RESEARCH-ARTICLE

How Users Interact with Generative Information Retrieval Systems: A Study of User Behavior and Search Experience

YIDONG LIANG, Beijing Institute of Technology, Beijing, China

ZHIJING WU, Beijing Institute of Technology, Beijing, China

FAN ZHANG, Wuhan University, Wuhan, Hubei, China

DANDAN SONG, Beijing Institute of Technology, Beijing, China

HEYAN HUANG, Beijing Institute of Technology, Beijing, China

Open Access Support provided by:

Beijing Institute of Technology

Wuhan University

Published: 13 July 2025

SIGIR '25: The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval
July 13 - 18, 2025
Padua, Italy

Conference Sponsors:
SIGIR

https://doi.org/10.1145/3726302.3729998
ISBN: 9798400715921

How Users Interact with Generative Information Retrieval Systems: A Study of User Behavior and Search Experience

Yidong Liang
School of Computer Science and Technology, Beijing Institute of Technology
Beijing, China
lyd@bit.edu.cn

Zhijing Wu*
School of Computer Science and Technology, Beijing Institute of Technology
Beijing, China
zhijingwu@bit.edu.cn

Fan Zhang
Wuhan University
Wuhan, China
frankyzf94@gmail.com

Dandan Song
School of Computer Science and Technology, Beijing Institute of Technology
Beijing, China
sdd@bit.edu.cn

Heyan Huang
School of Computer Science and Technology, Beijing Institute of Technology
Beijing, China
hhy63@bit.edu.cn

Abstract

The development of LLM has facilitated the emergence of generative information retrieval (IR) systems, such as "Bing Chat". Generative IR systems return generated text with citations rather than a list of ranked search results. User studies on IR systems are essential for understanding users' interaction patterns, evaluating and optimizing systems, and improving search experience, particularly in the context of generative IR systems with novel conversational interfaces and responses. However, systematic investigations into user behavior and search experience on generative IR systems are notably lacking. To address this gap, we conducted a user study using Bing Chat to explore user behavior and feedback on generative IR systems. The participants were required to accomplish three types of tasks using Bing Chat. During the search process, we collected their various behavior (e.g., click, query reformulation) and explicit feedback (e.g., satisfaction, credibility, and success). Additionally, the same study was conducted on traditional IR systems Bing for comparison. Analyses of these data show that Bing Chat can reduce the user's search effort and lead to a better search experience without any decrease in credibility compared with Bing. We believe that this work provides valuable insight into the design and evaluation of generative information retrieval systems.

CCS Concepts

• **Information systems → Users and interactive retrieval**.

Keywords

Generative Information Retrieval System, User Study, Search Behavior, Search Experience

ACM Reference Format:
Yidong Liang, Zhijing Wu, Fan Zhang, Dandan Song, and Heyan Huang. 2025. How Users Interact with Generative Information Retrieval Systems: A Study of User Behavior and Search Experience. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3726302.3729998

*Corresponding author

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3729998

1 Introduction

Web search engines have been essential tools for humans to find information quickly and efficiently online for decades. They enable users to search for specific information, documents, or resources on the Web. Recently, the advancement of large language models (LLMs) has driven the emergence of generative information retrieval (IR) systems (e.g., Bing Chat). Most of these systems primarily rely on retrieval systems to gather relevant information, which is then used by LLMs to generate responses [15, 16]. As shown in Figure 1, Bing Chat interacts with users in the form of a conversation, giving summarized answers with two types of citations: superscript links (sup-link) and "learn more" links (learn-link). Sup-link and learn-link usually appear at the end of a sentence in the summarized answers and at the bottom of the entire answer, respectively. In contrast, traditional IR systems (e.g., Bing) give a list of search results consisting of title, URL, and snippet on the Search Engine Result Page (SERP). Users need to find useful information from the SERP or landing pages and summarize answers by themselves. Therefore, generative IR systems have the potential to reduce users' effort and improve their search experience, such as their perceived satisfaction and success.

Generative IR systems have drawn significant attention in recent years. Most studies are system-centric, aiming to optimize the systems by improving the quality of generated responses and citations. For example, instructing models to simulate human search behavior, using retrieval systems to enhance answer generation

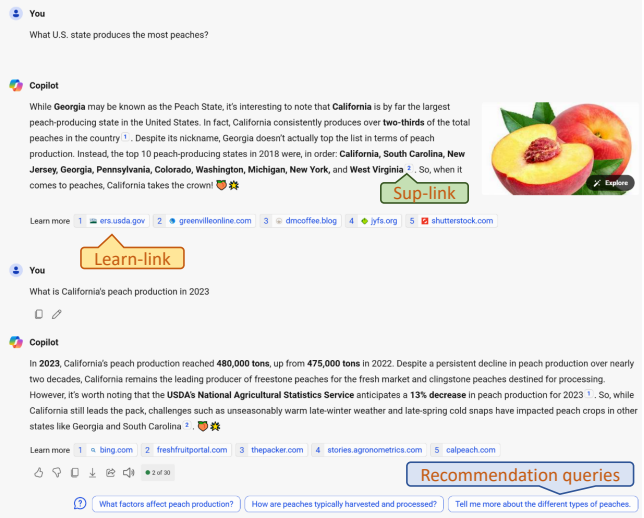

Figure 1: The interaction interface of Bing Chat. For convenience, we refer to this conversation interface as the search engine result page (SERP), and the webpage users are redirected to after clicking a link as the landing page.

through instruction fine-tuning or context learning [15, 23, 42, 49]. SearChain [57] generates reasoning chains, with the retrieval module interacting with the nodes of the chain. More strategies, such as post-hoc retrieval enhancement and re-generation, are also proposed to improve output quality [29, 59]. In contrast, user-centric research remains limited and typically focuses on specific scenarios or phenomena [39, 48, 50, 51, 61]. Gienapp et al. [16] provided theoretical insights into the user model and evaluation on generative IR systems, offering a foundation for further research. For users, generative IR systems pose risks such as increasing selective exposure, creating echo chambers, and contributing to opinion polarization [46]. However, there is a lack of systematic research on user behavior and experience in generative IR systems.

User-centric research on user search behavior and search experience is crucial for the development of IR systems. On the one hand, understanding how users interact with IR systems helps designing user-friendly and efficient interaction interfaces [55, 62]. On the other hand, user-centric studies contribute to several information retrieval tasks such as relevance estimation [1, 20], satisfaction prediction [17, 25, 26, 58], document ranking [31, 41], and search evaluation [7, 10]. Common search behaviors (e.g., click, mouse movement) can be extracted into interaction sequences to predict user satisfaction [40], and incorporating features derived from these behaviors can improve ranking performance [2]. Based on click behavior, many popular evaluation metrics have been proposed [10, 11, 60]. However, these studies are conducted in the scenario of traditional IR systems. There is still a lack of investigation of user behavior and search experience in generative IR systems where the form of interaction is different from traditional ones.

In this paper, we aim to investigate users' information-seeking process in generative IR systems, and compare user behavior and search experience (e.g., satisfaction) in generative IR systems with those in traditional ones. Our research questions are as follows:

- **RQ1:** How do users interact with the generative IR system, and how do user behavior patterns differ between generative IR systems and traditional IR systems?
- **RQ2:** Compared to traditional IR systems, can generative IR systems enhance users' experience in search?

To address these research questions, we carried out a controlled lab-based user study to collect users' behavior and explicit feedback when using both the generative IR system and the traditional one. Through data analysis, we find that the generative IR system can reduce the search effort of users and lead to higher users' perceived search satisfaction and success than the traditional IR system without any decrease in credibility, especially in exploratory tasks. To summarize, the main contributions are as follows:

- We construct a lab-study based search dataset[1] that includes 960 search sessions on the generative IR system (i.e., Bing Chat) and traditional IR system (i.e., Bing). The dataset contains various user behavior data and explicit feedback such as clicks, query reformulations, cumulative satisfaction, and objective/subjective task success (Section 2).
- We conduct a thorough investigation into how users interact with the generative IR system, concentrating on user behavior patterns and search experiences. We believe this research offers valuable insights into user behavior and search experiences within this novel IR system paradigm, which can help the development and evaluation of generative IR systems (Section 3 and Section 4).

2 Data Collection

We aim to investigate users' search behavior and experiences when using generative IR systems. **User behavior** refers to the various interactions of the user during the search process, including mouse movement, clicking, and query reformulation. **Search experience** generally refers to user perceptions and responses from using a search system [21]. We focus on the search experience reflected by users' explicit feedback, including satisfaction, credibility, and subjective success. Considering the potential gap between subjective experience and objective success [33], we also annotated objective success for each search task. The data collection procedure is illustrated in Figure 2. It consists of two steps: I. User Study and II. Objective Success Assessment.

2.1 Task design

Given that task complexity and difficulty have been recognized as important factors that affect user behavior [4, 5, 12, 44], we design tasks with different complexity levels. According to Marchionini [37], search tasks are commonly divided into two categories: lookup and exploratory. The lookup task search is more straightforward than exploratory search tasks, which have high objective task complexity [52]. In this paper, we focus on users' information-seeking process in generative IR systems. We design two types of lookup search tasks (Factual tasks and Mis-Factual tasks) and exploratory tasks.

[1] https://github.com/BITLYDG/User-study-dataset

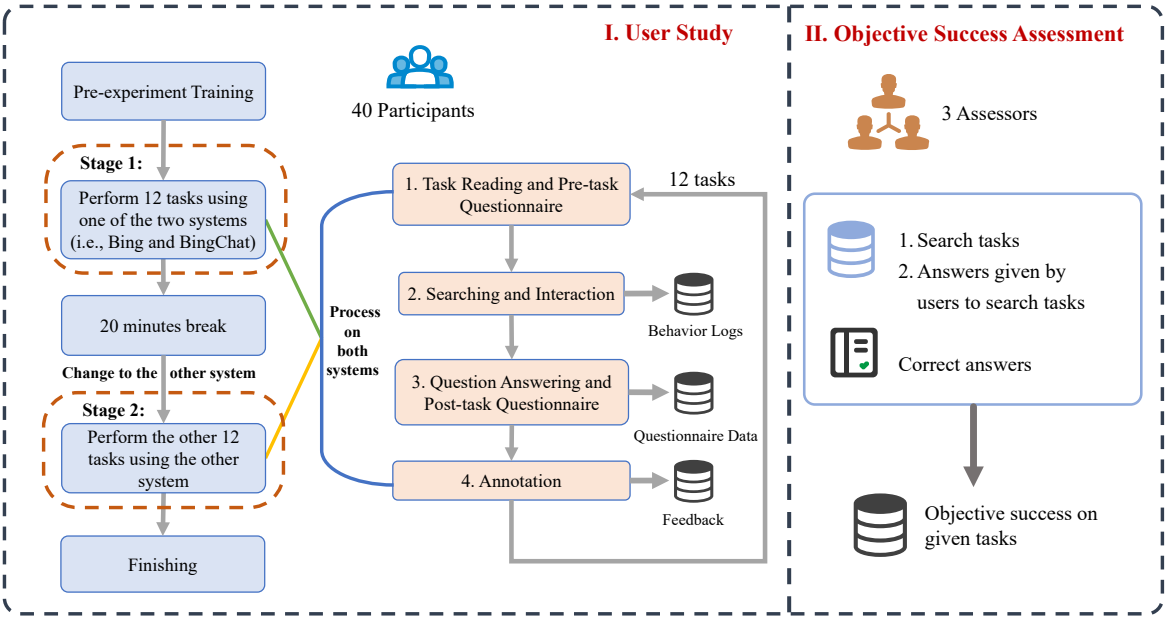

Figure 2: The procedure of our user study and task objective success assessment.

- **Factual tasks**. Factual tasks aim to locate facts or other similar information items. Most of these tasks can be solved with specific entity-based answers (e.g., providing a representative composer of Neapolitan opera).
- **Mis-Factual tasks**. Misconceptions are facts and views that are usually believed to be correct but are wrong. We define factual tasks with misconceptions as mis-factual tasks (e.g., is it illegal to chew gum in Singapore?).
- **Exploratory tasks**. Exploratory tasks usually contain an open-ended information-seeking goal and require more cognitive, learning, and information-gathering skills (e.g., why ultraviolet disinfection cannot completely supplant chlorination when disinfecting drinking water?).

Compared to factual tasks, mis-factual and exploratory tasks are more complex in two dimensions: the reliability of information and the amount of information required. Mis-factual tasks are characterized by the presence of conflicting viewpoints within search results, requiring users to expend additional effort to discern the reliability of the information. Exploratory tasks demand a greater quantity of information than factual tasks. Users need to collect, understand, and integrate information to form answers. These factors may contribute to distinct user behaviors and search experiences. We designed factual tasks based on T^2Ranking [56], a large-scale dataset with over 300k queries sourced from real-world search engine interactions. We avoid common-sense tasks to minimize the impact of the user's self-knowledge. We designed mis-factual tasks based on the TruthfulQA [32] dataset and the Wikipedia page "List of common misconceptions". The TruthfulQA [32] dataset and the Wikipedia page "List of common misconceptions" contain many facts and views that are usually believed to be correct but are wrong. We first chose some questions and submitted related queries to some popular commercial search engines (i.e., Google and Baidu). Based on the SERPs, we selected questions with more incorrect results than factual questions. The exploratory tasks are designed according to Liu et al. [33], utilizing the same 6 tasks and 2 additional tasks. The evaluation criteria (key points and weights) used in the following sections are consistent with Liu et al. [33]. We manually selected 24 search tasks, composed of 8 factual, 8 mis-factual, and 8 exploratory tasks, as our search tasks.

2.2 User study

The procedure of the user study is shown in Figure 2. We developed a user study platform using Django. Participants can log into the platform and complete search tasks. We recorded participants' explicit feedback and behavior, including mouse movements, scrolling, click behaviors, and timestamps. We divided the 24 search tasks into two groups, *G1* and *G2*. Each group contains 12 search tasks, comprising 4 factual tasks, 4 mis-factual tasks, and 4 exploratory tasks. To ensure fairness and avoid bias between the two systems, participants were divided into groups of four. Each participant in the same group conducted a two-stage study, completing tasks in one of the following orders: [(G1, Bing), (G2, Bing Chat)], [(G2, Bing), (G1, Bing Chat)], [(G1, Bing Chat), (G2, Bing)], or [(G2, Bing Chat), (G1, Bing)]. [(G1, Bing), (G2, Bing Chat)] means that the participant completed tasks in G1 using Bing in the first stage, then completed tasks in G2 using Bing Chat in the second stage, with the other orders following the same pattern. The tasks in G1 or G2 were presented in random order. In this way, each task was tested on both systems and prevented order effects.

In each stage, the study began with a warm-up task to familiarize participants with the procedure and experiment platform. Participants were instructed to find the answer using IR systems with no time limit for each task. Breaks were allowed between tasks. The

tasks were presented in random order to minimize order bias. The procedure for each search task is as follows:

Task Reading and Pre-task Questionnaire. First, the participant was asked to read the task and memorize it. After fully understanding the task, the participant needed to fill out the pre-task questionnaire assessing their familiarity with the task and its perceived difficulty. The participant gives feedback through a 5-point Likert scale (1: not at all, 2: slightly, 3: somewhat, 4: moderately, 5: very).

Searching and Interaction. Second, participants freely interacted with IR systems according to their habits and completed search tasks. While using Bing, they could reformulate queries, examine or click on a result to read a landing page, click on the buttons of next or previous pages (there are 10 search results per page), and so on as they wish. When using Bing Chat, they can complete tasks by engaging in conversation with the system and clicking on links. When participants felt they had gathered enough information to complete the search task or could not find more, they could end the search. We recorded user logs, including mouse movements and scrolling, clicking behaviors, and the timestamps corresponding to the various behaviors.

Question Answering and Post-task Questionnaire. After participants finished searching, they needed to answer the following three questions: 1) Do you feel that the information you have gathered can successfully answer the questions of the search task? (5-point Likert scale); 2) the credibility of results. (5-point Likert scale); 3) Please give your answer to the search task. This explicit feedback from participants helped us better understand their search experience. Participants could leave the answer blank if they had not found the answer.

Annotation. For both IR systems, participants were asked to annotate the usefulness of the results/links they examined (0: not at all, 1: somewhat, 2: fairly, 3: very). In addition to this, users were required to highlight text spans on SERPs they considered useful and annotate the cumulative satisfaction (5-point Likert scale) for each system response (on Bing Chat)/session (on Bing, we refer to session as the response to facilitate the subsequent explanation). Formally, the cumulative satisfaction for the ith response r_i, denoted as $CSat_i$, represents the overall satisfaction for responses $[r_1, r_2, r_3, ..., r_i]$. The user's cumulative satisfaction on a given task is represented as $[CSat_1, CSat_2, CSat_3, ..., CSat_n]$, where n is the number of system responses. Cumulative satisfaction reflects changes in satisfaction during interactions. The task-level satisfaction is the last value in the cumulative satisfaction list (In the following text, we will refer to task-level satisfaction as **user satisfaction**). The instruction about cumulative satisfaction is as follows: *Please provide a cumulative satisfaction score for each response/session. This score represents your overall satisfaction with the interaction from the beginning up to and including the current response/session.* We confirmed that users understood our instructions by asking them to explain cumulative satisfaction and differentiate between cumulative satisfaction and single-response satisfaction.

In our collected data, we observed that query reformulations on Bing were infrequent. Therefore, the cumulative satisfaction lists for most tasks on Bing contain only a single value. Moreover, users considered most of the useful text to be located on landing pages rather than on SERPs. Consequently, there are few highlighted text spans on SERPs. Therefore, in Section 3, we only report the analyses of highlighted text spans on Bing Chat.

Participants repeated the above procedure until all 12 tasks were finished with one IR system. The participants then repeated the same procedure with the other IR system to complete the other group of 12 tasks. We ensure that each task was performed an equal number of times on both systems to avoid bias.

2.3 Objective success assessment

In the post-task questionnaire, users rated their perceived success on each task, which we refer to as subjective success. Additionally, three external assessors evaluated participants' answers, and the scores were used as objective success. For factual tasks and mis-factual tasks, there is a single answer for each task, so the user's response is evaluated as either correct or incorrect (i.e., a binary assessment, 0: wrong, 1: correct). The value of Fleiss kappa [14] was 0.896, indicating a high level of consistency [28]. We used the majority vote of three assessors as the objective success.

For exploratory tasks, answers are more complex and require several answer sentences or passages. Users may provide a portion of correct answers, making binary evaluation unsuitable. We take the same approach as Liu et al. [33], using the same key points and criteria to evaluate the user's answer. For each exploratory task, we have key points listed by experts and evaluate the answer based on how many key points are included in the user's answer. Formally, all key point importance for a task can be represented as $[s_1, s_2, s_3, ..., s_n]$, where s_i is the importance score of the ith key point. The existence of key points in the answer can be represented as $[p_1, p_2, p_3, ..., p_n]$, where $p_i = 0/1$ means that the ith key point is not included/included in the answer. The final score for an answer can be calculated as equation 1.

$$Objective\_Success = \frac{\sum_{i=1}^{n} s_i \cdot p_i}{\sum_{i=1}^{n} s_i}. \quad (1)$$

We used the Intraclass Correlation Coefficient (ICC1) to evaluate the assessment results. An inter-assessor agreement of 0.864 was obtained, indicating good agreement.

2.4 Data statistics

We recruited 40 campus participants (23 men, 17 women), all undergraduate or graduate students. We designed 24 tasks in total, each participant was required to complete 12 tasks using Bing and the other 12 tasks using Bing Chat, receiving $15 compensation. The following three kinds of data were collected. (1) user behavior: users' interactions during the search process, including queries, click behaviors, mouse movements, etc. (2) explicit feedback, which includes the search experience: what users fill out in the questionnaires. (3) objective success: three external assessors evaluate the answers submitted by participants. In total, we collected 960 search sessions, with 480 sessions per IR system and 320 sessions per type of task. Based on the statistics of the pre-task questionnaire, there is a significant difference in participants' perceived difficulty across the three task types. The difficulty value is 1.54 on factual tasks, 1.85 on mis-factual tasks, and 2.83 on exploratory tasks. Through tasks with varying degrees of difficulty, the user study can comprehensively explore the behavior and experience. In terms of familiarity,

the average values are all lower than somewhat familiar (3 out of 5-point Likert scale) in three types of tasks. It shows that these tasks are unfamiliar to participants and can reduce the impact of participants' self-knowledge.

3 Behavior Patterns

In this section, we investigate users' behavior patterns on Bing Chat and Bing to address RQ1. There are two variables in our comparative analysis: IR system and task type. We use non-parametric tests rather than the t-test or ANOVA analysis for the significance test because the distributions of most of the data are not normal. When comparing data from different IR systems, we use the Wilcoxon signed-rank test [53]/Mann-Whitney U test [36] when pairwise conditions are met/not met. For example, there is task time on both systems (meeting pairwise condition) for each task completed by each user, but not necessarily click behavior (not meeting pairwise condition). When comparing the data from the three types of tasks, we use the Kruskal-Wallis test [27]. When conducting a two-way analysis, we use the Scheirer–Ray–Hare test [45] with one variable being the IR system and the other being the type of tasks.

3.1 Task time

Table 1 shows the statistics of task time. When using Bing Chat, users spend nearly identical time on factual and mis-factual tasks, with mis-factual tasks even taking slightly less time (on Bing, Mis-Factual tasks require 24.3% more time). This difference may arise from Bing Chat providing a clear, singular answer, reducing confusion from contradictory information. On exploratory tasks, users spend significantly more time due to task complexity. In addition, there is no significant difference in the percentage of time users spent on SERPs across the three types of tasks.

On average, users spend less time using Bing Chat than using Bing on all search tasks, mis-factual tasks, and exploratory tasks. This suggests that Bing Chat can reduce user effort and time costs. Moreover, on Bing Chat, users spend 83% of dwell time on the SERPs and only 17% browsing the landing pages. This is a significant difference compared to Bing, where users spend over 50% of dwell time browsing landing pages. It indicates that users tend to focus more on text, rather than reference links that lead to landing pages on Bing Chat. It is worth noting that the task time on factual tasks on Bing Chat is slightly higher than that on Bing. It may be because Bing Chat retrieves the relevant web pages before generating an answer, which takes more time than Bing. Users spend more time waiting for responses on Bing Chat. Therefore, on factual tasks, where the tasks are relatively easy to address, Bing Chat does not exhibit a shorter task time than Bing. Additionally, we can observe a larger reduction in time on the exploratory task compared to the Mis-Factual task (16.1% vs 12.4%).

3.2 Reformulation

From the results in Table 1, we observe that the number of queries on Bing Chat is more than that on Bing. The generative IR system in the form of conversation may slightly improve users' willingness to submit more queries. Maybe it is because the SERP of Bing has more results, and users tend to explore the results rather than submit additional queries. In Bing Chat, users can reformulate queries manually

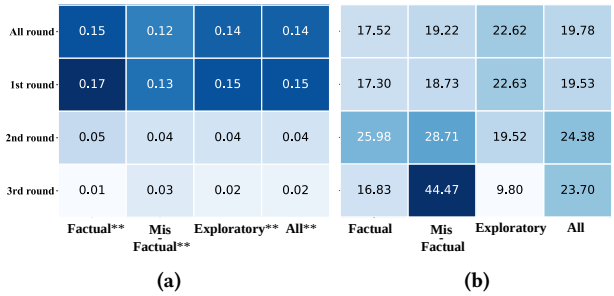

Figure 3: Statistics of (a) click rate and (b) dwell time of clicks in the different interaction rounds with Bing Chat. '*/" indicates that the differences among rounds are statistically significant at $p < 0.05/0.01$ level.**

or by clicking recommendations. All queries following the initial one are considered query reformulation. Manual reformulation accounts for 59% and the ratio of clicking on a recommendation is 41%. It indicates that recommendation queries are also an important part of the interaction between users and Bing Chat.

3.3 Click behavior

Table 1 shows the statistics of click behavior and the usefulness of clicked links. When analyzing user click behavior, considering that users are likely to click more than once on a task, we calculate the maximum, minimum, and average click dwell time, as well as the usefulness of the clicks. The reported values are averages of these statistics across all tasks, while other task-level statistics, which have only one value per task, are directly averaged across all tasks.

Number of clicks. From Table 1, we find that users click on fewer links on Bing Chat. The Scheirer–Ray–Hare test revealed that both the IR system and task type significantly affect the number of clicks. On Bing Chat, it is interesting to note that there is no obvious difference in the number of clicks on different types of tasks. Even on complex exploratory tasks, users clicked only 1.04 links on average. However, users click significantly more on the exploratory tasks than on the factual and mis-factual tasks when using Bing. The number of clicks on Bing Chat does not necessarily indicate the complexity of the task. It is worth noting that about 53% of sessions on Bing Chat involved no clicks, compared to only 9% on Bing. This suggests that users often find sufficient information through text on Bing Chat, with no need to click in over half of the tasks. The ratios of tasks without clicks on Bing Chat follow a similar trend to the number of clicks, with no significant difference across task types. Even on complex exploratory tasks, only 47% of sessions involve click behaviors (98% of sessions on Bing). It indicates that in many cases, users do not click on reference links at all on Bing Chat.

Click dwell time. The results in Table 1 show a slight increase in dwell time across all task types when using Bing Chat. However, when using Bing, the differences in the dwell time are more pronounced. Users spend significantly less time on landing pages when using Bing Chat on all types of search tasks. It indicates that Bing Chat can reduce user effort and time costs on landing pages. Bing Chat provides summarized answers to help users understand

Table 1: Statistics of user behavior and self-annotated usefulness. "↑/↓" indicates that the value on Bing Chat is greater/smaller than that on Bing. "*/" after "↑"or"↓" indicates that the differences between Bing Chat and Bing are statistically significant at $p < 0.05/0.01$ level. "†/‡/∗" in the first column indicates the system/task type/their interaction has a statistically significant at $p < 0.05$ level.**

	Factual			Mis-Factual			Exploratory			All		
	Bing	Chat		Bing	Chat		Bing	Chat		Bing	Chat	
Task time (s)‡*	61.02	67.45	↑	75.83	66.45	↓	121.55	101.95	↓*	86.13	78.61	↓
SERP time ratio†‡*	0.55	0.82	↑**	0.46	0.82	↑**	0.35	0.85	↑**	0.46	0.83	↑**
Number of queries†‡	1.03	1.31	↑**	1.03	1.21	↑**	1.14	1.32	↑**	1.07	1.28	↑**
Ratio of clicking recommendation queries	-	0.39	-	-	0.35	-	-	0.48	-	-	0.41	-
Number of clicks†‡*	1.69	1.07	↓**	1.78	0.97	↓**	2.56	1.04	↓**	2.01	1.03	↓**
Ratio of tasks without click	0.18	0.52	↑	0.11	0.53	↑	0.02	0.53	↑	0.09	0.53	↑
Max.click dwell time (s)†‡*	26.92	26.26	↓	38.66	24.06	↓**	44.34	31.30	↓**	37.15	27.22	↓**
Min.click dwell time (s)†‡*	15.47	11.85	↓*	28.50	14.75	↓*	20.60	15.95	↓**	21.63	14.17	↓**
Avg.click dwell time (s)†‡*	20.78	17.52	↓	33.39	19.22	↓**	31.57	22.62	↓**	28.94	19.78	↓**
Max.usefulness of clicks†‡	2.86	2.65	↓*	2.77	2.67	↓	2.73	2.44	↓**	2.79	2.58	↓**
Min.usefulness of clicks†‡*	2.16	2.07	↓	1.68	2.19	↑**	1.64	1.93	↑	1.83	2.06	↑*
Avg.usefulness of clicks‡*	2.52	2.35	↓	2.24	2.41	↑*	2.18	2.20	↑	2.31	2.32	↑

the content of web pages and enhance browsing efficiency. The IR system and the type of tasks both have a significant influence on the click dwell time according to the Scheirer–Ray–Hare test.

Usefulness of clicked links. On all tasks, the average usefulness is similar for both IR systems, there is no significant difference. However, on mis-factual tasks, Bing Chat has a higher average usefulness. This aligns with the characteristics of mis-factual tasks, where there are often incorrect answers on the internet, causing confusion for users and leading to lower usefulness. Bing Chat alleviates this by providing an aggregated answer. On all tasks, the minimum and maximum usefulness of the two systems show opposite patterns. Bing Chat has a higher minimum usefulness, especially on mis-Factual tasks, while Bing has a higher maximum usefulness, especially on exploratory tasks. Bing Chat employs a degree of filtration when presenting reference links, resulting in relatively higher minimum usefulness. Bing provides a significantly greater number of links, leading to a broader range of usefulness and higher maximum usefulness.

Click behavior in multi-round interaction. We further investigate how click rate and dwell time of clicks change as the conversation rounds increase. As mentioned in 2.2, query reformulations by users on Bing were infrequent. Given this and our study focuses on the generative IR system, we only conducted analysis on Bing Chat. Our analysis focuses on the first three turns of conversations, which account for the majority of interactions. Figure 3 shows the statistic results. As the number of conversation rounds increases, the click rate decreases significantly, and users show a decreasing desire to click. The reduction is more pronounced on the factual task. Due to increased costs and improved responses, they may lean towards not clicking and directly accepting the textual answers. However, it is worth noting that the average dwell time does not follow this trend, there is no significant difference across rounds. This suggests that while users' desire to click decreases, the time spent on landing pages remains roughly the same.

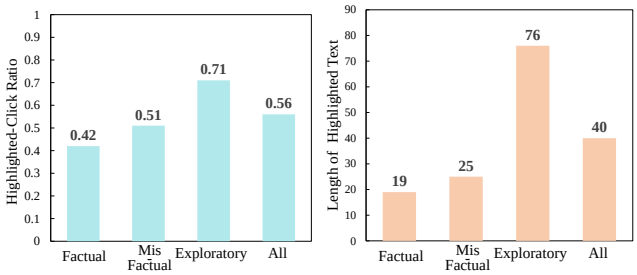

Figure 4: (a) The ratio of clicked links whose corresponding text is highlighted as useful text. (b) The length of highlighted text.

Highlighted text and clicked links. Previous studies show that users' click behavior correlates with snippet usefulness. We explore whether this still holds under Bing Chat. We plot the proportion of clicked links whose corresponding text is highlighted as useful in Figure 4a. We find that except for exploratory tasks, the ratio of highlighted clicks is close to 50%. It shows that even if the corresponding text is not useful, users may want to examine the landing page of the link for more information. We also plot the length of the highlighted text in Figure 4b. We find that the highlighted text is much longer on exploratory tasks than on the other types of tasks. This could be why there is a higher ratio of highlighted clicks on the exploratory tasks: answers to exploratory tasks are more complex, resulting in the highlighted useful text being longer and more likely to contain links.

3.4 Summary

In this section, we conduct an in-depth analysis of user behavior on the generative IR system. We summarize our findings as follows:

- The generative IR system can reduce the effort of users (e.g., shorter task duration and fewer clicks than those on the traditional IR system) by providing summarized answers in natural language (refer to Table 1).
- On the generative IR system, users' clicking behavior shows smaller variations across different types of tasks than those on the traditional IR system, particularly between factual and exploratory tasks (refer to Table 1).
- In query reformulation, the proportion of clicking on recommendation queries averages 41% across all tasks. This indicates that recommendation queries are crucial for user interaction with Bing Chat (refer to Table 1).
- On the generative IR system, the number of clicks decreases significantly as the interaction rounds increase (refer to Figure 3).
- On the generative IR system, users clicking on a link does not indicate that they think the corresponding text to the link is useful (refer to Figure 4).

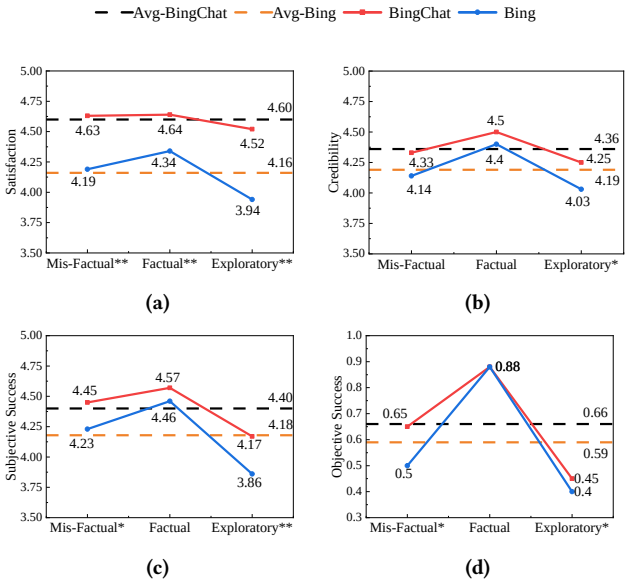

Figure 5: Users' subjective feedback and objective success. Differences between two IR systems in all tasks are statistically significant at $p < 0.01$ level. "*/" behind the task type indicates that differences between two IR systems in tasks of this type are statistically significant at $p < 0.05/0.01$ level.**

4 Search Experience

In this section, we investigate users' search experience to address RQ2. We collected explicit feedback including user satisfaction, subjective task success, and credibility from participants, and objective success annotated by external assessors. A detailed analysis of explicit feedback is presented to explore whether generative IR systems can better assist users. The statistics of explicit feedback are shown in Figure 5.

Satisfaction. The satisfaction is shown in Figure 5a. When using Bing Chat, users are most satisfied with factual tasks and least satisfied with exploratory tasks. Moreover, user satisfaction varied less across all types of tasks (0.12 on Bing Chat vs 0.4 on Bing). This finding aligns somewhat with the user behavior in Section 3. User behavior on Bing Chat shows lower efforts, such as task time and number of clicks, with these behaviors showing a smaller gap across the three task types compared to Bing. Compared to Bing, we find that user satisfaction with Bing Chat is significantly higher across all types of tasks ($p < 0.01$). In terms of the gap between the two IR systems, mis-factual and exploratory tasks are higher than factual, especially exploratory tasks (0.58). For mis-factual tasks, users usually can access a summarized answer that does not include conflicting viewpoints or facts through Bing Chat, which helps to reduce confusion and improve user search experience. For exploratory tasks, Bing Chat delivers well-summarized answers directly, saving the effort required for secondary processing such as viewing search results and summarizing content.

Credibility. As shown in Figure 5b, we find that Bing Chat is considered most credible for factual tasks, while its credibility is perceived as lowest for exploratory tasks. Although mis-factual tasks are generally more misleading, users still assign lower credibility to exploratory tasks. On Bing Chat, users give higher levels of credibility, even though people usually don't trust AI enough because of hallucinations. However, through integrating search technologies and citations, user credibility on generative IR systems is not diminished compared to traditional IR systems (even higher in our study). Regarding the gap between the two IR systems, mis-factual and exploratory tasks are higher than factual. However, significant differences are observed only in exploratory tasks.

Subjective Success. Figure 5c shows that users are more confident in factual tasks, experience a slight decrease in subjective success in mis-factual tasks, and encounter a notable decrease in exploratory tasks when using Bing Chat. Users are more confident that they would be able to complete the task with Bing Chat on any one type of task compared. As with satisfaction and credibility, the gap between the two IR systems is higher for mis-factual and exploratory tasks than for factual tasks. Moreover, satisfaction and subjective success, as two important subjective ratings, show slight differences. The gap in subjective success on both IR systems is relatively smaller compared to the satisfaction (0.22 vs 0.44). This suggests that users' high satisfaction with Bing Chat may be due in part to factors other than the content, such as friendly and natural ways of interaction. However, when it comes to the effectiveness of the returned content in resolving issues, users tend to be more cautious, although their subjective success still exceeds that of Bing.

Objective Success. The objective success is shown in Figure 5d. Overall, users get higher objectively assessed success using Bing Chat. On factual tasks, close scores were obtained for both systems. On the other two types of tasks, Bing Chat achieved statistically significantly higher scores. The objective success shows the largest discrepancy (0.15) on mis-factual tasks. Generative IR systems, trained on extensive corpora, possess inherent knowledge enabling them to discern certain misconceptions. Therefore, Bing Chat can assist users in more effectively completing mis-factual tasks. Additionally, the conversation-style responses of Bing Chat contribute to the resolution of exploratory tasks. However, both IR systems struggle

Table 2: Statistics of users' subjective feedback and objective success under tasks with click and w/o click. "*/" in the "w/o click" row indicates that the differences between click and w/o click are statistically significant at $p < 0.05/0.01$ level.**

Task type	Behavior	Satis-faction	Credi-bility	Subjective-success	Objective success
Factual	with click	4.48	4.39	4.47	0.86
	w/o click	4.80**	4.60	4.66	0.90
Mis-Factual	with click	4.45	4.16	4.36	0.65
	w/o click	4.79**	4.47*	4.53	0.65
Exploratory	with click	4.38	4.24	4.08	0.46
	w/o click	4.64*	4.26	4.25	0.44
All	with click	4.44	4.26	4.30	0.66
	w/o click	4.74**	4.44*	4.48*	0.66

with exploratory tasks (0.45 on Bing Chat and 0.4 on Bing). Despite Bing Chat's ability to summarize search results into natural language answers, manual inspection and comprehensive searching are still necessary for exploratory tasks.

In conclusion, when using Bing Chat, users' subjective ratings and objective success are highest on factual tasks and lowest on exploratory tasks. We find that users give higher subjective ratings to Bing Chat on any one type of task. Users are more satisfied with the results of Bing Chat, more confident that they would be able to complete the task and give higher levels of credibility. In terms of the gap between the two IR systems, mis-factual and exploratory tasks are higher than factual, especially exploratory tasks. For mis-factual tasks, there are more wrong results in SERPs on Bing. This can lead to confusion and make it challenging for users to identify the correct result. Therefore, the search experience can be negatively impacted. However, Bing Chat mitigates this by providing a summarized answer, improving the search experience. For exploratory tasks, users need to manually examine and integrate the information to form an answer. Bing Chat simplifies these tasks by providing conversational summaries, which make it easier for users to find the information they need. Therefore, the differences in subjective feedback between the two IR systems tend to be more noticeable on mis-factual and exploratory tasks.

To gain a deeper understanding of the search experience on Bing Chat, we further analyze the relationship between search experience and click behavior. We categorize search sessions into two groups: those with clicks and those without clicks. The results are shown in Table 2. We find that users' explicit feedback is significantly higher on any type of search task without clicks than those with clicks. It indicates that "no click" is a potential signal for a satisfying search experience. When users choose not to click, they likely believe summarized answers contain enough credible information and do not require further examining, in which case the search experience is usually better. There is no significant difference in the objective assessed success when considering the click as variable.

Finally, we conduct analysis on users' cumulative satisfaction. We find that cumulative satisfaction could decrease throughout interactions (with 14.9%/38.1% of adjacent rounds showing a decrease/increase). We conducted interviews with users. They reported that certain responses made them dissatisfied, leading to a

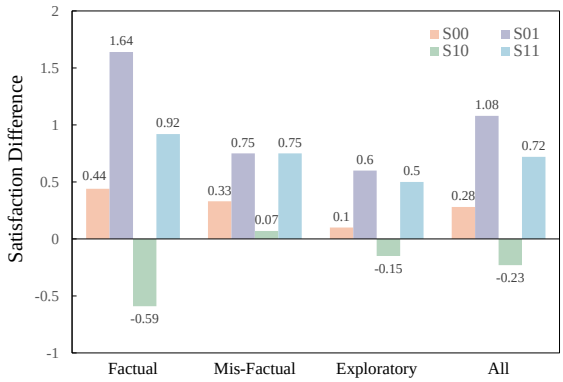

Figure 6: Cumulative satisfaction differences between two adjacent rounds of interactions (current round minus the previous round). The notion "Sij" with i (or j) = 0/1 means that there is not/is highlighted content in the previous (or current) response.

decrease in their cumulative satisfaction scores. We further analyze the relationship between users' highlighted useful text and cumulative satisfaction. There are four scenarios based on whether two adjacent responses contain highlighted text. "Sij" with $i = 0/1$ represents that there is no/is highlighted text in the previous response, and $j = 0/1$ represents there is no/is highlighted text in the current response. For example, $S01$ denotes that the previous response does not contain useful information to users but the current response contains useful information to users. The cumulative satisfaction changes under these four scenarios are shown in Figure 6. We find that when users get useful information in the previous response but there is no useful information in the current response, their cumulative satisfaction decreases obviously in factual tasks. If the current response contains useful information, users' cumulative satisfaction increases on all tasks.

In this section, we summarize our findings as follows:

- Users' subjective ratings and objective success are highest on factual tasks and lowest on exploratory tasks on Bing Chat. This finding also holds true on Bing. (refer to Figure 5).
- Users give higher subjective ratings to Bing Chat than to Bing on any type of task. The improvement is relatively small on factual tasks and more significant on the other two types, especially on exploratory tasks (refer to Figure 5).
- Bing Chat significantly improves the objectively assessed success on mis-factual and exploratory tasks. Interestingly, although users perceive a higher success on factual tasks when using Bing Chat than using Bing, the objective success remains unchanged (refer to Figure 5d).
- When using Bing Chat, users report higher satisfaction with search tasks without clicks than those with clicks, although there is no significant difference in objective success (refer to Table 2).
- Cumulative satisfaction may decrease during search. It tends to increase/decrease when the current response is more/less useful than the previous response (refer to Figure 6).

5 Related Work

5.1 Generative IR systems

Generative IR refers to a range of approaches that combine retrieval and generative components to address a task [16]. These approaches can be categorized into generation-augmented retrieval and retrieval-augmented generation. In this paper, we focus on users' interactions with the system rather than the implementation. Traditional IR systems present search results as a list of documents in SERPs [8, 19]. However, the generative IR system interacts with users through a conversation. For each query, the system generates a summarized text response that includes citations, pointing to the sources of the referenced webpage [13, 30]. Conversational IR systems are defined as IR systems designed to enable interactive communication between a user and an agent [43]. Although both systems are conversational, conversational IR systems do not emphasize specific response formats, such as including citations.

The research on generative IR systems has focused on system optimization. Early methods instruct models to simulate human search behavior for autonomous information retrieval [42, 49]. Retrieved content enhances generation by guiding LLMs through instruction fine-tuning or context learning [15, 23]. Subsequently, more advanced strategies have been proposed. For example, SearChain [57] generates reasoning chains, with the retrieval module interacting with the nodes of the chain. Other strategies, such as post-hoc retrieval enhancement and re-generation, aim to improve output quality. For example, CEG [29] applies the retrieval augmentation module after the generation process, and the model regenerates the answer when necessary. Additionally, NLI models are used to validate the consistency between generated answers and retrieved documents [59]. Nowadays, many applications combine generative models with search, such as Bing Chat, You Chat, and Gemini.

Some user-centric research has been conducted in the generative IR scenario [6, 34, 48, 50, 51, 61]. In the image geolocation task, participants using traditional IR systems outperformed those relying on generative IR systems [51]. In this scenario, query reformulation on generative IR systems proved to be more challenging. Zelch et al. [61] conducted a user study on the searcher side effects of ads in generative IR systems and found that integrating ads with content often goes unrecognized by users. Generative IR systems may amplify confirmatory querying and opinion polarization, especially when an opinionated LLM reinforces users' perspectives, potentially creating echo chambers [46]. Gienapp et al. [16] integrates the characteristics of generative IR systems with user modeling theories from traditional IR systems. It considers the utility model, browsing model, and accumulation models during users' interactions with generative IR systems. Building on this, a theoretical framework for evaluating generative IR systems was also developed.

These user-centric studies focus on specific scenarios or particular phenomena. There is a lack of comprehensive user studies to understand how users interact with generative IR systems.

5.2 User behavior and search experience

Researchers have conducted thorough investigations into user behavior and search experience in traditional IR systems [12, 24, 47]. User behavior is typically analyzed through interaction data such as queries [9] and clicks [3]. While search experience is often analyzed through subjective feedback, such as user satisfaction [38]. He and Yilmaz [18] conducted a field study in which participants' daily search and browsing activities were recorded. Users were asked to annotate the tasks they performed. Wu et al. [54] conducted a field study in an image search context. Through the study, a large volume of real-world search behavior data was collected with extensive first-level annotations. Jiang et al. [22] carried out a user study to collect user behavior from search sessions involving complex tasks. Participants were asked to perform different types of search tasks, with a focus on user behavior. Wu et al. [55] explored how the inclusion of answer modules in search engine results pages (SERPs) affected user behavior by comparing factoid and non-factoid queries across different SERP settings. Sharma et al. [46] set up controversial topics. They collected users' attitudes before and after using different search systems to uncover the risks of selective exposure and echo chambers in LLM-powered search systems.

We conducted a user study in a controlled lab setting to gather both user behavior and explicit feedback while interacting with both generative and traditional IR systems.

6 Conclusion and Limitations

User behavior and search experience play a crucial role in various information retrieval tasks. However, user behavior and search experience on generative IR systems have not been thoroughly investigated as an emerging type of IR system. In this study, we conducted a user study in which participants were asked to complete three types of search tasks using a generative IR system and a traditional IR system, respectively. The behavior data and feedback were collected directly from the participants. We study patterns of user behavior on Bing Chat, a generative IR system, and results show that the generative IR system can reduce the search effort and bring in a better search experience. As with any research, there are potential limitations to our experiments. First, it is a controlled, lab-based user study with only 40 participants from a university. However, it is unclear how a broader population of searchers might or might not employ similar search strategies. There are only 24 tasks in this study, and artificially designed search tasks can potentially affect user behavior. In addition, the study is performed under a controlled environment and may not accurately reflect users' true search intent or practical behavior patterns. To enhance the generalizability of our findings, it would be interesting to reaffirm our results by conducting a large-scale field study to directly obtain data from users' own daily information needs. Another limitation is that the interface in our experiments differs slightly from the commercial engine. Our study excluded sponsored results on Bing due to their demonstrated impact on user behavior on SERPs [35]. The interface of Bing Chat has been updated to a newer version (Copilot), and the previous version has been discontinued. Despite these differences, we believe the main results remain valid, as the basic interaction mode of Bing Chat has remained unchanged.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62302040), the China Postdoctoral Science Foundation (No. 2022TQ0033), and the Beijing Institute of Technology Research Fund Program for Young Scholars.

References

[1] Mikhail Ageev, Dmitry Lagun, and Eugene Agichtein. 2013. Improving search result summaries by using searcher behavior data. In *Proceedings of the 36th international acm sigir conference on research and development in information retrieval*. 13–22.

[2] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 19–26.

[3] Ioannis Arapakis, Xiao Bai, and B. Barla Cambazoglu. 2014. Impact of response latency on user behavior in web search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (Gold Coast, Queensland, Australia) *(SIGIR '14)*. Association for Computing Machinery, New York, NY, USA, 103–112. doi:10.1145/2600428.2609627

[4] Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2635–2651.

[5] Katriina Byström. 2002. Information and information sources in tasks of varying complexity. *Journal of the American Society for information Science and Technology* 53, 7 (2002), 581–591.

[6] Robert Capra and Jaime Arguello. 2023. How does AI chat change search behaviors? *arXiv preprint arXiv:2307.03826* (2023).

[7] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 621–630.

[8] Danqi Chen, Weizhu Chen, Haixun Wang, Zheng Chen, and Qiang Yang. 2012. Beyond ten blue links: enabling user click modeling in federated web search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (Seattle, Washington, USA) *(WSDM '12)*. Association for Computing Machinery, New York, NY, USA, 463–472. doi:10.1145/2124295.2124351

[9] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of the web conference 2021*. 743–755.

[10] Aleksandr Chuklin and Maarten de Rijke. 2016. Incorporating clicks, attention and satisfaction into a search engine result page evaluation model. In *Proceedings of the 25th acm international on conference on information and knowledge management*. 175–184.

[11] Aleksandr Chuklin, Pavel Serdyukov, and Maarten De Rijke. 2013. Click model-based information retrieval metrics. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 493–502.

[12] Michael J Cole, Chathra Hendahewa, Nicholas J Belkin, and Chirag Shah. 2015. User activity patterns during information search. *ACM Transactions on Information Systems* 33, 1 (2015), 1–39.

[13] Hanane Djeddal, Pierre Erbacher, Raouf Toukal, Laure Soulier, Karen Pinel-Sauvagnat, Sophia Katrenko, and Lynda Tamine. 2024. An Evaluation Framework for Attributed Information Retrieval using Large Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 5354–5359. doi:10.1145/3627673.3679172

[14] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[15] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6465–6488. doi:10.18653/v1/2023.emnlp-main.398

[16] Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Evaluating Generative Ad Hoc Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1916–1929. doi:10.1145/3626772.3657849

[17] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2019–2028.

[18] Jiyin He and Emine Yilmaz. 2017. User Behaviour and Task Characteristics: A Field Study of Daily Information Behaviour. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 67–76. doi:10.1145/3020165.3020188

[19] Marti A Hearst. 2009. *Search user interfaces*. Cambridge university press.

[20] Jeff Huang, Ryen W White, Georg Buscher, and Kuansan Wang. 2012. Improving searcher models using mouse cursor activity. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 195–204.

[21] BSEN ISO and BRITISH STANDARD. 2010. Ergonomics of human-system interaction. *British Standards Institution* (2010).

[22] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 607–616.

[23] Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883* (2023).

[24] Diane Kelly and Leif Azzopardi. 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* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 183–192. doi:10.1145/2766462.2767732

[25] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. 2014. 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*. 895–898.

[26] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM international conference on Web search and data mining*. 193–202.

[27] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952), 583–621.

[28] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[29] Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. 2024. Citation-Enhanced Generation for LLM-based Chatbots. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 1451–1466. doi:10.18653/V1/2024.ACL-LONG.79

[30] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2024. From Matching to Generation: A Survey on Generative Information Retrieval. *CoRR* abs/2404.14851 (2024). doi:10.48550/ARXIV.2404.14851 arXiv:2404.14851

[31] Xiangsheng Li, Jiaxin Mao, Chao Wang, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Teach machine how to read: reading behavior inspired relevance estimation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 795–804.

[32] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3214–3252. doi:10.18653/v1/2022.acl-long.229

[33] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. "Satisfaction with Failure" or "Unsatisfied Success": Investigating the Relationship between Search Success and User Satisfaction. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1533–1542. doi:10.1145/3178876.3186065

[34] Sijie Liu, Yuyang Hu, Zihang Tian, Zhe Jin, Shijin Ruan, and Jiaxin Mao. 2024. Investigating Users' Search Behavior and Outcome with ChatGPT in Learning-oriented Search Tasks. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Tokyo, Japan) *(SIGIR-AP 2024)*. Association for Computing Machinery, New York, NY, USA, 103–113. doi:10.1145/3673791.3698406

[35] Zeyang Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. 2015. Influence of Vertical Result in Web Search Examination. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 193–202. doi:10.1145/2766462.2767714

[36] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[37] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[38] David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. 2017. A Study of Snippet Length and Informativeness: Behaviour, Performance and User Experience. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 135–144. doi:10.1145/3077136.3080824

[39] Kerstin Mayerhofer, Rob Capra, and David Elsweiler. 2025. Blending Queries and Conversations: Understanding Tactics, Trust, Verification, and System Choice in Web Search and Chat Interactions. *arXiv preprint arXiv:2504.05156* (2025).

[40] Rishabh Mehrotra, Imed Zitouni, Ahmed Hassan Awadallah, Ahmed El Kholy, and Madian Khabsa. 2017. User interaction sequences for search satisfaction

prediction. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*. 165–174.

[41] Taesup Moon, Georges Dupret, Shihao Ji, Ciya Liao, and Zhaohui Zheng. 2010. User behavior driven ranking without editorial judgments. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 1473–1476.

[42] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

[43] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 117–126. doi:10.1145/3020165.3020183

[44] Tuukka Ruotsalo, Jaakko Peltonen, Manuel JA Eugster, Dorota Głowacka, Patrik Floréen, Petri Myllymäki, Giulio Jacucci, and Samuel Kaski. 2018. Interactive intent modeling for exploratory search. *ACM Transactions on Information Systems* 36, 4 (2018), 1–46.

[45] C James Scheirer, William S Ray, and Nathan Hare. 1976. The analysis of ranked data derived from completely randomized factorial designs. *Biometrics* (1976), 429–434.

[46] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1033, 17 pages. doi:10.1145/3613904.3642459

[47] Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User intent, behaviour, and perceived satisfaction in product search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 547–555.

[48] Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Sathish Manivannan, Nagu Rangan, and Longqi Yang. 2024. The Use of Generative Search Engines for Knowledge Work and Complex Tasks. *CoRR* abs/2404.04268 (2024). doi:10.48550/ARXIV.2404.04268 arXiv:2404.04268

[49] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

[50] Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. 2024. What do Users Really Ask Large Language Models? An Initial Log Analysis of Google Bard Interactions in the Wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2703–2707. doi:10.1145/3626772.3657914

[51] Albatool Wazzan, Stephen MacNeil, and Richard Souvenir. 2024. Comparing Traditional and LLM-based Search for Image Geolocation. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval* (Sheffield, United Kingdom) *(CHIIR '24)*. Association for Computing Machinery, New York, NY, USA, 291–302. doi:10.1145/3627508.3638305

[52] Ryen W White and Resa A Roth. 2009. *Exploratory search: Beyond the query-response paradigm*. Number 3. Morgan & Claypool Publishers.

[53] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*. Springer, 196–202.

[54] Zhijing Wu, Yiqun Liu, Qianfan Zhang, Kailu Wu, Min Zhang, and Shaoping Ma. 2019. The Influence of Image Search Intents on User Behavior and Satisfaction. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 645–653. doi:10.1145/3289600.3291013

[55] Zhijing Wu, Mark Sanderson, B. Barla Cambazoglu, W. Bruce Croft, and Falk Scholer. 2020. Providing Direct Answers in Search Results: A Study of User Behavior. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1635–1644. doi:10.1145/3340531.3412017

[56] Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, et al. 2023. T2ranking: A large-scale chinese benchmark for passage ranking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2681–2690.

[57] Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024. Search-in-the-Chain: Interactively Enhancing Large Language Models with Search for Knowledge-intensive Tasks. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1362–1373. doi:10.1145/3589334.3645363

[58] Ya Xu and David Mease. 2009. Evaluating web search using task completion time. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 676–677.

[59] Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective Large Language Model Adaptation for Improved Grounding and Citation Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6237–6251. doi:10.18653/v1/2024.naacl-long.346

[60] Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. 2010. Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (Toronto, ON, Canada) *(CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 1561–1564. doi:10.1145/1871437.1871672

[61] Ines Zelch, Matthias Hagen, and Martin Potthast. 2024. A User Study on the Acceptance of Native Advertising in Generative IR. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval* (Sheffield, United Kingdom) *(CHIIR '24)*. Association for Computing Machinery, New York, NY, USA, 142–152. doi:10.1145/3627508.3638316

[62] Jie Zou, Mohammad Aliannejadi, Evangelos Kanoulas, Maria Soledad Pera, and Yiqun Liu. 2023. Users meet clarifying questions: Toward a better understanding of user interactions for search clarification. *ACM Transactions on Information Systems* 41, 1 (2023), 1–25.