105] J. Teevan, S. T. Dumais, and E. Horvitz. Characterizing the value of personalizing search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 757–758, 2007.
- [106] C. Trattner, D. Moesslang, and D. Elswiler. On the predictability of the popularity of online recipes. *EPJ Data Science*, 7(1):1–39, 2018.
- [107] P. Tseng et al. Coordinate ascent for maximizing nondifferentiable concave functions. 1988.
- [108] J. W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.
- [109] P. van Beek, R. Cohen, and K. Schmidt. From plan critiquing to clarification dialogue for cooperative response generation. *Computational Intelligence*, 9(2):132–154, 1993.
- [110] E. M. Voorhees. Overview of the trec 2001 question answering track. In *In Proceedings of the Tenth Text REtrieval Conference (TREC*. Citeseer, 2001.
- [111] E. M. Voorhees, D. K. Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 63. Citeseer, 2005.
- [112] J. Wang and W. Li. Template-guided clarifying question generation for web search clarification. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3468–3472, 2021.
- [113] X. Wang, M. Bendersky, D. Metzler, and M. Najork. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 115–124, 2016.

- [114] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.
- [115] C. Wissler. The spearman correlation formula. *Science*, 22(558):309–311, 1905.
- [116] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, 2010.
- [117] Z. Wu, B. Kao, T.-H. Wu, P. Yin, and Q. Liu. Perq: Predicting, explaining, and rectifying failed questions in kb-qa systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 663–671, 2020.
- [118] J. Xu, Y. Wang, D. Tang, N. Duan, P. Yang, Q. Zeng, M. Zhou, and S. Xu. 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)*, pages 1618–1629, 2019.
- [119] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 73–80, 2007.
- [120] J. Yi, Y. Chen, J. Li, S. Sett, and T. W. Yan. Predictive model performance: Offline and online evaluations. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1294–1302, 2013.
- [121] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and effort: An analysis of document utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 91–100, 2014.
- [122] H. Zamani, S. Dumais, N. Craswell, P. Bennett, and G. Lueck. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, pages 418–428, 2020.
- [123] H. Zamani, S. Dumais, N. Craswell, P. Bennett, and G. Lueck. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020, WWW '20*, page 418–428, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380126.

- [124] H. Zamani, G. Lueck, E. Chen, R. Quispe, F. Luu, and N. Craswell. Mimics: A large-scale data collection for search clarification. In *Proceedings of CIKM*, page 3189–3196, 2020. ISBN 9781450368599.
- [125] H. Zamani, G. Lueck, E. Chen, R. Quispe, F. Luu, and N. Craswell. Mimics: A large-scale data collection for search clarification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3189–3196, 2020.
- [126] 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 SIGIR*, page 1181–1190, 2020.
- [127] 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*, pages 1181–1190, 2020.
- [128] H. Zamani, J. R. Trippas, J. Dalton, and F. Radlinski. Conversational information seeking. *ArXiv*, abs/2201.08808, 2022.
- [129] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 15–24, 2009.
- [130] F. Zhang, K. Zhou, Y. Shao, C. Luo, M. Zhang, and S. Ma. How well do offline and online evaluation metrics measure user satisfaction in web image search? In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 615–624, 2018.
- [131] Y. Zhang, X. Chen, Q. Ai, L. Yang, and W. B. Croft. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 177–186, 2018.
- [132] Z. Zhang and K. Zhu. Diverse and specific clarification question generation with keywords. In *Proceedings of the Web Conference 2021*, pages 3501–3511, 2021.
- [133] H. Zheng, D. Wang, Q. Zhang, H. Li, and T. Yang. Do clicks measure recommendation relevancy? an empirical user study. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 249–252, 2010.

- [134] 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*, pages 369–378, 2019.
- [135] J. Zou, E. Kanoulas, and Y. Liu. An empirical study of clarifying question-based systems. In *29th ACM International Conference on Information and Knowledge Management (CIKM)*, 2020.

Appendices

Appendix A

Instructions and Examples of Crowd-sourcing Tasks

TASK 1 – Offline Rating

Page 1: This page provide an introduction about the survey, essential information about IRB approval and contact info, participant information and some information about the number of questions and pages in this survey.

0% ————— 100%

Thanks for choosing this survey. This survey aims to enhance user satisfaction when searching a query (the information need) on Google. This is a research project being conducted by Ms. Leila Tavakoli at RMIT University. Your participation in this research study is voluntary. **This is the second round of this study and you are allowed to participate.**

The procedure involves filling out an online survey that will take a few minutes. Your responses will be confidential and we do not collect identifying information such as your name, gender, age or email address. All data is stored in an RMIT password-protected electronic format. To help protect your confidentiality, the surveys will not contain information that will personally identify you. The results of this study will be used for scholarly purposes only and may be shared with RMIT University representatives.

If you have any questions about the research, please contact Ms. Leila Tavakoli via email at leila.tavakoli@rmit.edu.au.

This research has been reviewed according to RMIT University IRB procedures for research involving human subjects. The IRB approval number is 66-19/22334. Please read the full participant information details which are available here:

[Participant Information Sheet](#)

Please read the following instruction and answer 3 questions given in 5 pages to complete the task. **You have 1 hour to complete this survey. Loading the next page may take a few seconds.**

By taking this survey you agreed with this electronic consent. Please note that there are several attention check questions in the task. If you do not answer them correctly, you will NOT be compensated.

Page 2: This page provide the instruction of the survey and the steps which need to be done by the workers. There is also a CAPTCHA to detect

Survey Completion

0%100%

Introduction


Assume whenever Google cannot understand the information you request from the query you submit, it asks you a multi-choice clarification question. By answering this clarification question, you receive the information you need much faster.

In this study, we are going to assess the quality of clarification questions generated by the Google search engine for given queries from the users' points of view. The study will be conducted in two sequential steps:

- In the first step, you are given a query submitted to Google and the top search results shown by Google. You are expected to review all search results and answer one question.
- In the second step, several multi-choice clarification questions generated by Google are shown. You are expected to judge the clarification questions and their answers.

Check the box to go to the next page.

☐ I'm not a robot


reCAPTCHA
[Privacy](#) - [Terms](#)

Page 3: To ensure workers pay attention to the different aspects of the query and the document summaries, we show them eight relevant summaries and one irrelevant summary in addition to the query. Workers are then asked to identify the irrelevant document summary, which is placed in a random rank position for each task. For this example, document #8 is irrelevant.

Survey Completion

0%100%

Assume you have googled a query and Google has shown you several results on the first page. The query and the search results are shown below. Please read the results carefully. (Some results might be blank)

Query:

12 gauge shotgun

Result #1:
Shotguns Bass Pro Shops
Browse for the exciting deals on our wide selection of 12 b gauge b shotguns semi automatic pump action and more at Bass Pro Shops online or in store near you

Result #2:
12 Gauge Semi Automatic Shotguns Cheaper Than Dirt
EAA GIRSAN MC312 12 b Gauge Semi Auto Shotgun 28 quot Vent Rib Barrel 3 1 2 quot Chamber EAA GIRSAN MC312 12 b Gauge Semi Auto Shotgun 28 quot Ve Our Low Price 396 57

Result #3:
12 Gauge amp 20 Gauge Pump Shotguns Academy
Inherently more reliable under adverse conditions such as dirt and sand exposure and climate extremes than other types of shotguns pump shotguns are available in 12 b gauge 20 gauge 28 gauge and 410 bore

Result #4:
12 Gauge Pump Action Shotguns Cheaper Than Dirt
Shop Cheaper Than Dirt for wide variety of 12 b gauge pump action shotguns from top brands like Mossberg Winchester Remington Stevens and more Due to High Order Volumes Expect Delays in Processing Orders For the Fastest Delivery select Express Shipping at Checkout

Result #5:
Shotguns for Sale Father 39 Day Deals at DICK 39 S
Powerful 12 b gauge weapons are the most popular choice However for youth and smaller frame shooters or someone who spends hours shooting skeet the reduced recoil of 20 gauge may be better choice Once you 39 ve decided on gauge make choice from the following styles Pump Action classic and popular choice Easy to operate and reliable

Result #6:
12 Gauge Semi Automatic Shotguns For Sale Hinterland
Shop online for the best selection and prices of 12 b Gauge Semi Auto Shotguns at Hinterland Outfitters from top brands like Browning Remington Winchester Savage and many more We also carry full range of shotgun shells and accessories

Result #7:
12ga Shotguns Walmart com
Product Title Tetra ValuPro III 12 Gauge Shotgun Cleaning Kit 750i Average rating 0 out of 5 stars based on 0 reviews Current Price 25 56 25 56 List Price 28 95 28 95

Result #8:
Free Bill of Sale Forms PDF Word doc
Free b Bill b of Sale b Forms PDF Templates b bill b of sale represents receipt for an exchange of goods between two 2 parties buyer and seller The buyer offers cash or trade to seller for personal property with the most popular being vehicles

Result #9:
Mossberg 12 GA Shotguns For Sale Vance Outdoors
Mossberg 500 Turkey 12 b Gauge Pump Action Shotgun with Mossy Oak Obsession Finish 429 99 In Stock Brand Mossberg Item Number 52280 Mossberg 88 Special Purpose 12 Gauge Shotgun 229 99 Notify Me When Available Brand Mossberg Item Number 31046 Mossberg 88 Maverick SP 12 GA 20 Inch Pump Action Shotgun

Question 1: One result does not belong to the given query and is irrelevant. Please write down the number of the irrelevant result in the box below.

144

Page 4: In this page we show all generated clarification panes for the shown query in the previous page and ask the workers to write down the number of clarification pane they see. This is an attention check question to ensure then workers does not answer the questions randomly. The correct answer for this question is 4.

Survey Completion

0%100%

To understand your query accurately, Google has generated several multi-choice clarification questions.
(Note: There might be some multi-choice clarification questions with only 2 or 3 answers.)
(Note: Rotate your screen to landscape if you do it on your mobile.)

Query:

12 gauge shotgun

Multi-choice clarification question #1

Select one to refine your search

12 gauge fully automatic
shotgun

single
shot

12 gauge revolver
shotgun

Multi-choice clarification question #2

Select one to refine your search

lever action

bolt action

single shot

pump action

muzzleloader

Multi-choice clarification question #3

Select one to refine your search

bolt action

lever action

single shot

Multi-choice clarification question #4

Select one to refine your search

savage

browning

mossberg

winchester

cز

Question 2: How many Multi-choice clarification questions have been generated for the shown query? **(Note: Some multi-choice clarification questions might be blank, do NOT count them.)**

Page 5: In this page we ask the workers to rate all shown clarification panes for the given query based on their preferences in star-rating format.

Survey Completion

0%100%

Question 3: Knowing the query and its different aspects, please rate the multi-choice clarification questions shown below based on your preference for the given query.
(Note: 5 start means very good and 1 star means very bad.)
(Note: Rotate your screen to landscape if you do it on your mobile.)

Query:

12 gauge shotgun

Select one to refine your search

12 gauge fully automatic shotgun

single shot

12 gauge revolver shotgun

Multi-choice clarification question 1

★

★

★

★

★

Select one to refine your search

lever action

bolt action

single shot

pump action

muzzleloader

Multi-choice clarification question 2

★

★

★

★

★

Select one to refine your search

bolt action

lever action

single shot

Multi-choice clarification question 3

★

★

★

★

★

Select one to refine your search

savage

browning

mossberg

winchester

cz

Multi-choice clarification question 4

★

★

★

★

★

Final Page: This is the last page and we ask workers to take the randomly generated code shown here in their submission on Mechanical Turk. This the final stage of quality check.

Survey Completion

0%100%

Thank you for your time.
Please provide any feedback. It will help us to improve this survey.

Here is your Mechanical Turk code:

CQ-95869-CC

←

Submit

TASK 2 – Quality Labelling

Page 1: This page provide an introduction about the survey, essential information about IRB approval and contact info, participant information and some information about the nnumber of questions and pages in this survey.

Survey Completion
0% ————— 100%

Thanks for choosing this survey. This survey aims to enhance user satisfaction when searching a query (the information need) on Google. This is a research project being conducted by Ms. Leila Tavakoli at RMIT University. Your participation in this research study is voluntary. **This is the third round of this study and you are allowed to participate.**

The procedure involves filling out an online survey that will take a few minutes. Your responses will be confidential and we do not collect identifying information such as your name, gender, age or email address. All data is stored in an RMIT password-protected electronic format. To help protect your confidentiality, the surveys will not contain information that will personally identify you. The results of this study will be used for scholarly purposes only and may be shared with RMIT University representatives.

If you have any questions about the research, please contact Ms. Leila Tavakoli via email at leila.tavakoli@rmit.edu.au.

This research has been reviewed according to RMIT University IRB procedures for research involving human subjects. The IRB approval number is 66-19/22334. Please read the full participant information details which are available here:

[Participant Information Sheet](#)

Please read the following instruction and answer 4 questions given in 6 pages to complete the task. You have 1 hour to complete this survey. **Loading the next page may take a few seconds.**

By taking this survey you agreed with this electronic consent. Please note that there are several attention check questions in the task. If you do not answer them correctly, you will NOT be compensated.

Page 2: This page provide the instruction of the survey and the steps which need to be done by the workers. There is also a CAPTCHA to detect

Survey Completion

0%100%

Introduction


Assume whenever Google cannot understand the information you request from the query you submit, it asks you a multi-choice clarification question. By answering this clarification question, you receive the information you need much faster.

In this study, we are going to assess the quality of clarification questions generated by the Google search engine for given queries from the users' points of view. The study will be conducted in two sequential steps:

- In the first step, you are given a query submitted to Google and the search result shown by Google. You are expected to review all search results and answer one question.
- In the second step, a multi-choice clarification question asked by Google is shown. You are expected to judge the clarification question and its answers.

Check the box to go to the next page.

☐ I'm not a robot


reCAPTCHA
[Privacy](#) - [Terms](#)

Page 3: To ensure workers pay attention to the different aspects of the query and the document summaries, we show them eight relevant summaries and one irrelevant summary in addition to the query. Workers are then asked to identify the irrelevant document summary, which is placed in a random rank position for each task. For this example, document #4 is irrelevant.

Survey Completion
0% ————— 100%

Assume you have googled a query and Google has shown you several results on the first page. The query and the search results are shown below. Please read the results carefully. (Some results might be blank)

Query:

create a drop down list in excel

Result #1:

How to Create Drop Down List in Excel with Pictures

This wikiHow teaches you how to create a drop down list in Microsoft Excel spreadsheet using computer. This feature allows you to create a list of items to choose from and insert a drop down selector into any empty cell on your spreadsheet. The drop down feature is only available on desktop versions of Excel.

Result #2:

Create drop down list Office Support

After you create your drop down list, make sure it works the way you want. For example, you might want to check to see if you change the column width and row height to show all your entries. If the list of entries for your drop down list is on another worksheet and you want to prevent users from seeing it or making changes, consider hiding and protecting that worksheet.

Result #3:

Create Drop down List in Excel Easy Excel Tutorial

Create a Drop down List. To create a drop down list in Excel, execute the following steps: 1. On the second sheet, type the items you want to appear in the drop down list. Note: If you don't want users to access the items on Sheet2, you can hide Sheet2. To achieve this, right-click on the sheet tab of Sheet2 and click on Hide.

Result #4:

Wizards of the Coast

Producer of hobby gaming systems including Magic: The Gathering, Dungeons and Dragons, and Star Wars TCG. Has detailed game information, books, events, company history, and news.

Result #5:

How to add drop down list to an Excel cell TechRepublic

To the right, you see labels and formats in preparation for creating the Excel drop down list. To create the Region list, do the following: Select H2. Click the Data tab and then click Data.

Result #6:

How to create drop down list in Excel to manage data

The best practice is to create a separate worksheet for your drop down list. To create a new tab, click the '+' icon next to the last tab in your spreadsheet. Double-click the tab to rename it.

Result #7:

How to create drop down list in Excel TechRadar

Creating a simple drop down list in Excel might sound a bit intimidating at first, but it's actually very simple. It's so simple in fact that even an elementary school kid can do it all.

Result #8:

How to Create Drop Down List in Excel in 60 Seconds or

The drop down list is a great way to seem like a superuser and impress your co-workers and boss. At the same time, it's a very user-friendly asset in almost all custom-made Excel sheets. In this tutorial, I'm going to show you the 5 steps to create a drop down in 1 minute or less. Call it the 1 Minute Drop Down.

Result #9:

Excel: How to create simple and dependent drop down lists

Drop down lists in Excel let you create a list of valid choices that you can select for a given field. We'll show you how to use tables, named ranges, formulas, data validation, and table styles.

Question 1: One result does not belong to the given query and is **irrelevant**. Please write down the number of the irrelevant result in the box below.

Page 4: In this page we show the workers the query and a clarification pane (question and candidate answers) and ask the workers to write down the number of candidate answer they see. This is an attention check question to ensure then workers does not answer the questions randomly. The correct answer for this question is 4.

Survey Completion
 0% 100%

To understand your query accurately, Google has asked you a multi-choice clarification question as below.
 (Note: Rotate your screen to landscape if you do it on your mobile.)

Query:
 create a drop down list in excel

Multi-choice clarification question:

| | | | | |
|-------------------------------|--|---------------------------------------|------------|------------|
| Clarification Question | What version of Excel are you looking for? | | | |
| Answers | create a drop down list in excel 2016 | create a drop down list in excel 2010 | excel 2013 | excel 2007 |

Question 2: How many answers are there for this multi-choice clarification question?

Page 5: In this page we ask the workers to label the quality of each candidate answer individually.

Survey Completion

0%100%

Question 3: Given the query, how do you rate the quality of each multi-choice clarification answer?
(Note: Rotate your screen to landscape if you do it on your mobile.)

Query:

create a drop down list in excel

Multi-choice clarification question:

| What version of Excel are you looking for? | | | |
|--|---------------------------------------|------------|------------|
| create a drop down list in excel 2016 | create a drop down list in excel 2010 | excel 2013 | excel 2007 |

| | Very Good Quality | Good Quality | Fair Quality | Bad Quality | Very Bad Quality |
|---------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| create a drop down list in excel 2016 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| create a drop down list in excel 2010 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| excel 2013 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| excel 2007 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Page 6: In this page we ask the workers to label the overall quality of the clarification pane (clarification question and candidate answers altogether).

Survey Completion

0%100%

Question 4: If you want to give an overall quality label to this multi-choice clarification question (question and answers altogether), what would be the quality label?
(Note: Rotate your screen to landscape if you do it on your mobile.)

Query:

create a drop down list in excel

Multi-choice clarification question:

| What version of Excel are you looking for? | | | |
|--|---------------------------------------|------------|------------|
| create a drop down list in excel 2016 | create a drop down list in excel 2010 | excel 2013 | excel 2007 |

☐ Very Good

☐ Good

☐ Fair

☐ Bad

☐ Very Bad

Final Page: This is the last page and we ask workers to take the randomly generated code shown here in their submission on Mechanical Turk. This the final stage of quality check.

Survey Completion

0%100%

Thank you for your time.
Please provide any feedback. It will help us to improve this survey.

Here is your Mechanical Turk code:


CQ-77633-CC

←

Submit

TASK 3 – Aspect Labelling

Page 1: This page provide an introduction about the survey, essential information about IRB approval and contact info, participant information and some information about the nnumber of questions and pages in this survey.

Survey Completion
0%  100%

Thanks for choosing this survey. This survey aims to enhance user satisfaction when searching a query on Google. This is a research project being conducted by Ms. Leila Tavakoli at RMIT University. Your participation in this research study is voluntary. **This is the first round of this study and you are allowed to participate.**

The procedure involves filling out an online survey that will take a few minutes. Your responses will be confidential and we do not collect identifying information such as your name, gender, age or email address. All data is stored in an RMIT password-protected electronic format. To help protect your confidentiality, the surveys will not contain information that will personally identify you. The results of this study will be used for scholarly purposes only and may be shared with RMIT University representatives.

If you have any questions about the research, please contact Ms. Leila Tavakoli via email at leila.tavakoli@rmit.edu.au.

This research has been reviewed according to RMIT University IRB procedures for research involving human subjects. The IRB approval number is 66-19/22334. Please read the full participant information details which are available here:


[Participant Information Sheet](#)

Please read the following instruction and answer 6 questions given in 9 pages to complete the task. Loading the next page may take a few seconds.

By taking this survey you agreed with this electronic consent. Please note that there are several attention check questions in the task. If you do not answer them correctly, you will NOT be compensated.

Page 2: This page provide the instruction of the survey and the steps which need to be done by the workers. There is also a CAPTCHA to detect

Survey Completion

0%  100%


Introduction

Assume whenever Google cannot understand the information you request from the query you submit, it asks you a multi-choice clarification question. By answering this clarification question, you receive the information you need much faster.

In this study, we are going to assess the quality of clarification questions generated by the Google search engine for given queries from the users' points of view. The study will be conducted in two sequential steps:

- In the first step, you are given a query submitted to Google and the search result shown by Google. You are expected to review all search results and answer one question.
- In the second step, a multi-choice clarification question asked by Google is shown. You are expected to judge the clarification question and its potential answers in terms of different aspects.

Check the box to go to the next page.

☐ I'm not a robot 
reCAPTCHA
[Privacy](#) - [Terms](#)

Page 3: To ensure workers pay attention to the different aspects of the query and the document summaries, we show them eight relevant summaries and one irrelevant summary in addition to the query. Workers are then asked to identify the irrelevant document summary, which is placed in a random rank position for each task. For this example, document #6 is irrelevant.

Survey Completion

0%100%

Assume you have googled a query and Google has shown you several results on the first page. The query and the search results are shown below. Please read the results carefully. (Some results might be blank)

Query:

yucca

Result #1:
Yucca Wikipedia
Yucca is genus of perennial shrubs and trees in the family Asparagaceae subfamily Agavoideae Its 40 50 species are notable for their rosettes of evergreen tough sword shaped leaves and large terminal panicles of white or whitish flowers They are native to the hot and dry parts of the Americas and the Caribbean Early reports of the species were confused with the cassava Manihot esculenta

Result #2:
Yucca Uses Side Effects Interactions Dosage and Warning
Yucca is the common name for the more than 40 species of plants in the Yucca genus The root of the non flowering plant is used to make medicine Yucca is used for osteoarthritis high blood

Result #3:
Yucca Better Homes amp Gardens
The yucca plant has co evolved alongside several species of moths their symbiotic relationship benefits both plants and moths The yucca emits fragrance at night to attract the moths to pollinate them As the moths begin to mate the female finds freshly opened bloom and works her way down to the ovary of the flower

Result #4:
The Health Benefits of Yucca
Yucca shouldn be confused with yuca which is root vegetable also known as cassava Yucca offers numerous health benefits and is often used medicinally Parts of the yucca plant can be

Result #5:
Yucca Uses Benefits amp Side Effects Drugs com Herbal
Yucca plants are characterized by stiff evergreen sword shaped leaves crowded on stout trunk There is dense terminal flowerhead faintly resembling candle The flowers are white or greenish All yucca plants depend for pollination on nocturnal yucca moths Each variety of moth is adapted to single species of yucca Scientific Name

Result #6:
Amazon com 22 inch smart tv
VIZIO Series 24 Inch Class 1080p Full HD LED Smart TV D24F G1 with Built in HDMI USB SmartCast Voice Control Bundle with XRYX 6 5 ft HDMI Cable and Accessories 3 8 out of 5 stars 14 199 99 199 99

Result #7:
How to Grow and Care for Yucca Plants Garden Design
Yucca aloifolia Commonly called Spanish bayonet or aloe yucca Zones 8 11 Height Spread Up to 20 feet tall 4 feet wide and with clumps to 10 feet wide Exposure Partial to full sun Bloom time Early summer Color Green leaves cream to white flowers The leaves of this yucca are extremely sharp hence the common name of Spanish bayonet

Result #8:
Yucca Growing How to Care for Yucca Plants Outside
Yucca Growing Outdoors As it is native of the southwestern United States yucca thrives in soil that drains well and can be in full sun It is also able to withstand temperatures as cold as 10 12 so you can grow b yucca plant in many different climates

Result #9:
Yuca Recipes Food Network Food Network
This hardy root vegetable fries up wonderfully Find recipes for yuca fries chips and more

Question 1: One result does not belong to the given query and is irrelevant. Please write down the number of the irrelevant result in the box below.

Page 4: In this page, we show the workers the query and a clarification pane that is generated for the given query and inform the workers that they will answer a few questions about it in next few pages.

Survey Completion

0%100%

To understand your query accurately, Google has asked you a multi-choice clarification question as below.

Please answer the questions in the next pages. (Note: Rotate your screen to landscape if you do it on your mobile.)

Query:

yucca

Multi-choice clarification question:

| Select one to refine your search | | | | |
|----------------------------------|----------------|--------------|------------|--------------|
| yucca valley | yucca mountain | yucca desert | yucca lake | yucca canyon |

Page 5: In this page, we ask the workers to provide their opinions about “*coverage*” of the clarification pane. We give them the definition of the *coverage* and show them a clarification pane with high coverage and a clarification pane with low coverage as examples.

Survey Completion


0% 100%

Checking the *coverage* of the multi-choice clarification question.
(Note: Rotate your screen to landscape if you do it on your mobile.)

Definition: A multi-choice clarification question has high coverage if its potential answers cover every potential aspect of the query.

Examples:
[more](#)

*Example of a multi-choice clarification question with **high coverage**:*



Zyprexa side effects

← Query 🔊 📷 🔍

Select one to refine your search

← Clarification question

In elderly

In men


In women

In children

In adults

← Candidate Answer

*Example of a multi-choice clarification question with **low coverage**:*



Zyprexa side effects

← Query 🔊 📷 🔍

Select one to refine your search

← Clarification question

In men

In women

← Candidate Answer

Question 2: Does the multi-choice clarification question have a **high coverage** for the given query?

Query:
yucca

Multi-choice clarification question:

| Select one to refine your search | | | | |
|----------------------------------|----------------|--------------|------------|--------------|
| yucca valley | yucca mountain | yucca desert | yucca lake | yucca canyon |

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

Page 6: In this page, we ask the workers to provide their opinions about “*diversity*” of the clarification pane. We give them the definition of the *diversity* and show them a clarification pane with high diversity and a clarification pane with low diversity as examples.

Survey Completion

0%100%

Checking the **diversity** of the multi-choice clarification question.
(Note: Rotate your screen to landscape if you do it on your mobile.)

Definition: A multi-choice clarification question has high diversity if its potential answers do not contain redundant information.

Examples:
[more](#)

Example of a multi-choice clarification question with **high diversity**:

Google

Gift for grandfather

Query

Select one to refine your search

birthdayChristmas

Clarification question

Example of a multi-choice clarification question with **low diversity**:

Google

reading quote

Query

Select one to refine your search

Clarification question

babykids

Candidate Answer

Question 3: Does the multi-choice clarification question have a **high diversity** for the given query?

Query:

yucca

Multi-choice clarification question:

| Select one to refine your search | | | | |
|----------------------------------|----------------|--------------|------------|--------------|
| yucca valley | yucca mountain | yucca desert | yucca lake | yucca canyon |

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

Page 7: In this page, we ask the workers an attention check question to ensure that the workers do not answer the questions randomly.

Survey Completion

0%100%

Question 4: Today is Friday. Which day was yesterday?

☐ Thursday

☐ Saturday

☐ Sunday

☐ Tuesday

Page 8: In this page, we ask the workers to provide their opinions about “*understandability*” of the clarification pane. We give them the definition of the *understandability* and show them an understandable clarification pane and a non-understandable clarification pane as examples.

Survey Completion
 0% 100%

*Checking the **understandability** of the multi-choice clarification question.*
 (Note: Rotate your screen to landscape if you do it on your mobile.)

Definition: A multi-choice clarification question is understandable if it is easily readable and meaningful. (Assume this is the query you have submitted and you understand the meaning of your query. Now the question is to check whether the multi-choice clarification question is understandable for your query or not.)

Examples:
[more](#)

*Example of a multi-choice clarification question with **understandable**:*

Google florsheim shoes ← Query

Florsheim shoes for whom? ← Clarification question

florsheim shoes for men florsheim shoes for women ← Candidate Answer

*Example of a multi-choice clarification question with **non-understandable**:*

Google Words with the letters ← Query

Select one to refine your search ← Clarification question

6 7 ← Candidate Answer

Question 5: Is the multi-choice clarification question **understandable** for the given query?

Query:
 yucca

Multi-choice clarification question:

| Select one to refine your search | | | | |
|----------------------------------|----------------|--------------|------------|--------------|
| yucca valley | yucca mountain | yucca desert | yucca lake | yucca canyon |

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

Page 9: In this page, we ask the workers to provide their opinions about “*importance order*” of the candidate answers in a clarification pane. We give them the definition of the *importance order* and show them a clarification pane with correct order for candidate answers and a clarification pane with incorrect order for the candidate answer as examples.

Survey Completion


0% 100%

Checking the *importance order* of the multi-choice clarification question.
(Note: Rotate your screen to landscape if you do it on your mobile.)




Definition: A multi-choice clarification question has the correct order if the most relevant and important answers to the different aspects of the query are positioned from left to right.

Examples:
[more](#)

Example of a multi-choice clarification question with *correct order*:



Open gift file ← Query



Select one to refine your search

← Clarification question

Windows 10

Windows 8

Windows 7

Windows vista

Windows xp

← Candidate Answer



Open gift file ← Query



Select one to refine your search

← Clarification question

Windows 7

Windows xp

Windows 10

Windows vista

Windows 8

← Candidate Answer

Query:
yucca

Select one to refine your search

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

Final Page: This is the last page and we ask workers to take the randomly generated code shown here in their submission on Mechanical Turk. This the final stage of quality check.

Survey Completion

0%100%

Thank you for your time.
Please provide your feedback. It will help us to improve this survey.

Here is your Mechanical Turk code:

CQ-73675-CC

←

Submit

Appendix B

Publications

Publications used in this thesis:

Doctoral Consortium

Tavakoli, L. (2020, October). Generating clarifying questions in conversational search systems. *In Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3253-3256.

Journal Paper

Tavakoli, L., Zamani, H., Scholer, F., Croft, W. B., & Sanderson, M. (2022). Analyzing clarification in asynchronous information-seeking conversations. *Journal of the Association for Information Science and Technology*, 73(3), pp. 449-471.

Conference Paper

Tavakoli, L., Trippas, J. R., Zamani, H., Scholer, F., & Sanderson, M. (2022, July). MIMICS-Duo: Offline & Online Evaluation of Search Clarification. *In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3198-3208.

Additional Publications:

Throughout the duration of this PhD, I contributed to two publications; however, they have not been included in this thesis since they are not directly relevant to the research topic.

Conference Papers

Cambazoglu, B. B., **Tavakoli, L.**, Scholer, F., Sanderson, M., Croft, B. (2021, March). An intent taxonomy for questions asked in web search. *In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pp. 85-94.

Cambazoglu, B. B., Baranova, V., Scholer, F., Sanderson, M., **Tavakoli, L.**, Croft, B. (2021, March). Quantifying Human-Perceived Answer Utility in Non-factoid Question Answering. *In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pp. 75-84.

Appendix C

Ethics Approval Letter

Notice of Approval

Date: **16 October 2019**

Project number: **66-19/22334**

Project title: **Information storage, analysis and retrieval**

Risk classification: **Negligible risk**

Chief investigator: **Professor Mark Sanderson**

Status: **Approved**

Approval period: From: **16/10/2019** To: **16/10/2022**

The following documents have been reviewed and approved:

| Title | Version | Date |
|--|---------|----------------|
| Risk Assessment and Application Form | 2 | 29 August 2019 |
| Participant Information Sheet and Consent Form | 2 | 29 August 2019 |

The above application has been approved by the RMIT University CHEAN as it meets the requirements of the *National Statement on Ethical Conduct in Human Research* (NHMRC, 2007).

Terms of approval:**1. Responsibilities of chief investigator**

It is the responsibility of the above chief investigator to ensure that all other investigators and staff on a project are aware of the terms of approval and to ensure that the project is conducted as approved by CHEAN. Approval is valid only whilst the chief investigator holds a position at RMIT University.

2. Amendments

Approval must be sought from CHEAN to amend any aspect of a project. To apply for an amendment, use the request for amendment form, which is available on the HREC website and submitted to the CHEAN secretary. Amendments must not be implemented without first gaining approval from CHEAN.

3. Adverse events

You should notify the CHEAN immediately (within 24 hours) of any serious or unanticipated adverse effects of their research on participants, and unforeseen events that might affect the ethical acceptability of the project.

4. Annual reports

Continued approval of this project is dependent on the submission of an annual report. Annual reports must be submitted by the anniversary of approval of the project for each full year of the project. If the project is of less than 12 months duration, then a final report only is required.

5. Final report

A final report must be provided within six months of the end of the project. CHEAN must be notified if the project is discontinued before the expected date of completion.

6. Monitoring

Projects may be subject to an audit or any other form of monitoring by the CHEAN at any time.

7. Retention and storage of data

The investigator is responsible for the storage and retention of original data according to the requirements of the *Australian Code for the Responsible Conduct of Research* (R22) and relevant RMIT policies.

8. Special conditions of approval

Nil.

In any future correspondence please quote the project number and project title above.

Yours sincerely,

Dr Lauren Saling
Deputy Chair, Science Engineering & Health
College Human Ethics Advisory Network