

Asking Clarifying Questions: to Benefit or to Disturb Users in Web Search?

Jie Zou^{a,b}, Aixin Sun^{a,b,*}, Cheng Long^{a,b,*}, Mohammad Aliannejadi^c and Evangelos Kanoulas^c

^a*Singtel Cognitive and Artificial Intelligence Lab for Enterprises@NTU, 50 Nanyang Ave, 639798, Singapore*

^b*School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Ave, 639798, Singapore*

^c*Informatics Institute, University of Amsterdam, Science Park 904, Amsterdam 1098 XH, The Netherlands*

ARTICLE INFO

Keywords:

User Study

Clarifying Questions

Information-seeking Systems

ABSTRACT


Modern information-seeking systems are becoming more interactive, mainly through asking Clarifying Questions (CQs) to refine users' information needs. System-generated CQs may be of different qualities. However, the impact of asking multiple CQs of different qualities in a search session remains underexplored. Given the multi-turn nature of conversational information-seeking sessions, it is critical to understand and measure the impact of CQs of different qualities, when they are posed in various orders. In this paper, we conduct a user study on CQ quality trajectories, i.e., asking CQs of different qualities in chronological order. We aim to investigate to what extent the trajectory of CQs of different qualities affects user search behavior and satisfaction, on both query-level and session-level. Our user study is conducted with 89 participants as search engine users. Participants are asked to complete a set of Web search tasks. We find that the trajectory of CQs does affect the way users interact with Search Engine Result Pages (SERPs), e.g., a preceding high-quality CQ prompts the depth users to interact with SERPs, while a preceding low-quality CQ prevents such interaction. Our study also demonstrates that asking follow-up high-quality CQs improves the low search performance and user satisfaction caused by earlier low-quality CQs. In addition, only showing high-quality CQs while hiding other CQs receives better gains with less effort. That is, always showing all CQs may be risky and low-quality CQs do disturb users. Based on observations from our user study, we further propose a transformer-based model to predict which CQs to ask, to avoid disturbing users. In short, our study provides insights into the effects of trajectory of asking CQs, and our results will be helpful in designing more effective and enjoyable search clarification systems.

1. Introduction

Information-seeking systems are expected to deliver information to users efficiently and effectively in order to satisfy users' information needs. Benefiting from the recent developments in dialogue systems, the conversational information-seeking system is increasingly becoming a vital factor in information retrieval (IR). Such systems are now enabled to ask their users Clarifying Questions (CQs), in order to enhance their ability to understand the users' underlying information needs and retrieve the right information for them.

Asking CQs in information-seeking systems has been formalized and studied in different ways in the recent IR literature, including CQ generation (Zamani et al., 2020a,b), next CQ selection (Aliannejadi et al., 2019; Hashemi et al., 2020), and CQ engagement level prediction (Sekulic et al., 2021). Although these studies have demonstrated the significance and applicability of CQs, their main focus was on the development of effective algorithms. That is, studying the extent to which CQs affect users' performance and satisfaction in a search session is still underexplored, while equally important. Along this line, Zamani et al. (2020c) present a user study, demonstrating that CQs have both functional and emotional benefits. However, when generating CQs by a machine learning model in an information-seeking system, the model inevitably generates CQs of different qualities (Zamani et al., 2020a).¹ For instance, for a user who issues a query "Welcome to the Jungle" (which can be a film name or a song name) looking for the corresponding film, the system may generate a high-quality CQ like "Do you refer to the movie or the song?" to the user. The system may also ask a low-quality CQ such as "What song information are you looking for?". Accordingly, Wang and Ai (2021) point out that posing low-quality CQs, instead of directly showing search results, brings in the

*Corresponding author

 jie.zou@ntu.edu.sg (J. Zou); axsun@ntu.edu.sg (A. Sun); c.long@ntu.edu.sg (C. Long); m.aliannejadi@uva.nl (M. Aliannejadi); E.Kanoulas@uva.nl (E. Kanoulas)

¹The quality of CQs measures the degree of clarifying user intents and usefulness for completing the search task.

risk of user dissatisfaction. They model the risk of CQs using reinforcement learning. However, the authors do not provide a systematic study on how CQs of different qualities could impact user behavior and satisfaction. Zou et al. (2022), on the other hand, conduct a large-scale user study, demonstrating the effect of asking off-topic CQs on the users' interactions with the search engine and compare that with asking good or at least relevant CQs. Although they show that asking a good CQ leads to improved user performance and satisfaction, they study CQs in isolation of each other, and therefore it is unclear how a series of CQs affects each other and the user in a search session.² In a search session, CQs are not independent, each of which contribute to the overall structure of the conversation. Therefore, CQs may affect each other, while the order of CQs and their quality may affect the user behavior and satisfaction in different ways.

For example, what is the impact of displaying a low-quality CQ after a high-quality CQ and vice versa, on user satisfaction and behavior interacting with the Search Engine Result Pages (SERPs)? Because of the interactive nature of conversations, the system can infer the relevance of a CQ by the user's response (Zou et al., 2022). Therefore, knowing the relative impact of high-quality CQs after low-quality CQs (or vice versa) can lead to significantly improved choices in conversational information-seeking systems. In other words, while Zou et al. (2022) demonstrate the effectiveness of asking CQs in Web search, they do not provide insights into how system designers can utilize user feedback for a better search experience. For instance, what if the system asks a CQ and receives negative feedback from the user? Should the system avoid asking, or take the risk of asking another CQ? In a different scenario, if the system asks a CQ and receives positive feedback, should that affect the system's decision on asking another CQ or not?

To fill this research gap, in this work we explore the effect of various CQ trajectories in terms of CQ quality in information-seeking systems. Although the effect of question order has been studied in some domains such as social sciences (DeMoranville and Bienstock, 2003; Schuman and Presser, 1996), the effect of CQ trajectories of CQ quality has not been studied for information-seeking systems to the best of our knowledge. In more details, we aim to explore to what extent the trajectory of CQs of different qualities affects user search behavior and satisfaction. We first investigate the query-level impact of showing different-quality CQs on user behavior to understand the immediate impact caused by the adjacent query. In addition to the query-level impact, we also explore the session-level impact of different quality trajectories of CQs on user behavior and satisfaction, to understand the effect of showing different-quality CQs across the entire session. The quality of conversational information-seeking systems asking CQs is usually evaluated based on the entire session, and thus trajectories are important at the session level. This way we are able to capture the impact at the session level and compare it with query-level impact. To this end, we conduct a user study involving 89 participants, asking them to complete a set of Web search tasks, following a standard laboratory-based user study setup (Edwards and Kelly, 2017; Harvey and Pointon, 2017). In particular, we simulate various quality trajectory patterns that a user and a system would encounter where we study the effect of these trajectory patterns on user behavior and satisfaction.

Our user study leads to the following findings:

- Users interact with SERPs less extensively when they have seen preceding low-quality CQs before high-quality CQs. Similarly, users interact with SERPs more extensively if they have seen preceding high-quality CQs before low-quality CQs.
- A low-quality CQ leads to low search performance and user satisfaction. Asking a follow-up high-quality CQ improves search performance and user satisfaction, alleviating the earlier negative impact caused by asking a low-quality CQ.
- Compared with always showing CQs, only showing high-quality CQs while hiding other CQs leads to a better search experience: better search performance and better user satisfaction with shorter sessions and lower time efforts. That is, always showing CQs may be risky and low-quality CQs do disturb users.

Based on the above findings, we further develop a transformer-based model on top of user behavior data, named TranShow. TranShow predicts which CQs we should show to users, in order not to disturb users. To the best of our knowledge, TranShow is the first effort to utilize user behavior data for predicting whether a system should ask a CQ or not.

The rest of this paper is organized as follows. In Section 2, we introduce the related work. In Section 3, we explain the details of our method and study design. In Section 4, we describe the research questions, and provide a detailed analysis as well as a discussion of the results. Section 5 expresses our CQ showing prediction model and performance analysis, while Section 6 concludes the paper.

²A search session is an entire session for a user completing a search task, which may involve multiple actions from the user.

2. Related Work

In this section, we summarize the related work. Here we mainly review the work that is closely related to our study and focus on the literature of user studies on asking CQs. We first introduce the applications and the algorithms developed for asking CQs to highlight the importance of CQs. We then describe related work on user studies on asking CQs.

Thanks to the great power of collecting users' explicit feedback, asking CQs has recently been employed in a wide range of areas, such as conversational recommender systems (Sepiarskaia et al., 2018; Zou et al., 2020a), conversational product search (Zou and Kanoulas, 2019; Zhang et al., 2018), question answering (De Boni and Manandhar, 2003; Xu et al., 2019), and information-seeking systems (Aliannejadi et al., 2019; Hashemi et al., 2020; Zamani et al., 2020a; Wang and Ai, 2021; Lipani et al., 2021; Radlinski and Craswell, 2017; Sekulić et al., 2022; White and Iivonen, 2001).

In information-seeking systems, an increasing number of approaches have been developed to model CQs and enable information-seeking conversations. Early exploration attempts to provide choices in a search session (Belkin et al., 1995). More recently, researchers begin to work on a wider range of topics, such as generating CQs in information-seeking systems (Zamani et al., 2020a; Hashemi et al., 2020; Wang and Li, 2021), predicting CQ engagement levels (Sekulic et al., 2021), modeling the risk of CQs (Wang and Ai, 2021), and collecting benchmark datasets (Aliannejadi et al., 2019; Zamani et al., 2020b; Aliannejadi et al., 2020; Ren et al., 2021; Chu et al., 2022). These studies highlight the benefit of asking CQs and further prompt the research of asking CQs in search systems. Different from the aforementioned studies, which primarily focus on the algorithm developed for asking CQs, we explore the impact of asking CQs and study the underlying mechanism of user interactions with CQs, providing insights into the design of these algorithms.

Another research direction related to asking CQs is to conduct user studies on asking CQs, which is most related to our study. Research discussing user studies is broad, from user perception of algorithms (e.g., privacy concerns in the context of personalized algorithms (Shin et al., 2022b), the effect of cultural values on algorithms (Shin et al., 2022a), and algorithmic literacy on user acceptance (Shin et al., 2021)) to blockchain affordances (Shin and Hwang, 2020) and conversational journalism (Shin, 2021). In this study, we focus on the user study on examining CQs in information-seeking systems.

At the early stage, Vtyurina et al. (2017) present a user study to compare user behavior for three different conversational search agents: humans, assistants, and wizards. They find that users are glad to use the three conversational search agents as long as their expectations about accuracy are met. Instead of performing clarification over texts, Kiesel et al. (2018) and Trippas et al. (2017) study the effect of clarification over voices. They find that it would be beneficial to select the best response method in different scenarios. More recently, with the unknown impact of document ranking models incorporating CQs, Krasakis et al. (2020) analyze the effect of CQs on the performance of document ranking. They highlight the importance of understanding and incorporating explicit conversational feedback. Avula et al. (2022) investigate user behaviors in the context of collaborative search and discuss important directions on mixed-initiative conversational search systems to support collaborations. While most algorithms are designed with the assumption that users are willing to provide answers for CQs, and are able to answer CQs correctly, Zou et al. (2020b) empirically quantify and validate the extent of users' willingness to provide answers, and the ability to provide correct answers to CQs, in existing CQ-based systems. They demonstrate that users provide noisy answers sometimes, and future research on CQ-based algorithms should take noisy answers into consideration. At the same time, Zamani et al. (2020a) perform an in-situ analysis showing that asking CQs is useful, and asking CQs is superior to simply showing some options. They further conduct a large-scale study, analyzing the characteristics leading to a high engagement rate of CQs, including search query properties, CQ template types, and answer attributes (Zamani et al., 2020c). Similarly, Tavakoli et al. (2022) attempt to identify the characteristics of more useful CQs in terms of types and patterns. Instead of conducting an in-situ study, Zou et al. (2022) perform a controlled laboratory user study to understand the user interactions with CQs for search clarification. They focus on the effect of clicking CQs in different categories on user behaviors, and the effect of user background, CQ category, task types, query index, SERP quality, and screen size on CQ engagement (i.e., the click-through rate of CQs). While we are the first to explore the effects of trajectories of asking CQs in Web search, this is the closest work to ours. Different from their study (Zou et al., 2022), our study focuses on the effects of trajectories of asking CQ by exploring different trajectory patterns. Moreover, we focus on the effect of showing patterns instead of clicks of CQ in different categories, which is considered in Zou et al. (2022). Furthermore, we develop a transformer model to predict the appearance of CQs.

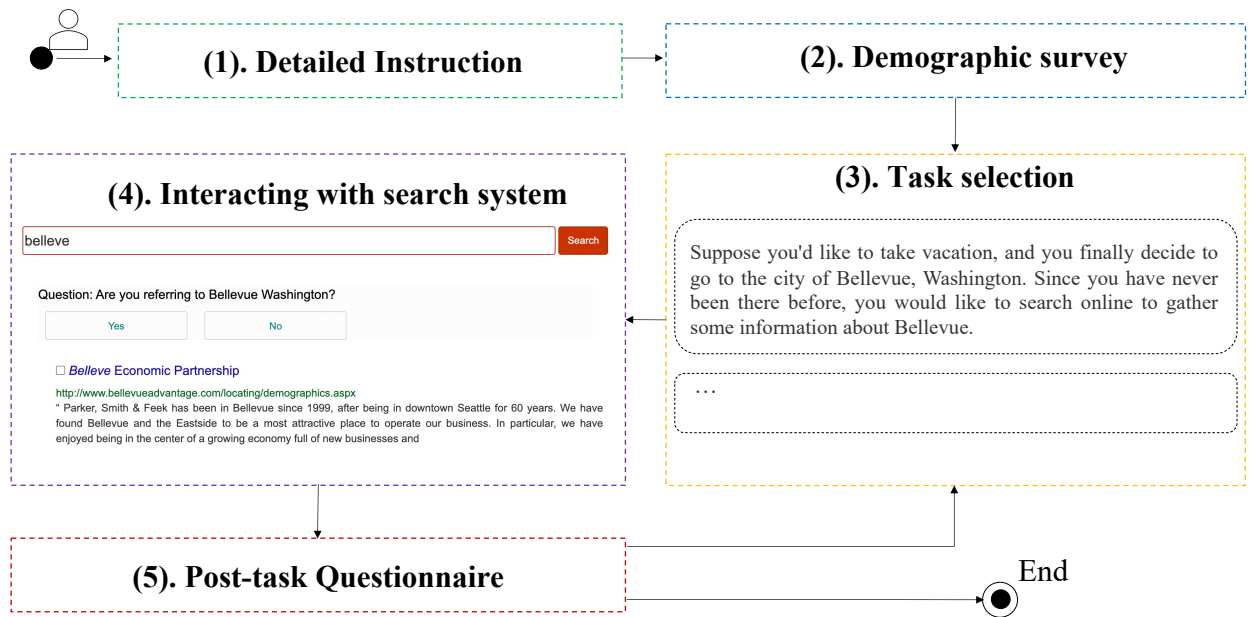


Figure 1: Study protocol overview. An example Web search task is shown in Step (3), and the search interface is shown in Step (4). After completing Step (5), users may choose to complete another task from Step (3).

3. Method

In this section, we present our method, including the study protocol, study system, CQs and trajectory patterns, behavior and satisfaction measures, and participants.

3.1. Study Protocol

We have developed an online Web system to conduct the user study. The study protocol is shown in Figure 1 as follows:

- (1) Participants are presented with detailed instructions regarding this study. They are also trained to be familiar with the system through a detailed video.
- (2) We ask the participants to complete a demographic survey on their gender, age, career field, English language proficiency, and highest education level completed.
- (3) Participants are presented with a list of Web search tasks. They are guided to read through all the task descriptions. Then they select a Web search task with which they feel most comfortable, to relieve task assignment bias (Ho and Vaughan, 2012).
- (4) Participants start to search for relevant information for the selected task, either by submitting queries themselves or by interacting with presented CQs. Once they find some relevant results, they can bookmark the results as relevant.
- (5) After the participant completes a Web search task, he/she is asked to complete a post-task questionnaire, on his/her experience and satisfaction, including the overall satisfaction rating, and perceived helpfulness.
- (6) Participants are asked to consider selecting another Web search task to complete, by returning to step (3), or they can finish the study participation.

The flow of study processing for asking CQs in Web search is also provided in Figure 2. All participants are informed that their data is securely encrypted and is only used in a collective manner. We did not collect any data to breach their privacy. Moreover, this study is approved by the ethics committee of the institute.

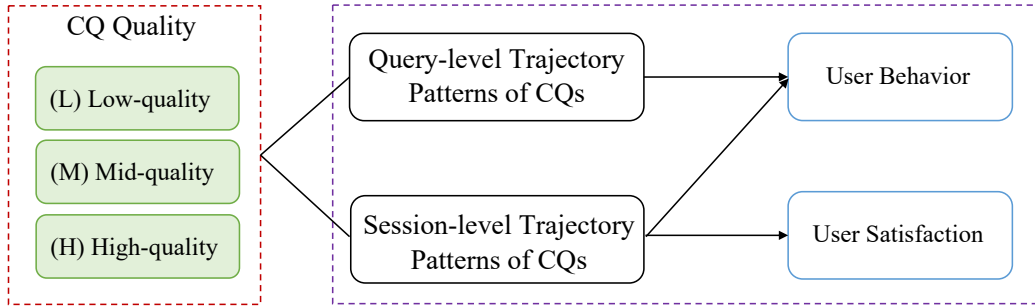


Figure 2: The flow of study processing for asking CQs in Web search. We aim to explore the effect of CQ trajectories (query-level and session-level) in terms of CQ quality on user behavior and satisfaction. CQ quality, query-level and session-level trajectory patterns of CQs are described in Section 3.3. User behavior and satisfaction measures are introduced in Section 3.4.

Table 1

Example CQs of different qualities. The generated CQs are from the search task “Bellevue Washington.”

Quality	Examples of CQs & answers
(L) Low	Q1: What would you like to know about Bellevue Hospital Center in New York? A1: 1. History; 2. Facilities; 3. Address; 4. Homepage; 5. Contact info.
(M) Mid	Q2: Do you need information about Bellevue in Nebraska? A2: 1. Yes; 2. No, Bellevue in Washington.
(H) High	Q3: Which Bellevue are you interested in? A3: 1. Bellevue in Nebraska; 2. Bellevue Hospital Center in New York; 3. Bellevue University in Nebraska; 4. Bellevue Theater in Amsterdam; 5. Bellevue in Washington.

3.2. Study System

We follow Zou et al. (2022) to mimic the search interface of a commercial search system (Bing.com) (Zamani et al., 2020a) and use ChatNoir³ Web search engine (Potthast et al., 2012; Bevendorff et al., 2018) to produce SERPs. ChatNoir is a widely used Web search engine indexing the entire ClueWeb09/12 corpus (Hagen et al., 2017; Vakkari et al., 2019). As shown in step (4) in Figure 1, the system includes a typical search interface and embeds a CQ pane at the top of the search page. Users are allowed to reformulate their queries by themselves or interact with the CQ pane (if presented).

We design four search tasks derived from the Text Retrieval Conference (TREC) Web Track 2009 – 2012,⁴ with an expanded description to offer participants a search context as done in previous work (Borlund, 2003; White et al., 2007).⁵ These search tasks are about collecting information regarding certain topics (e.g., a tourist city, an airport, a movie, a famous quote, or a fun fact) from various Web sources. An task example is shown in step (3) in Figure 1.

3.3. CQs and Trajectory Patterns

In the absence of a reasonable model to automatically generate CQs and candidate answers of different qualities, we manually construct a pool of CQs along with their candidate answers beforehand, same as other studies in the community (Aliannejadi et al., 2019, 2020; Ren et al., 2021; Zou et al., 2022).

In this work, we use three-level CQ quality categories similar to Zou et al. (2022) and Zamani et al. (2020b): (L) low-quality, (M) mid-quality, and (H) high-quality. The CQ generation pipeline is as follows. First, two expert annotators generate CQs in three quality categories as well as candidate answers, for each search task. In case of disagreement, they would discuss and agree on a better formulation. Second, the expert-generated CQs are deployed

³<https://www.chatnoir.eu>

⁴<https://trec.nist.gov/data/webmain.html>

⁵From our post-hoc analysis, distribution of the selected search tasks is balanced. There is no obvious preference for any particular task. Furthermore, 93% of participants indicate that the task description is very clear.

Table 2
Trajectory Patterns.

Query-level		Session-level	
Trajectory	Description	Trajectory	Description
L2L	... L → L ...	Lb4L	... L → ... → L ...
L2M	... L → M ...	Lb4M	... L → ... → M ...
L2H	... L → H ...	Lb4H	... L → ... → H ...
M2L	... M → L ...	Mb4L	... M → ... → L ...
M2M	... M → M ...	Mb4M	... M → ... → M ...
M2H	... M → H ...	Mb4H	... M → ... → H ...
H2L	... H → L ...	Hb4L	... H → ... → L ...
H2M	... H → M ...	Hb4M	... H → ... → M ...
H2H	... H → H ...	Hb4H	... H → ... → H ...

in a pilot study, and all CQs are labeled by the actual users independently, again in the three CQ quality categories. Third, we remove those CQs with conflict category labels between expert annotators and pilot study users. Hence, only the CQs with consistent quality labels are used in the actual study, to ensure the label quality. In the end, we have six CQs with their respective candidate answers for each search task. Each CQ has at most five answers, with each answer corresponding to a reformulated query for the next turn, following Bing’s setting (Zamani et al., 2020c; Zou et al., 2022). When generating and labeling the three-level CQ quality categories, we provide detailed guidelines and examples for each of the low-, mid-, and high-quality CQs to expert annotators and pilot study users. Specifically, the guideline indicates that high-quality CQs should meet the following criteria (Zamani et al., 2020a): (1) highly helpful for completing the search task, (2) correctly clarifying user intents, and (3) being fluent and grammatically correct. If a CQ fails to meet some of these criteria but is still an acceptable CQ to be asked, it should be labeled as mid-quality. Otherwise, low-quality labels should be provided by expert annotators and pilot study users.⁶ The general principle in constructing such CQ quality categories is to cover a variety of CQ quality categories and explore the potential effects of CQ trajectories under these quality categories. High-quality CQ is designed as the highest CQ quality which is highly useful and can aid users, whereas low-quality CQ is the lowest CQ quality which is usually useless or off-topic so that may elicit user dissatisfaction. CQ examples under each CQ quality category are shown in Table 1.

To study trajectory effects in information-seeking systems, we also define a series of query-level and session-level trajectory patterns, as shown in Table 2. Specifically, we define the following nine trajectory patterns to study the query-level impact of trajectory patterns: L2L, L2M, L2H, M2L, M2M, M2H, H2L, H2M, and H2H. Take L2H as an example of query-level trajectory. It means that a low-quality CQ is asked first, followed by a high-quality CQ. Accordingly, we define the following nine session-level trajectory patterns: Lb4L, Lb4M, Lb4H, Mb4L, Mb4M, Mb4H, Hb4L, Hb4M, and Hb4H. For example, Lb4H means that in a session, there is a low-quality CQ before a high-quality CQ, where a search session may involve multiple CQs. In more detail, assume a search session contains three CQs to form a CQ trajectory pattern $L \rightarrow M \rightarrow H$, i.e., asking a low-quality CQ as the first CQ (e.g., Q1 in Table 1), then asking a mid-quality CQ as the second CQ (e.g., Q2 in Table 1), then asking a high-quality CQ as the third CQ (e.g., Q3 in Table 1). This search session would belong to each of the following session-level trajectory patterns: Lb4M, Mb4H, and Lb4H, and each of the following query-level trajectory patterns: L2M and M2H.

The participants were split into two groups. One group of participants completes the search task with a plain interface, i.e., without CQs shown in the search session. Another group of participants completes the search task with a CQ pane, similar to the one in step (4) in Figure 1. If a participant was assigned to the group with a CQ pane, CQs would be shown by following a randomly assigned CQ trajectory pattern, following the assignment setup of related literature (Collins-Thompson et al., 2016; Harvey and Pointon, 2017; Kelly and Azzopardi, 2015; Kelly, 2009). For each trajectory pattern, the CQ to be shown to the user was randomly selected from CQs under the corresponding CQ quality category for a certain search task (Collins-Thompson et al., 2016; Harvey and Pointon, 2017; Kelly and Azzopardi, 2015; Kelly, 2009).⁷ Users can answer multiple CQs in one search session. When a CQ is shown to the user, the user can choose to reformulate their query again or click on a CQ answer. If the user clicks on a CQ answer,

⁶We allowed participants to report issues with CQs; none was reported.

⁷From our post-hoc analysis, the distribution of each trajectory pattern is balanced.

the answer is concatenated to the user's query and resubmitted to the search engine, following Bing's setting (Zamani et al., 2020c). For participants in the group with a CQ pane, once they issue a new query, or answer a CQ, a new CQ is selected to be displayed to them.

3.4. Behavior and Satisfaction Measures

In our user study, we use the user behavior and satisfaction measures, following Kelly and Azzopardi (2015) and Zou et al. (2022). Specifically, regarding behavior measures, we include the following metrics: the CQ engagement rate (click-through rate), number of SERP scrolls, number of SERP hovers, number of SERP clicks, number of queries issued, number of query terms, number of results marked relevant (# bookmarks), number of correct bookmarks (# hit), and SERP quality measured by nDCG@10 (normalized discounted cumulative gain from rank 1 to 10), dwell time on SERPs per query, and the overall task time for a Web search task. Regarding satisfaction, we use explicit feedback collected through the post-task questionnaires, including the overall satisfaction rating, and user-perceived helpfulness. Specifically, in the post-task questionnaires, the options for perceived helpfulness include "positive," "negative" and "neutral." The overall satisfaction rating is on a scale from 1 to 5. The query-level measures are calculated based on the current search page while the session-level measures are calculated accumulatively throughout the full search session.

3.5. Participants

We recruit participants through an academic crowdsourcing platform, called Prolific.⁸ Before the actual study, we ran a pilot study with 23 participants, to iterate over the experimental design and confirm the quality labels of CQs. For the actual user study, 89 participants are recruited after quality control filtering.

To ensure data quality from crowdsourcing participants, we perform quality control filtering to avoid including results from some less trustworthy participants. Specifically, we deploy two quality checks: (i) we ask participants questions regarding the study descriptions to make sure that they have read and understood the instructions, and (ii) we evaluate the time participants spent reading the textual descriptions of Web search tasks. We then filter out participants who read the task descriptions in less than 10 seconds (a minimal expected threshold for a trustworthy worker (Han et al., 2020)). The removed participants are further checked manually to ensure they are not filtered out wrongly.

Participants were paid around 3.33 pounds each to complete the study. Their demographic data is briefed as follows:

- Gender: 55 females, 30 males, 4 non-binary.
- Age: 12 participants are in the age group 18–24, 41 are in 25–34, 21 are 35–44, and 15 are older than 44 years old.
- Career field: 21 in science, computers and technology, 11 in management, business and finance, 10 in education and social services, 13 in healthcare, 9 in arts and communications, 4 in law and law enforcement, and 1 in trades and transportation; 20 do not specify.
- English language proficiency: 82 native, 5 proficient, and 2 beginners.
- Highest education level completed: 11 high school, 23 college, 33 bachelor's, 10 master's, and 2 doctorate; 10 do not specify.

4. Results

The analysis is based on 897 search requests made between our system and 89 crowd workers on 241 search sessions. The statistics of the collected data are detailed in Table 3. Unless otherwise reported, we perform t-tests (Kim, 2015) and one-way analysis of variance (ANOVA) for statistical analysis in this study, assuming the independence of different groups (Xie et al., 2018). Specifically, we use t-tests for comparisons between two groups only, and use one-way ANOVA and Tukey's HSD tests for comparisons with more than two groups, following past user studies (Turpin and Scholer, 2006). The presence or absence of CQs is a between-subjects variable, whereas the CQ quality category is a within-subjects variable.

Through analyzing the collected data, we answer the following research questions:

⁸<https://www.prolific.co/>

Table 3

Statistics of collected data through the user study.

# users	89
# search tasks	4
# search sessions	241
# search requests	897
# user bookmarks	1,074
avg. # search tasks per user	2.71
# CQs	24
# CQ showing/hiding times	760/137
# CQ clicks	388
# CQ showing times in L/M/H	297/224/239
# CQ clicks in L/M/H	76/147/151
# user cursor hovering records	7,867
# user page scrolling records	6,616

Table 4

User behavior measures by different query-level trajectory patterns of showing CQs. Corresponding values are reported by means, followed by standard deviations in parentheses. * denote significant difference with No CQs (p -value < 0.05).

	No CQs	L2L	M2L	H2L	L2M	M2M	H2M	L2H	M2H	H2H
SERP hovers	11.36(10.49)	7.99(11.24)	9.80(12.52)	10.29(7.26)	5.96(10.11)	6.80(6.58)	7.31(9.28)	6.27(7.10)	6.79(8.48)	10.11(12.03)
SERP scrolls	11.30(12.54)	7.75(12.38)	8.78(10.72)	9.84(9.90)	4.53(11.21)*	5.35(8.88)	6.37(10.94)	5.77(10.07)	5.23(8.78)	7.23(11.93)
SERP clicks	0.67(1.36)	0.17(0.71)*	0.11(0.42)*	0.28(0.55)	0.11(0.72)*	0.27(0.85)	0.16(0.57)*	0.07(0.25)*	0.30(1.07)	0.25(0.66)
dwel time(s)	47.34(48.89)	29.50(35.56)	32.52(33.94)	37.23(30.39)	19.41(30.20)*	22.39(25.01)*	26.07(34.36)*	18.39(17.15)*	28.52(47.75)	30.95(41.51)
nDCG@10	0.35(0.29)	0.28(0.26)	0.42(0.32)	0.42(0.33)	0.31(0.25)	0.41(0.35)	0.35(0.32)	0.30(0.24)	0.35(0.34)	0.41(0.37)
# bookmarks	1.98(2.19)	1.35(3.28)	1.42(2.24)	1.60(2.21)	0.49(1.13)*	0.76(1.12)	0.90(1.90)	0.77(1.29)	1.11(2.05)	1.48(3.18)
# hits	1.02(1.30)	0.84(2.20)	0.68(1.10)	0.66(0.97)	0.28(0.74)	0.33(0.62)	0.34(0.74)	0.32(0.85)	0.48(1.10)	0.65(1.25)

- **RQ1:** To what extent do different trajectory patterns affect user behavior at the query level?
- **RQ2:** To what extent do different trajectory patterns affect user behavior and satisfaction at the session level?
- **RQ3:** Does always showing CQs benefit or disturb users?

4.1. Query-level Impact of Trajectory (RQ1)

We first investigate the query-level impact of CQ trajectory patterns on user search behaviors. The results of CQ engagement are plotted in Figure 3a, and results of other user behavior measures categorized by different query-level trajectory patterns are reported in Table 4.

CQ engagement. Observe from Figure 3a, the engagement of first showing CQs is higher than others. A possible reason is that users focus more on their first seen CQs. Showing better preceding CQs (mid- or high-quality CQs) before low-quality CQs increases the CQ engagement of low-quality CQs, indicating CQs might affect user interactions of the following CQs, and low-quality CQs are beneficial from the last seen mid- or high-quality CQs.

SERP scrolls and hovers. SERP scrolls and hovers are valuable signals of user behavior in Web search (Huang et al., 2012). Table 4 shows that the number of scrolls and hovers increase from L2L to H2L, from L2M to H2M, and from L2H to H2H. This observation for scrolls and hovers indicates that users scan a SERP more extensively when they have seen a preceding high-quality CQ and users scan a SERP less extensively when they are shown a preceding low-quality CQ. That is, users are more confident about SERP when they have seen a preceding high-quality CQ.

SERP clicks. We observe that the number of SERP clicks on the H2L condition is more than that of the L2L condition. Also, we see that the number of SERP clicks on H2H is higher than that on L2H, suggesting that the quality of a preceding CQ plays a role in the number of SERP clicks on the current CQ. That is, asking a preceding high-quality CQ leads to higher SERP clicks for the current CQ, no matter what the quality of the current CQ is.

Dwell time. For SERP dwell time, we see $L2L < M2L < H2L$, $L2M < M2M < H2M$, and $L2H < M2H < H2H$. This result again confirms that users spend more time and scan a SERP more extensively when they have seen a preceding

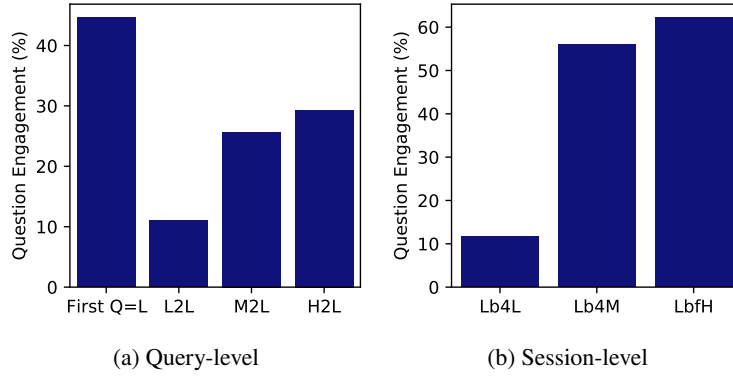


Figure 3: CQ engagement of different (a) query-level and (b) session-level trajectory patterns for low-quality CQs.

high-quality CQ, while users spend less time and scan a SERP less extensively when they are shown a preceding low-quality CQ. A reason is that users tend to interact with the high-quality CQ when they see a preceding high-quality CQ, leading to an improved SERP quality. The improved SERP quality then makes users locate more relevant SERP results, leading to an increase in dwell time.

SERP quality. For SERP quality, measured by $nDCG@10$, we observe $L2L < H2L$, $L2M < H2M$, and $L2H < H2H$. This indicates that showing a preceding low-quality CQ before a high-quality CQ decreases SERP quality while showing a preceding high-quality CQ before a low-quality CQ increases SERP quality. A possible reason is query reformulation, i.e., users reformulate their queries toward better queries when they have seen a preceding high-quality CQ.

Bookmark quality. Bookmark quality, in terms of # bookmarks and # hits, reflects user search performance for finding relevant information. From Table 4, we observe that both the number of bookmarks and the number of hits decrease from H2H to M2H to L2H. This demonstrates that showing a preceding low-quality CQ before a high/mid-quality CQ lowers user search performance for finding relevant information, as a preceding low-quality CQ leads to low SERP quality. In contrast, showing preceding high/mid-quality CQs before low-quality CQs increases the number of bookmarks.

Summary. Overall, as for the query-level impact of different trajectories of asking CQs on user behavior, we observe the trend that adding preceding low-quality CQs before high-quality CQs lowers user behavior measures. Adding preceding high-quality CQs before low-quality CQs leads to increases in user behavior measures. That is, when users have seen preceding low-quality CQs before high-quality CQs, they interact with SERPs less extensively, including less CQ engagement, mouse hovering behavior, mouse scrolling behavior, SERP clicking behavior, bookmark behaviors (number of bookmarks and number of hits), and spend less dwell time. In contrast, users interact with SERPs more extensively when they have seen preceding high-quality CQs before low-quality CQs.

4.2. Session-level Impact of Trajectory (RQ2)

We now explore the session-level impact of CQ trajectory patterns on user search behaviors. In addition, we investigate how user satisfaction for the whole session is affected by CQ trajectory patterns.⁹

4.2.1. User Behavior

We analyze user behavior in a session where we display a high-quality CQ after a low-quality CQ (or vice versa) and see how it affects various user behavior measures on a session-level scale. The results of CQ engagement are shown in Figure 3b and the results of other user behavior measures are listed in Table 5. To make the comparison easier, we report the results of four more descriptive trajectory patterns, i.e., Lb4L, Lb4H, Hb4L, and Hb4H, in Table 5.

CQ Engagement. From Figure 3b, we observe that a low-quality CQ leads to a low CQ engagement, and showing a high-quality CQ after a low-quality CQ in a session increases CQ engagement. This observation that users prefer to

⁹Users perceive satisfaction from the whole search session and they only rate satisfaction measures after they complete a search session. Therefore, we explore the effects of CQ trajectory on user satisfaction from the session level only.

Table 5

Objective behavior measures by different session trajectory patterns. Corresponding values are reported by means, followed by standard deviations in parentheses. * denote significant difference with No CQs (p -value < 0.05).

	No CQs	Lb4L	Lb4H	Lb4L-Lb4H	Hb4L	Hb4H	Hb4L-Hb4H
SERP hovers	20.83(16.20)	41.82(30.75)*	46.79(31.16)*	4.97	40.95(30.70)*	45.73(30.62)*	4.78
SERP scrolls	20.72(17.21)	32.59(30.41)	35.52(30.04)	2.93	33.11(28.22)	36.86(30.06)	3.75
SERP clicks	1.22(1.84)	0.72(1.83)	0.80(1.79)	0.08	0.84(1.86)	0.91(1.99)	0.07
# query terms	6.83(6.96)	20.04(14.52)*	20.29(14.32)*	0.25	20.57(14.41)*	20.62(14.07)*	0.05
# queries	1.83(1.21)	5.89(3.54)*	6.29(3.51)*	0.4	5.95(3.45)*	6.06(3.38)*	0.11
session time(s)	66.97(65.17)	120.71(113.10)	120.45(103.10)	0.26	114.70(109.54)	122.73(111.51)	8.03
nDCG@10	0.64(0.69)	1.97(1.88)*	2.10(2.06)*	0.13	2.09(1.89)*	2.07(1.98)*	0.02
# bookmarks	3.64(2.24)	4.93(4.14)	5.91(4.34)	0.98	5.03(3.93)	6.19(5.27)	1.16
# hits	1.86(1.57)	2.61(2.96)	2.64(2.93)	0.03	2.49(2.64)	2.82(3.23)	0.33

engage with high-quality CQs, is in line with previous work (Zou et al., 2022, 2020b). We see a considerable difference in CQ engagement when comparing sessions of Lb4L and Lb4H. Starting with a low-quality CQ and followed by higher-quality CQs lead to better engagement. This suggests that there is a good chance for a system to recover from an initial low-quality CQ by asking CQs of higher quality. Therefore, systems can take into account user feedback on a low-quality CQ, and make a more informed decision on showing the next CQs.

SERP scrolls and hovers. From Table 5, we observe that numbers of SERP scrolls and hovers increase from Lb4L to Lb4H, and from Hb4L to Hb4H. This indicates that users are more engaged with SERP when CQs of higher quality are posed to them after a preceding CQ. We observe a somewhat symmetric difference in terms of hover and scroll when comparing the four conditions and keep the initial CQ fixed ($lLb4L - Lb4Hl$ vs. $lHb4L - Hb4Hl$). The effect of showing a high-quality CQ after a low-quality CQ is an inverse of showing a low-quality CQ after a high-quality CQ. Therefore, as much as the system can recover from a low-quality CQ (by asking a high-quality CQ), it faces the risk of losing user interest by asking low-quality CQs (after a high-quality CQ).

SERP clicks. The number of SERP clicks increases from Lb4L to Lb4H, and from Hb4L to Hb4H. The trend is similar to scrolls and hovers. Therefore, the same symmetric effect exists in this case too.

User queries. Also, numbers of user queries and query terms increase from Lb4L to Lb4H, and from Hb4L to Hb4H.

Session time. The session time increases from Hb4L to Hb4H while remaining a similar number between Lb4L and Lb4H. Here, we observe an interesting phenomenon: a considerable drop in session time when a low-quality CQ is shown after a high-quality CQ (Hb4L). This can be due to the fact that users find most of the relevant results when they interact with high-quality CQs; therefore, when they see a low-quality CQ, there is less motivation to interact with the SERP, leading to a shorter session time.

SERP quality. The SERP quality increases from Lb4L to Lb4H while remaining a similar number between Hb4L and Hb4H.

Bookmark quality. Numbers of marked results and hits increase from Lb4L to Lb4H, and from Hb4L to Hb4H. We see a higher impact when comparing the difference between Hb4L and Hb4H, as opposed to the difference between Lb4L and Lb4H. There is more risk in showing a low-quality CQ after a high-quality CQ, compared with the gain one receives when showing a high-quality CQ after a low-quality CQ. This can be due to the first impression of users when they interact with a low-quality CQ, and that impacts their behavior and interest all over the session.

Summary. As for session impact, compared with showing low-quality CQs, showing follow-up high-quality CQs leads to better user behavior; users spend similar time to find more relevant information (# bookmarks and # hits). This suggests that while a low-quality CQ leads to low user search performance and user satisfaction, asking a follow-up high-quality CQ improves search performance and user satisfaction, alleviating the earlier negative impact caused by the low-quality CQ. After showing high-quality CQs, showing follow-up high-quality CQs leads to finding more relevant information, but also more time, compared with showing follow-up low-quality CQs. However, we observe that even though the user effort may not change much in these two cases, their efficiency is different. There are various

Table 6

User satisfaction measures by different session trajectory patterns. Satisfaction are represented by its means followed by standard deviations in parenthesis. User-perceived helpfulness are represented by ratio of positive/negative ratings.

	No CQs	Lb4L	Lb4H	Hb4L	Hb4H
satisfaction	3.33(1.20)	2.97(1.16)	3.04(1.32)	3.28(1.20)	3.33(1.14)
helpfulness(%)	25.00/50.00	47.95/46.58	56.36/40.00	64.00/26.67	65.38/26.92

possible reasons, e.g., a bad first impression, or misleading CQs that affect the user's next query in a session. Therefore, even though the system can recover from a low-quality CQ, it still cannot fully recover because of the impact of the low-quality CQ has on users' behavior and interest.

4.2.2. User Satisfaction

Observe from Table 6 that, when there are low-quality CQs in a session (Lb4L, Lb4H, and Hb4L), user satisfaction ratings are lower than no CQs showing. This confirms that low-quality CQs lead to low user satisfaction. Showing high-quality CQs after low-quality CQs in a session improves the user satisfaction ratings, and user-perceived helpfulness (i.e., higher positive percentage and lower negative percentage) (Lb4L vs. Lb4H). Showing low-quality CQs after high-quality CQs in a session lowers the user satisfaction ratings, and user-perceived helpfulness (Hb4L vs. Hb4H). In short, when a low-quality CQ leads to low user satisfaction, a follow-up high-quality CQ improves user satisfaction and alleviates the earlier negative impact of the low-quality CQ.

4.3. Session-level Impact of Always Showing CQs (RQ3)

Next, we explore the session-level impact on user behavior in cases where a system would always show CQs to users. Our goal is to demonstrate the importance of effective prediction of clarification needs, in a search session, and the negative effect it may bring in on the user's experience.

4.3.1. User Behavior

Table 7 reports session-level behavior measures for always vs. not-always showing CQs. For the condition of always showing CQs, we constantly show a CQ (low- or mid- or high-quality CQ) for each search request in the search session. For the condition of not-always showing CQs, we show only one category of CQs (low- or mid- or high-quality CQ) and hide CQs for search requests if they are not low/mid/high-quality CQs in the search session, respectively. For instance, for search sessions showing high-quality CQs only ('H-only'), we show the CQ if it is a high-quality CQ and hide the CQ if it is not a high-quality CQ.

CQ engagement. We observe that CQ engagement decreases for low-quality CQ from always showing CQs (48.05%) to not-always showing CQs (28%), while increases for high-quality CQ from always showing CQs (50.41%) to not-always showing CQs (54.10%). Showing various low-quality CQs leads to more engagement from users, which can also increase the risk of dissatisfaction because these CQs are more likely to be useless.

SERP scrolls and hovers. Observe from Table 7, numbers of SERP scrolls and hovers increase from low-quality CQs to mid-quality CQs to high-quality CQs. Compared with L and M (always showing), L-only and M-only (not-always showing) achieve a lower number of SERP scrolls and hovers, respectively. H and H-only contain similar numbers of SERP scrolls and hovers. This again confirms that showing more CQs lead to more user engagement and bears more cost to users, leading to the risk of user dissatisfaction.

SERP clicks. Number of SERP clicks changes slightly from low-quality CQs to mid-quality CQs to high-quality CQs. M-only and H-only (not-always showing) achieve a higher number of SERP clicks than M and H (always showing), respectively. Despite a higher CQ engagement as well as hovering and scrolling, users click on fewer results. This suggests that even though users are engaged more with SERP and spend more time, their performance to click the interesting results is lower when CQs are always shown to them.

User queries. Numbers of user queries and query terms increase from low-quality CQs to mid-quality CQs to high-quality CQs. L-only, M-only, and H-only (not-always showing) achieve a lower number of user queries and query terms than L, M, and H (always showing), respectively.

Table 7

Objective behavior measures by always showing CQs and not-always showing CQs. L/M/H means sessions always showing CQs and showed CQs contain low/mid/high-quality CQs, respectively. L/M/H-only means sessions showing low/mid/high-quality CQs only and hiding CQs if it is not low/mid/high-quality CQs, respectively. * denote significant difference with No CQs (p -value < 0.05).

	No CQs	Always showing			Not-always showing		
		L	M	H	L-only	M-only	H-only
SERP hovers	20.83(16.20)	34.70(27.03)	35.70(27.48)*	36.87(28.31)*	23.37(20.10)	27.16(15.13)	35.04(28.30)
SERP scrolls	20.72(17.21)	28.64(25.05)	29.70(24.88)	31.02(26.28)	22.10(24.46)	24.74(18.15)	32.92(27.25)
SERP clicks	1.22(1.84)	0.90(1.75)	0.87(1.67)	0.86(1.78)	0.93(1.50)	1.26(1.68)	1.00(1.84)
# query terms	6.83(6.96)	15.41(12.46)*	16.79(12.79)*	17.09(12.96)*	7.70(6.19)	12.68(9.75)	14.12(11.90)
# queries	1.83(1.21)	4.44(3.14)*	4.82(3.12)*	4.83(3.12)*	2.13(1.50)	3.11(1.71)	3.50(2.12)
session time(s)	66.97(65.17)	102.47(98.44)	107.81(97.52)	108.86(105.84)	69.36(88.44)	55.78(45.48)	89.53(106.07)
nDCG@10	0.64(0.69)	1.54(1.58)*	1.66(1.66)*	1.68(1.71)*	0.64(0.62)	1.22(1.11)	1.36(1.52)
# bookmarks	3.64(2.24)	4.64(3.71)	4.71(3.64)	5.12(4.44)	3.83(3.56)	3.16(1.23)	5.77(6.57)
# hits	1.86(1.57)	2.30(2.36)	2.19(2.25)	2.47(2.70)	2.17(2.45)	1.58(1.27)	2.96(3.67)
rate of gain	0.0277	0.0224	0.0203	0.0229	0.0312	0.0283	0.0331

Table 8

User satisfaction measures by always showing CQs and not-always showing CQs. Satisfaction are reported by its means followed by standard deviations in parenthesis. User-perceived helpfulness are represented by ratio of positive/negative ratings.

	No CQs	Always showing			Not-always showing		
		L	M	H	L-only	M-only	H-only
satisfaction	3.33(1.20)	3.17(1.24)	3.26(1.25)	3.32(1.20)	3.00(1.29)	3.63(1.31)	3.65(0.87)
helpfulness(%)	25.00/50.00	51.88/38.35	54.03/34.68	61.54/27.69	21.74/69.57	50.00/33.33	72.00/12.00

Session time. Session time increases from low- to high-quality CQs. L-only, M-only, and H-only (not-always showing) achieve lower session time than L, M, and H (always showing), respectively. As mentioned earlier, more engagement with CQs as well as SERP leads to more time spent on SERP.

SERP quality. SERP quality increases from low-quality CQs to mid-quality CQs to high-quality CQs. L-only, M-only, and H-only (not-always showing) achieve lower SERP quality than L, M, and H (always showing), respectively. This is perhaps because engaging with more CQs leads to longer sessions and more SERPs, hence having higher nDCG.

Bookmark quality. Number of marked results and hits increase from low-quality CQs to high-quality CQs. L-only and M-only (not-always showing) achieve a lower number of marked results and hits than L and M (always showing), respectively, while H-only achieves a higher number of marked results and hits than H. Looking at session time, we notice a big difference in the time spent on pages for the two conditions (always showing vs. not-always showing). Therefore, we compute the rate of gain for a fair comparison of user performance. We follow Aliannejadi et al. (2021a) and define *rate of gain* as the number relevant documents per second, formally, $rate\ of\ gain = \#hits/session\ time$. As reported in the last row of Table 7, we see that all of the values are higher for not-always showing condition, indicating that users are more efficient when fewer CQs were shown to them.

Summary. Mostly user behavior metrics increase from low-quality CQs to mid-quality CQs then to high-quality CQs (L \rightarrow M \rightarrow H). This indicates that users from sessions involving high-quality CQs interact with SERPs more deeply and spend more time, and find more relevant information. Compared with low-quality CQs (L) or mid-quality CQs (M), showing low-quality CQs only or showing mid-quality CQs only (M-only) lowers user behavior measures (L vs. L-only & M vs. M-only). Compared with sessions containing low-quality CQs or mid-quality CQs, users from sessions showing low-quality CQs only or showing mid-quality CQs only interact with SERPs less deeply and spend less time, finding less relevant information. Higher efficiency is observed when CQs are not always shown to users. Moreover, the highest improvement in the rate of gain is observed for high-quality CQs (H vs. H-only), suggesting that a selective

posing of high-quality CQs would help users achieve a much higher rate of gain. This is in line with Aliannejadi et al. (2021b) where the necessity of predicting clarification needs in a conversational session is highlighted. This result also suggests that hiding low-quality and mid-quality CQs for a session is beneficial. Our finding is in agreement with Wang and Ai (2021), that low-quality CQs could bring risk to users, suggesting that we should model the risk of low-quality CQs and predict the necessity of showing CQs.

4.3.2. User Satisfaction

From Table 8, we see that user satisfaction rating improves from low-quality CQs to mid-quality CQs to high-quality CQs. Also, based on the percentage of positive and negative ratings across groups, user-perceived helpfulness improve from low-quality CQs to mid-quality CQs to high-quality CQs. Compared with sessions containing high-quality CQs (H), H-only receives better user satisfaction and helpfulness. These observations again confirm that always showing CQs may be risky, especially when showing low-quality CQs. Moreover, L-only receives the lowest user satisfaction ratings and helpfulness, suggesting that low-quality CQs lead to a negative impact on user's experience, which is in line with previous work (Zou et al., 2022).

4.4. Discussion

In this section, we provide a deeper discussion based on our observations and aim to answer the research questions posed earlier.

RQ1: To what extent do different trajectory patterns affect user behavior at the query level? From the results in Figure 3a and Table 4, we observe that in general users tend to engage more with a CQ if it is the first CQ shown in the session. Therefore, showing a high-quality CQ at the beginning of the session is crucial. However, this is a challenging task (Aliannejadi et al., 2019; Zamani et al., 2020a), but leads to useful information about user information needs and preferences as the users provide a response to the initial CQ. We then observe that in case the system fails in asking a high-quality CQ, it can still use the feedback it gets from the CQ to pose a second CQ. Studying the results in the table, we see that asking a higher-quality CQ would lead to an improved user experience.

RQ2: To what extent do different trajectory patterns affect user behavior and satisfaction at the session level? After studying the immediate effect of asking multiple CQs in the RQ1, in this RQ we focus on the effect of CQ trajectories at a session-wide level. This is crucial to understand as it sheds light on various aspects of user behavior and its relation to the RQs, e.g., how much asking a high-quality CQ after a low-quality CQ would affect the user's satisfaction of the session as a whole? From Figure 3b, we see the same pattern during a session, i.e., asking a high-quality CQ after a low-quality CQ leads to more engagement. As we compare the CQ engagement for Lb4H=61% and LbfL=12%, we see that the system can gain up to 5 times higher user engagement in a session if a higher-quality CQ is shown after a lower-quality CQ, giving the opportunity for improved user experience in case of an initial failure. Moreover, as observed from Table 5, for all CQ qualities, an informed decision on the trajectory of showing CQs to the users can lead to a higher rate of gain, which reiterates the importance of predicting the risk (Wang and Ai, 2021) of asking a low-quality CQ.

RQ3: Does always showing CQs benefit or disturb users? From Table 7, we clearly see that systems should employ a method to model the risk of asking a CQ (Wang and Ai, 2021) and predict the clarification need (Aliannejadi et al., 2021b) in a session by taking into account various factors, e.g., query ambiguity (Guo et al., 2021). Showing more CQs clearly leads to higher engagement with the CQs (see Figure 3); therefore, there is a higher risk when users engage with lower-quality CQs. Even though in some cases always showing low-quality CQs lead to a higher number of hits in a session, it comes with the cost of nearly twice as much time as they spend in a session with fewer shown CQs. This is in line with the findings of Aliannejadi et al. (2021a) where they discuss the trade-off between showing more results vs. asking CQs and find that asking more CQs does not always return the same amount of gain (i.e., lower rate of gain).

4.5. Implications

In this section, we also highlight the implications arising from this study, including both theoretical and practical implications.

Theoretical implications Our study advances the understanding of user interactions with CQs in information-seeking systems. Specifically, our results contribute to the understanding of how and why different trajectories of CQs with

different qualities affect user behavior and satisfaction at the query level and session level. To the best of our knowledge, our work is the first effort to discuss the effects of trajectories of asking CQs, and thus enriches the literature on asking CQs in information-seeking systems. Unlike ad-hoc retrieval where a single query would be enough, conversational information-seeking systems by asking CQs are usually evaluated based on a whole session, and thus trajectories are crucial in this setting. By quantifying user behavior and satisfaction caused by asking CQs of different quality, we further disentangle the relationships between the quality of CQs and various aspects of user behavior and satisfaction. Also, our results reinforce the importance of asking CQs in information-seeking systems. We confirm that asking high-quality CQs is beneficial. However, we also provide insights into the extent to which asking low-quality CQs disturbs users and what negative impacts low-quality CQs bring to users. Moreover, our study provides viewpoints on how multiple CQs in a search session affect each other, and how a trajectory of CQs can be used for a better user experience. For example, we confirm that when the system receives negative feedback for the first low-quality CQ, the system can take follow-up actions such as asking CQs of higher quality to regain user satisfaction.

Practical implications Practitioners and researchers can gain practical lessons from these findings in terms of the effect of different trajectories to ask CQs at the query and session levels. Our results have design implications that will allow practitioners to support CQs, and help them to decide in which situations to show CQs so that users can benefit from asking for CQs. As asking CQs in information-seeking systems is still far from widespread adoption, the system providers need a better understanding of user interaction with CQs of different quality. Our study provides a foundation for the industry to develop a normative framework to evaluate the adoption potential of new services regarding asking CQs. Also, our results are able to support the development of CQ-based algorithms. Our study sheds light on asking a CQ with uncertain quality as it may bring risks. Although this is inevitable as the system may provide risky CQs unintentionally, our study provides motivations for modeling risk and urge the researchers to design suitable risk-aware algorithms based on asking CQs (Wang and Ai, 2021). Similarly, users may choose noisy answers when interacting with CQs, our results prompt the researchers to design noise-tolerant algorithms on the basis of asking CQs. Moreover, we investigate the effects of always showing CQs, indicating that always showing CQs leads to a higher user dissatisfaction risk. This motivates the problem of CQ showing prediction. Future research for CQ-based systems would benefit from incorporating CQ showing prediction into their models. Last, our study discovers the relationships between user behavior measures and CQ quality, which can be used for assessing CQ quality. We demonstrate that user behavior and satisfaction are affected by the quality of CQs. This would advance the research on algorithms for generating CQs and selecting CQs in information-seeking systems; the algorithm should incorporate a quality-aware module, either from CQ quality assessment models or from real CQ quality labels (e.g., three-level CQ quality labels in the real dataset for asking CQs in information-seeking systems, MIMICS (Zamani et al., 2020b)), when generating and selecting CQs.

5. CQ Showing Prediction

As we demonstrate that always showing CQs is not optimal, it is necessary to predict whether the system should or not display a CQ to users, to avoid disturbing users with low-quality CQs. To this end, we develop a transformer-based algorithm on top of the user study data to predict the necessity of showing CQs.¹⁰

5.1. Model

We adopt the Transformer framework to predict CQ Showing. We suppose there is a set of CQs, denoted by $C = \{c_1, c_2, \dots, c_{|C|}\}$. Each CQ $c_i \in C$ has a set of corresponding candidate answers A_i . Given an initial query q , a selected CQ c along with its candidate answers A , a list of search results S , and user behavior data Y , we aim to predict the CQ showing signal of the current selected CQ c and its candidate answers A .

$$P(q, c, A, S, Y) = \text{Sigmoid}(\text{Dropout}(W \times \mathbf{h}_{\text{CLS}} + b)) \in [0, 1], \quad (1)$$

$$\mathbf{h}_{\text{CLS}} = \text{ALBERT}(\mathbf{q}, \mathbf{c}, \mathbf{A}, \mathbf{S}, \mathbf{Y}), \quad (2)$$

where $\mathbf{q}, \mathbf{c}, \mathbf{A}, \mathbf{S}$, and \mathbf{Y} are embeddings of q, c, A, S , and Y , respectively. The selected CQ c is a CQ selected by the system intended to show to a user before the CQ showing prediction. For search results S , we use SERP titles in this paper. For user behavior data Y , we use the query position (i.e., i -th query) and mouse hovering time on the previous

¹⁰This data will be released publicly for research purposes.

Table 9

Model performance on CQ showing prediction. Best performances are marked in Bold.

Model	Accuracy	F1
Always showing	0.494	0.661
Always hiding	0.506	0
TranShow	0.755	0.796
-w/o query	0.722	0.767
-w/o CQ pane	0.497	0.500
-w/o SERP titles	0.728	0.788
-w/o query position	0.742	0.794
-w/o hovering	0.517	0.681

showing CQs before the current search query in the search session. If there are no previous showing CQs or no mouse hovering behavior, it is set to zero. We use q , c , A , S , and Y as input factors, as the quality of q , c , A , and S contain cues on how necessary to show the CQ pane, while Y demonstrates the user preference to view the CQ pane. Our ablation experiments in Section 5.2 also demonstrate the importance and contribution of these factors. The model outputs the probability of showing the selected CQs based on the input representations. In our implementation, we use ALBERT (Lan et al., 2019), a lite BERT, as our pre-trained language model for generating embeddings of q , c , A , S , and Y . ALBERT takes the tokenized query q , CQs c , candidate answers A , SERP titles S , behavior information like the query position and mouse hovering time on the previous showing CQs Y as input, separated by the Separation SEP token. A classification token CLS is also inserted at the beginning of the input sequence. In this way, the hidden-state representation of the classification token CLS in the last layer \mathbf{h}_{CLS} is further passed to produce the prediction. The prediction is performed by a Sigmoid function with a linear layer and a dropout layer in between to produce an output probability over CQ showing. During training, we minimize Binary Cross Entropy loss as the nature of the task is a binary classification problem.

We train our model using Adam optimizer (Kingma and Ba, 2014) and Pytorch with a learning rate of 10^{-4} . The hidden dimension is 768, as in the pre-trained language model ALBERT. The maximum sequence length is set to 512 tokens. We use 5-fold cross-validation to evaluate our model. In each fold, the dataset is split into training and testing set by the ratio of 8:2. For ground truth generation, we use the SERPs in which the showed CQ receives a click-through as the positive CQ showing class while others as the negative CQ showing class.

5.2. Performance

From the results reported in Table 9, we observe that the performance is much higher than the straightforward baselines: always showing CQs and always hiding CQs. Hence, CQ showing is predictable and our model to predict CQ showing is effective.

To evaluate the importance of different factors, we further conduct an ablation study comparing our TranShow model with its ablation variants. We refer the TranShow by removing the information of query, CQ pane (i.e., CQs and candidate answers), SERP titles, query position, and CQ hovering time as “-w/o query”, “-w/o CQ pane”, “-w/o SERP titles”, “-w/o query position”, and “-w/o hovering”, respectively. Based on the results in Table 9, all factors contribute to the final performance. Among all factors, the CQ pane and user behavior data of mouse hovering time on the previous showing CQs are the most important factors as they contribute most to the prediction performance.

6. Conclusion and Limitations

In this paper, we conduct a user study to investigate the effects of trajectory of showing CQs in information-seeking systems. We explore the effect of asking different trajectory patterns showing different quality categories of CQs on user behavior and satisfaction at the query- and session-level. Moreover, we propose a transformer-based model, named TranShow, to predict CQ showing to alleviate the disturbance that low-quality CQs might bring to users.

From our analysis, we learn that low-quality CQs disturb users and systems should ask CQs only when necessary. Always showing CQs may be risky and systems should hide CQs with uncertain quality. When a low-quality CQ is shown to users, the system still has the chance to redeem user interest by asking follow-up high-quality CQs.

Furthermore, future research for CQ-based systems should incorporate CQ showing prediction or risk modeling of CQs into their models, instead of simply asking many CQs.

One limitation of this study is the small number (i.e., four) of Web search tasks used. Using a limited number of experimental tasks to control factors while maintaining the cost at a reasonable level is a widely used setting in many studies in the community (e.g., (Collins-Thompson et al., 2016; Harvey and Pointon, 2017; Edwards and Kelly, 2017; Kelly and Azzopardi, 2015; Kelly et al., 2015; Han et al., 2015; Cole et al., 2015; Zou et al., 2022; O'Brien et al., 2020)). Nevertheless, we leave the analysis for more Web search tasks as future work. Also, the number of participants is limited, while a similar setting has been adopted in some previous laboratory user study setups (Collins-Thompson et al., 2016; Harvey and Pointon, 2017; Edwards and Kelly, 2017; Kelly and Azzopardi, 2015; Han et al., 2015; Cole et al., 2015; Kiesel et al., 2018; Vtyurina et al., 2017; Zou et al., 2022; Liu et al., 2019; Xie et al., 2017; Kim, 2008; Yuan et al., 2014). It would be beneficial to extend the user study with more participants from different user groups in the future.

We propose a transformer-based model, TranShow, to predict whether or not to show CQs, based on the collected user study data. We understand that, as a controlled laboratory user study, the data is limited to train a large model. To this end, we utilize the pre-trained language model to alleviate this limitation and the experimental results demonstrate that our proposed model is effective. Nevertheless, it is worth collecting more training data from a large number of users to train a prediction model in the future.

Last, in this paper, we mainly focus on exploring the effects of trajectory patterns with the length of two (i.e., each trajectory pattern involves two CQs). We leave the trajectory patterns with the length of more than two for future work.

CRediT authorship contribution statement

Jie Zou: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Aixin Sun:** Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing - review & editing. **Cheng Long:** Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing - review & editing. **Mohammad Aliannejadi:** Conceptualization, Formal analysis, Software, Writing - review & editing. **Evangelos Kanoulas:** Conceptualization, Formal analysis, Writing - review & editing.

7. Acknowledgment

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from Singapore Telecommunications Limited (Singtel), through Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU). This study is also supported by the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), and the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green, And Integrated Transport (814961). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

- Aliannejadi, M., Azzopardi, L., Zamani, H., Kanoulas, E., Thomas, P., Craswell, N., 2021a. Analysing mixed initiatives and search strategies during conversational search, in: CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, ACM. pp. 16–26.
- Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M., 2020. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). arXiv preprint arXiv:2009.11352.
- Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M.S., 2021b. Building and evaluating open-domain dialogue corpora with clarifying questions, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Association for Computational Linguistics. pp. 4473–4484.
- Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B., 2019. 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, pp. 475–484.
- Avula, S., Choi, B., Arguello, J., 2022. The effects of system initiative during conversational collaborative search. Proceedings of the ACM on Human-Computer Interaction 6, 1–30.
- Belkin, N.J., Cool, C., Stein, A., Thiel, U., 1995. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. Expert systems with applications 9, 379–395.
- Bevendorff, J., Stein, B., Hagen, M., Potthast, M., 2018. Elastic chatnoir: Search engine for the cluweb and the common crawl, in: European Conference on Information Retrieval, pp. 820–824.

- Borlund, P., 2003. The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research* 8, 8–3.
- Chu, Z., Wang, Z., Liu, Y., Huang, Y., Zhang, M., Ma, S., 2022. Convsearch: A open-domain conversational search behavior dataset. *arXiv preprint arXiv:2204.02659*.
- Cole, M.J., Hendahewa, C., Belkin, N.J., Shah, C., 2015. User activity patterns during information search. *ACM Transactions on Information Systems (TOIS)* 33, 1–39.
- Collins-Thompson, K., Rieh, S.Y., Haynes, C.C., Syed, R., 2016. Assessing learning outcomes in web search: A comparison of tasks and query strategies, in: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pp. 163–172.
- De Boni, M., Manandhar, S., 2003. 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*, pp. 48–55.
- DeMoranville, C.W., Bienstock, C.C., 2003. Question order effects in measuring service quality. *International Journal of Research in Marketing* 20, 217–231.
- Edwards, A., Kelly, D., 2017. Engaged or frustrated?: Disambiguating emotional state in search, in: *SIGIR*, pp. 125–134.
- Guo, M., Zhang, M., Reddy, S., Alikhani, M., 2021. Abg-coqa: Clarifying ambiguity in conversational question answering, in: *3rd Conference on Automated Knowledge Base Construction, AKBC 2021*.
- Hagen, M., Potthast, M., Adineh, P., Fatehifar, E., Stein, B., 2017. Source retrieval for web-scale text reuse detection, in: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 2091–2094.
- Han, L., Maddalena, E., Checco, A., Sarasua, C., Gadiraju, U., Roitero, K., Demartini, G., 2020. Crowd worker strategies in relevance judgment tasks, in: *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 241–249.
- Han, S., Yue, Z., He, D., 2015. Understanding and supporting cross-device web search for exploratory tasks with mobile touch interactions. *ACM Transactions on Information Systems (TOIS)* 33, 1–34.
- Harvey, M., Pointon, M., 2017. Searching on the go: The effects of fragmented attention on mobile web search tasks, in: *SIGIR*, pp. 155–164.
- Hashemi, H., Zamani, H., Croft, W.B., 2020. Guided transformer: Leveraging multiple external sources for representation learning in conversational search, in: *SIGIR*, pp. 1131–1140.
- Ho, C.J., Vaughan, J., 2012. Online task assignment in crowdsourcing markets, in: *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Huang, J., White, R.W., Buscher, G., Wang, K., 2012. Improving searcher models using mouse cursor activity, in: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 195–204.
- Kelly, D., 2009. *Methods for evaluating interactive information retrieval systems with users*. Now Publishers Inc.
- Kelly, D., Arguello, J., Edwards, A., Wu, W., 2015. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework, in: *ICTIR*, pp. 101–110.
- Kelly, D., Azzopardi, L., 2015. How many results per page? a study of serp size, search behavior and user experience, in: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 183–192.
- Kiesel, J., Bahrami, A., Stein, B., Anand, A., Hagen, M., 2018. Toward voice query clarification, in: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1257–1260.
- Kim, K.S., 2008. Effects of emotion control and task on web searching behavior. *Information Processing & Management* 44, 373–385.
- Kim, T.K., 2015. T test as a parametric statistic. *Korean journal of anesthesiology* 68, 540.
- Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Krasakis, A.M., Aliannejadi, M., Voskarides, N., Kanoulas, E., 2020. Analysing the effect of clarifying questions on document ranking in conversational search, in: *Proceedings of the 2020 ACM SIGIR on International Conference on The Theory of Information Retrieval*, pp. 129–132.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
- Lipani, A., Carterette, B., Yilmaz, E., 2021. How am i doing?: Evaluating conversational search systems offline. *ACM Transactions on Information Systems*.
- Liu, J., Wang, Y., Mandal, S., Shah, C., 2019. Exploring the immediate and short-term effects of peer advice and cognitive authority on web search behavior. *Information Processing & Management* 56, 1010–1025.
- O’Brien, H.L., Arguello, J., Capra, R., 2020. An empirical study of interest, task complexity, and search behaviour on user engagement. *Information Processing & Management* 57, 102226.
- Potthast, M., Hagen, M., Stein, B., Graßegger, J., Michel, M., Tippmann, M., Welsch, C., 2012. Chatnoir: a search engine for the clueweb09 corpus, in: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 1004–1004.
- Radlinski, F., Craswell, N., 2017. A theoretical framework for conversational search, in: *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pp. 117–126.
- Ren, P., Liu, Z., Song, X., Tian, H., Chen, Z., Ren, Z., de Rijke, M., 2021. Wizard of search engine: Access to information through conversations with search engines, in: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 533–543.
- Schuman, H., Presser, S., 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.
- Sekulic, I., Aliannejadi, M., Crestani, F., 2021. User engagement prediction for clarification in search, in: *Proceedings of the European Conference on Informatin Retrieval (ECIR)*.
- Sekulić, I., Aliannejadi, M., Crestani, F., 2022. Evaluating mixed-initiative conversational search systems via user simulation, in: *WSDM*.
- Sepliaraskaia, A., Kiseleva, J., Radlinski, F., de Rijke, M., 2018. Preference elicitation as an optimization problem, in: *RecSys*, pp. 172–180.
- Shin, D., 2021. The perception of humanness in conversational journalism: An algorithmic information-processing perspective. *New Media & Society*, 1461444821993801.
- Shin, D., Chotiyaputta, V., Zaid, B., 2022a. The effects of cultural dimensions on algorithmic news: How do cultural value orientations affect how people perceive algorithms? *Computers in Human Behavior* 126, 107007.
- Shin, D., Hwang, Y., 2020. The effects of security and traceability of blockchain on digital affordance. *Online information review* 44, 913–932.

- Shin, D., Kee, K.F., Shin, E.Y., 2022b. Algorithm awareness: Why user awareness is critical for personal privacy in the adoption of algorithmic platforms? *International Journal of Information Management* 65, 102494.
- Shin, D., Rasul, A., Fotiadis, A., 2021. Why am i seeing this? deconstructing algorithm literacy through the lens of users. *Internet Research* .
- 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, 449–471.
- Trippas, J.R., Spina, D., Cavedon, L., Sanderson, M., 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis, in: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pp. 325–328.
- Turpin, A., Scholer, F., 2006. User performance versus precision measures for simple search tasks, in: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 11–18.
- Vakkari, P., Völske, M., Potthast, M., Hagen, M., Stein, B., 2019. Modeling the usefulness of search results as measured by information use. *Information Processing & Management* 56, 879–894.
- Vtyurina, A., Savenkov, D., Agichtein, E., Clarke, C.L., 2017. Exploring conversational search with humans, assistants, and wizards, in: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2187–2193.
- Wang, J., Li, W., 2021. Template-guided clarifying question generation for web search clarification, in: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 3468–3472.
- Wang, Z., Ai, Q., 2021. Controlling the risk of conversational search via reinforcement learning. *arXiv preprint arXiv:2101.06327* .
- White, M.D., Iivonen, M., 2001. Questions as a factor in web search strategy. *Information Processing & Management* 37, 721–740.
- White, R.W., Bilenko, M., Cucerzan, S., 2007. Studying the use of popular destinations to enhance web search interaction, in: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 159–166.
- Xie, X., Liu, Y., de Rijke, M., He, J., Zhang, M., Ma, S., 2018. Why people search for images using web search engines, in: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 655–663.
- Xie, X., Liu, Y., Wang, X., Wang, M., Wu, Z., Wu, Y., Zhang, M., Ma, S., 2017. Investigating examination behavior of image search users, in: *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, pp. 275–284.
- Xu, J., Wang, Y., Tang, D., Duan, N., Yang, P., Zeng, Q., Zhou, M., Xu, S., 2019. 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)*, pp. 1618–1629.
- Yuan, J., Sivrikaya, F., Marx, S., Hopfgartner, F., 2014. When to recommend what? a study on the role of contextual factors in ip-based tv services .
- Zamani, H., Dumais, S., Craswell, N., Bennett, P., Lueck, G., 2020a. Generating clarifying questions for information retrieval, in: *Proceedings of The Web Conference 2020*, pp. 418–428.
- Zamani, H., Lueck, G., Chen, E., Quispe, R., Luu, F., Craswell, N., 2020b. Mimics: A large-scale data collection for search clarification, in: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3189–3196.
- Zamani, H., Mitra, B., Chen, E., Lueck, G., Diaz, F., Bennett, P.N., Craswell, N., Dumais, S.T., 2020c. Analyzing and learning from user interactions for search clarification. *arXiv preprint arXiv:2006.00166* .
- Zhang, Y., Chen, X., Ai, Q., Yang, L., Croft, W.B., 2018. Towards conversational search and recommendation: System ask, user respond, in: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 177–186.
- Zou, J., Aliannejadi, M., Kanoulas, E., Pera, M.S., Liu, Y., 2022. Users meet clarifying questions: Toward a better understanding of user interactions for search clarification. *ACM Transactions on Information Systems (TOIS)* .
- Zou, J., Chen, Y., Kanoulas, E., 2020a. Towards question-based recommender systems, in: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 881–890.
- Zou, J., Kanoulas, E., 2019. Learning to ask: Question-based sequential bayesian product search, in: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 369–378.
- Zou, J., Kanoulas, E., Liu, Y., 2020b. An empirical study on clarifying question-based systems, in: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2361–2364.