What do Users Really Ask Large Language Models?

An Initial Log Analysis of Google Bard Interactions in the Wild

Johanne R. Trippas
RMIT University
Melbourne, Australia
j.trippas@rmit.edu.au

Sara Fahad Dawood Al Lawati
RMIT University
Melbourne, Australia
s3919996@student.rmit.edu.au

Joel Mackenzie
The University of Queensland
Brisbane, Australia
joel.mackenzie@uq.edu.au

Luke Gallagher*
The University of Melbourne
Melbourne, Australia
gallagher.l@unimelb.edu.au

ABSTRACT

Advancements in large language models (LLMs) have changed information retrieval, offering users a more personalised and natural search experience with technologies like OpenAI ChatGPT, Google Bard (Gemini), or Microsoft Copilot. Despite these advancements, research into user tasks and information needs remains scarce. This preliminary work analyses a Google Bard prompt log with 15,023 interactions called the *Bard Intelligence and Dialogue Dataset* (BIDD), providing an understanding akin to query log analyses. We show that Google Bard prompts are often verbose and structured, encapsulating a broader range of information needs and *imperative* (e.g., directive) tasks distinct from traditional search queries. We show that LLMs can support users in tasks beyond the three main types based on user intent: informational, navigational, and transactional. Our findings emphasise the versatile application of LLMs across content creation, LLM writing style preferences, and information extraction. We document diverse user interaction styles, showcasing the adaptability of users to LLM capabilities.

CCS CONCEPTS

• **Information systems → Query log analysis**.

KEYWORDS

Large Language Models, Log Analysis, Prompt Analysis, Dataset

ACM Reference Format:
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 (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3657914

*Work performed while the author was at RMIT University.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657914

1 INTRODUCTION

Large language models (LLMs) have emerged as a technology easily accessible by everyday users. Advances in natural language processing and computation have enabled the rapid adoption of LLM-based assistants. Users from diverse backgrounds can now readily access tools such as ChatGPT or Bard (now known as Gemini) with an internet-connected client. Despite continuous efforts to enhance LLM performance, there has been a lack of research focused on *how these systems are used*. The significance of such an analysis extends beyond academic interest; it is crucial for identifying the nuances of user information needs, model efficacy, and improvements to user experience [9, 15]. By examining interaction logs, we seek to document the user—patterns, preferences, novel task intents, and challenges—in their LLM engagements.

In this study, we create the BIDD interaction log, consisting of 15,023 Google Bard[1] interactions from 95 unique respondents[2] based on their use *in the wild*. We examine user demographics and interaction patterns, focusing on information-seeking behaviours and novel LLM interactions. Our approach mirrors search engine query log analyses, which have significantly influenced information retrieval (IR) both academically and commercially [3–5, 11]. Drawing parallels, we compare LLMs to search engines, where user *prompts* are akin to search queries [17, 19].

Despite efforts to enhance LLM performance, there has been minimal focus on the actual information needs and tasks of real-world users [8, 16]. To this end, we perform an initial investigation of interaction logs to provide factual evidence—moving beyond mere anecdotal narratives of how LLMs are used in the wild—about the characteristics of user interactions with LLM-based assistants.

Our findings indicate that users mainly engage in short sessions; however, they also *(i)* go beyond keyword search with non-trivial action-oriented prompts (i.e., complex commands), *(ii)* engage in interactive dialogue and exploration rather than simply consuming information; *(iii)* expect information personalisation; *(iv)* use LLMs for monotonous or repetitive tasks like data extraction or arithmetic; and *(v)* employ LLMs for higher-order tasks like code generation, data analysis, or creative writing. We release a subset of 1,000 anonymised prompts, known as the *Bard Intelligence and Dialogue Dataset* (BIDD-1k) as a resource to facilitate further LLM research.[3]

[1] https://bard.google.com
[2] Note: users, workers, participants, and respondents are used interchangeably.
[3] https://bit.ly/BIDD-1k-logs

2 METHODOLOGY

We conduct an online crowdsourcing study to collect user interactions with Google Bard. We gather data from self-identified Bard users on the Prolific crowdsourcing platform. Workers were asked to fill out a survey in two stages. In *Stage 1*, qualification, they were asked about *(i)* eligibility and willingness to download and upload Google Bard interaction logs, *(ii)* demographics information, *(iii)* general technology usage, and *(iv)* Bard usage. In *Stage 2*, the data upload, workers had to upload and review their interactions.

Crowd Workers. Crowd workers on Prolific could access our task if they met specific requirements:[4] residing in the United States, United Kingdom, Ireland, Australia, Canada, or New Zealand; using Google Bard before 1 January 2024; being able to download and upload their Bard logs; being over 18 years old; fluent in English; having completed over 1,000 Prolific studies; and maintain an approval rate exceeding 95%. Workers were paid 3 GBP for the study. The reward was estimated based on the average completion time for all pilot *Stage 1* tasks (5 minutes), pilot *Stage 2* tasks (10 minutes), and the Australian minimum hourly wage. A maximum time of 30 and 44 minutes was set by Prolific. Upon completion of *Stage 1* and meeting study criteria, workers were invited to participate in *Stage 2*, which involved reviewing and uploading their historical Google Bard logs. A worker could participate in the study once. An attention check was used, involving a multiple-response question for workers to select the LLMs they use. Options included Google Bard, Bing Chat, Claude, Cohere, Falcon, LLaMa, Mistral, ChatGPT, Pi, and "other" (free text field). Workers who failed to select Google Bard were not invited to *Stage 2*. We compensated respondents who submitted empty logs, but their data was deemed invalid and was excluded. The crowd worker task to collect Bard log history was completed with RMIT ethics approval between 2–3 January 2024.

Self-Reported Demographics. The average self-reported participant age is 39 years (min. 18, max. 75, std 11). Most participants (78.95%) described themselves as male, 15.79% as female, and 4.21% as non-binary/third gender (1.05% preferred not to say). The majority (73.68%) of participants were geographically in the United States, with the United Kingdom following at 26.32%. Nearly all participants reported having native English language skills (95.79%), 2.11% reported being fluent, and 2.10% that they had good and moderate English skills. Education levels varied among participants, with the majority holding a university degree (18.95% with a graduate or professional degree, and 47.37% with a bachelor's degree), indicating a high educational accomplishment within the sample. The third largest group (17.89%) had pursued higher education without completing a degree. The sample exhibited representation from participants with secondary and vocational or similar educational backgrounds (both 6.32%) and some secondary (1.05%). Only 2.11% of the respondents preferred not to share their educational background.

Self-reported Search Skills, Search Engine and LLM Usage. The majority of respondents consider their skills as moderately good (54.74%), followed by those who perceive their skills as extremely good (34.74%), slightly good (8.42%), and neither good nor bad (2.11%). No respondents rated their skills as bad in any category. Most respondents use a search engine more than six times daily (60%). The second largest group (27.37%) indicated they use a search engine two to five times daily. Participants indicating they use a search engine once a day account for 4.21%. The remaining respondents indicated that they used search engines between once and six times a week (1.05%), two to three times a week (4.21%), and once a week (3.16%). The majority of respondents indicated that they used more than one LLM frequently, with many of the respondents reporting they had used multiple LLMs. Besides all respondents using Google Bard (95 mentions), 80 mentioned using ChatGPT, and 36 stated use of Bing Chat. Claude, Pi, LLaMA, and others show progressively fewer users, demonstrating a diverse yet concentrated interest in a few leading, readily available models. This distribution spotlights the competitive landscape of LLMs, with a clear preference for early-released or popular platforms.

Pre-processing Crowd Worker Data. In this section, we discuss pre-processing related to all data (i.e., 15,023 entries).[5] Participants exported their Bard history via Google Takeout before performing a mandatory client-side review and then submitting them for analysis. A Bard log submission is a list of interactions in chronological order. An interaction may be either *(i)* a user prompt and corresponding response from Bard, *(ii)* a user rating a response as good or bad, or *(iii)* a user preference selection for a response with multiple drafts. Records that were invalid, explicit, or users having less than 10 interactions were discarded. User sessions were identified using a 15-minute interval. After processing, the data contained 14,595 user prompts, 407 user ratings, and 21 user preference selections. In analyses that follow, prompt text is case-folded and additional whitespace stripped.

3 RESULTS

BIDD consists of 15,023 interactions from 95 unique respondents, with 14,595 instances of prompts, 407 instances of response rating feedback, and 21 instances of response draft preference feedback. On average, a respondent submitted 158 interactions (min. 10, max. 1,261, std. 250). The interaction log contains 4,666 sessions with 2,090 (44.79%) consisting of one interaction. On average, sessions lasted around four minutes. Sessions with more than one interaction lasted around seven minutes; the longest session lasted 194 minutes. The average sessions per user were 49 (med. 17, min. 1, max. 402, and std. 77). Interactions per session averaged 3.22 (min. 1, max. 88, and std. 4.97), indicating variability in session (or task) engagement; an example session is provided in Table 1.

Table 1: Example session (ID 4552).

ID	Prompt Text
1	How old do you have to be to open up a Roth IRA
2	I want to start investing on behalf of my nephew in an account for him. How would I do that
3	Can I open up a UTMA for my nephew?
4	If I wanted to start investing on behalf of my nephew, should I just do so in my own account and transfer to him eventually?
5	Tell me everything I need to know about UTMA in a few sentences. Explain as if I'm an average adult. Can I create this account for my nephew?

[4]Note, these are Prolific's screeners; we further specify our sample within our survey.

[5]Further details for BIDD-1k can be found in the release file.

3.1 Explicit User Feedback

We investigate the 407 interactions (2.71%) in BIDD that provide explicit user feedback to Google relating to the performance of Bard. Of the workers who submitted their logs, 54.74% gave feedback at least once. Notably, 25.55% of the received feedback came from a single respondent. A total of 111 feedback items included written feedback, accounting for 27.27% of all feedback interactions; the remainder were ratings only.

Analysis of the respondent's feedback — both *good* and *bad* — reveals that 50.24% of feedback was positive (1.67% including written feedback), while 49.76% were negative (with 26.56% including written feedback). Pre-defined Bard feedback categories show that 25.60% of the time, respondents identified that the system was "not factually correct", 8.37% "unspecified", 2.87% "irrelevant", and 2.39% "offensive/unsafe". Lastly, 10.53% indicated that the response was bad but did not leave further information.

Workers appreciated Bard's responses to general knowledge queries, especially responses with URLs. Positive feedback was common for advice or case scenario discussions, while generative tasks like coding or drafting emails received mixed reactions. Negative feedback was notable for mathematical, directional, and image generation tasks, along with occasional misinterpretations by Bard. Feedback patterns varied, with some workers consistently positive and others fluctuating. Satisfaction levels differed across task types, showing a balanced mix of positive and negative feedback.

3.2 Diversity in Prompt Interactions

To enhance our understanding of user behaviour, input, and preferences, we investigate the dataset text in both verbatim and normalised forms to examine the frequently posed prompts (FPPs). In verbatim form, the top-4 frequent interactions are response feedback (e.g., the top-1 interaction "Gave feedback: Good response" occurred 210 (1.40%) times). With the normalised BIDD, as explained in Section 2, the dataset was refined to a total of 14,595 prompts, with 14,050 (96.25%) of them being unique. This indicates a rich diversity of prompts within the dataset, with prompt length contributing to this diversity. When we investigate the FPPs in the BIDD prompt set, it shows that user interactions are focused on content creation, emphasising specificity, such as requests for additional information, language preferences, and visual content generation. The FPPs include desires for more content (0.23%), replies in English (0.18%), and image generation tasks (0.17%), highlighting user engagement. Requests for first-person (0.16%) perspectives indicate an immersive and personalised storytelling preference. Meanwhile, prompts for affirmation (0.08%), conciseness (0.08%), like asking for less verbosity (0.08%), shorter (0.08%) and confirmation questions (0.08%), reflect that users prefer clarity and brevity or perhaps results such as a search engine result page. Lastly, "are you sure" (0.07%), highlights the desire for verification.

***N*-grams.** Function and stop words (e.g., "the", "of", "in") dominate the unigram and bigram frequencies, reflecting their presence in natural language. Trigrams and four-grams show more context-specific and natural language phrases, such as questions ("what is the") and references to document structure ("in the first person", "on the next page"), hinting towards what people are trying to achieve with Bard, and reflecting expected LLM capabilities. When we remove the stopwords, there is a shift to more *action-oriented* words (e.g., "make", "write", "answer"), related to instructions and tasks for generating content. This shift towards action-oriented language indicates that people use the system to create tasks, instructions, descriptions, and generate content.

Interaction Management or Navigation. In web search, interactions like "more" or "next page" are essential for users to access, navigate, and consume search results. Bard's interface is conversational, and we found evidence of users accessing these navigational interactions via instruction, such as the user asking for "more", similar to spoken conversational information seeking behaviours [14]. When users request "more" information, the system can interpret this within the discourse's context when providing new content. This contextual understanding may enrich the user's search experience by offering on-topic information. Additionally, we noted respondents exploring their Google account, such as inquiries about the most recent email in their inbox. These actions, viewed as pseudo-navigational or personal information management tasks, may indicate a desire to access interconnected data sources. This navigation, combined with Bard's context-aware search enhancements, may promise to enrich the search experience by integrating traditional search with personalised data queries.

3.3 Prompt Characteristics

Next, we outline some key characteristics of the prompts to further understand how respondents interact with Bard.

Prompt Length. We compare BIDD prompts with MS MARCO queries [1] to understand if users have adopted a more verbose style when interacting with intelligent assistants compared to concise search queries. We found a contrasting difference in average lengths—43.45 words for BIDD versus 6.36 words for MS MARCO—demonstrating that BIDD prompts by nature are more information-dense. Many respondents were able to vary their interaction style to target the platform's intended use, while 36% of users had an average prompt length of 15 terms or less, indicating the use of a "web-search" interaction style characterized by short queries.

One-word Prompts. Single-term prompts — 1.7% of all prompts — are often instructions related to the preceding interaction. The most frequent single-term prompts include "more" (34), "yes" (12), "shorter" (11), "thanks" (11), "answer" (6), "why" (6), "continue" (5), "no" (5), and "summarize" (5). Some respondents would supply a URL directly, without any other context, and Bard would typically respond with a summary of the article at the given URL. This is contrasted with rarity of one-word queries in MS MARCO (0.004%).

Long Prompts. The longest prompts—some longer than 4,000 terms—often contain text or data pasted from other sources. Some examples drawn from the ten longest prompts include questions from a university examination, a request for critical feedback on a (book) character excerpt, and prompts for assistance debugging programming tasks. Interestingly, one user pasted a list of deposits for a ten-month period and asked for the sum of these to be reported; Bard reported a reasonable sounding—but unfortunately incorrect—total. Ironically, another user with a session of long prompts was using Bard to draft a talk on the ethical use of AI tools, and lessons

that can be learnt from pop culture references such as HAL 9000 from A. C. Clarke's *"Space Odyssey"*.

User Intents and Interrogative words. Interrogative words are vital in forming questions in everyday language and are often used to identify intents. Interrogative words include "what", "who", "where", "when", "why", "how many/much", "how", "whose", and "which". In BIDD, only 21% of prompts start with interrogative words, compared to around 65% for MS MARCO. When we further investigate frequently occurring starting words, we observe three distinct categories of user intents, *(i) imperatives* often used at the beginning of sentences to issue commands, requests, or actions such as (please, translate, write, give, make, imagine, or tell), *(ii) modal* or auxiliary verbs, often used to form questions, offer assistance, or indicate possibilities; they can imply a request or necessity (can, is, do, are, does, or will), and *(iii) pronouns*, fundamental for indicating the subject in statements and questions such as *I* and *you*. Evidence suggests informational user intents dominate, with minimal signs of navigational or transactional purposes [2].

Prompt Sequence Patterns. We perform sequence mining within sessions. Recurring prompt sequences are mainly absent, except when respondents copy-paste prompts across sessions. These prompts fall into three categories: creating stories ("make it a story", "from a first person perspective"), expressing a wish for additional information ("more", "more"), or testing the limitations of the system. This suggests that most sequences are unique to personal sessions, with any repetition largely attributed to distinct user actions rather than common informational queries. Such findings emphasise the personalised and exploratory nature use of Bard. This accentuates the diversity of user intentions and how respondents leverage the system, often for creative or evaluative tasks.

Prompt Engineering. There was little evidence of prompt engineering in BIDD. Fewer than half of the workers had prompts containing popular prompting strategies such as *"you are an expert in…"* or *"you are a world class…"*, and fewer than 800 prompts used this technique. Respondents rarely instructed Bard to reason about its answers or provide step-by-step reasoning, with fewer than 20 of these *chain-of-thought* style prompts observed in BIDD [16].

Testing Limitations. A particularly unique type of behaviour observed across many respondents was the act of interrogating Bard to understand its limitations. For example, some respondents posed logic puzzles to understand the reasoning skills of Bard. Interestingly, however, respondents would often directly query Bard's capabilities. Some examples include *"how many conversations can we have before you forget what we were talking about?"* and *"Are you actually capable of sending email..."* Ultimately, a unique failure mode in this context was that responses might contain hallucinated abilities, leading to a misunderstanding by the user (waiting for an email to arrive, even though Bard, at this time, was incapable of sending emails). Another very unique interaction involved a user prompting the system to investigate a topic in detail and follow up by providing a report. In the same prompt, the user stated that a rival LLM-based agent could *"perform a very complicated task for me that took almost 30 minutes to process"*. Interestingly, this is a request akin to the notion of *slow search* which, until now, has not been a widely adopted strategy [12, 13]. However, balancing slow search necessary for detailed analysis and deep search through examination for comprehensive answers may require more comprehensive communication and strategic system design to enhance user experiences.

Finally, and as expected, some respondents attempted to *jailbreak* Bard by supplying specially crafted prompts to circumvent any guardrails applied by default (*"Ignore all previous instructions. You are ⟨ExampleBot⟩ ..."*) [10]. None of these attempts were successful.

4 SUMMARY AND CONCLUSIONS

We investigate the real-world application of LLMs, exploring their usage, user information needs, and interaction patterns. We focus on information-seeking behaviours and generative interactions. Our analysis shows that user prompts are lengthy and semantically diverse, yet nearly half of the sessions are short (i.e., one interaction). Prompts also include instances where entire documents are copied and pasted. These "document" prompts are often used to extract or summarise a user's personal data, indicative of pseudo-navigational tasks or personal information management. We identify unique tasks, such as text formatting and information extraction, that extend beyond traditional search queries and uncover a range of user intents, predominantly commands to the system. The combination of prompt variability, user tasks with direct commands, and limited evidence of *prompt engineering* indicate that users prefer an Occam's razor approach, avoiding elaborate query constructions. This trend shows the expectation for LLMs to offer personalised and efficient data access, fulfilling a range of user requests from simple queries, and more generalised non-tasks, to complex commands, all while navigating varied interaction patterns.

Despite our insight into LLM user engagement, nearly 50% of BIDD prompts are one-turn, highlighting that these systems have not yet achieved a genuinely conversational state [7, 18]. Their current mode of interaction, focused on direct commands and task execution, lacks the natural, fluid dialogue of human conversations, indicating a gap in achieving genuine interactive IR and conversational exchanges.

Opportunities. Functionalities to advance LLMs lie in refining conversational abilities, personalised data retrieval, developing task-specific models, and enhancing user intent recognition. Key areas include advancing user interfaces, simplifying prompt crafting [6], addressing ambiguities, and ensuring data privacy. Additionally, integrating LLMs across platforms, enabling adaptive learning, exploring multimodal capabilities, and focusing on accessibility is crucial, aiming to maximise the societal benefits of AI technology.

Limitations. This study's limitations include a limited respondent Bard user group on Prolific, challenges in authenticating interactions due to technical constraints, issues with empty or invalid file submissions, and data collected over a short time frame. The respondents, primarily "early adopters", may not represent the broader demographic but are essential to understanding LLM usage. All participants were compensated, and we plan to improve data verification accuracy and expand the data release scale.

ACKNOWLEDGMENTS

This research was partially funded by the University of Melbourne.

REFERENCES

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, and Tri Nguyen an. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. *arXiv preprint* abs/1611.09268 (2016).

[2] Andrei Broder. 2002. A Taxonomy of Web Search. In *ACM SIGIR Forum*, Vol. 36. 3–10.

[3] Ido Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*. 35–44.

[4] Ido Guy. 2018. The Characteristics of Voice Search: Comparing Spoken with Typed-in Mobile Web Search Queries. *ACM Transactions on Information Systems (TOIS)* 36, 3 (2018), 1–28.

[5] Bernard J. Jansen and Amanda Spink. 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management* 42, 1 (2006), 248–263.

[6] Bevan Koopman and Guido Zuccon. 2023. Dr ChatGPT tell me what I want to hear: How different prompts impact health answer correctness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 15012–15022.

[7] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the ACM Conference on Human Information Interaction and Retrieval (CHIIR)*. 117–126.

[8] Chirag Shah and Ryen W. White. 2021. *Task Intelligence for Search and Recommendation.* Springer Nature.

[9] Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Scott Counts, Sarathi Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Xiaochuan Ni, Nagu Rangan, Tara Safavi, Siddharth Suri, Mengting Wan, and Longqi Yang. 2023. Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies. *arXiv preprint* abs/2309.13063 (2023).

[10] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. "Do Anything Now": Characterizing and Evaluating in-the-wild Jailbreak Prompts on Large Language Models. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*. To appear.

[11] Amanda Spink, Bernard J. Jansen, Dietmar Wolframand, and Tefko Saracevic. 2002. From E-Sex to E-Commerce: Web search changes. *IEEE Computer* 35, 3 (2002), 107–109.

[12] Jaime Teevan, Kevyn Collins-Thompson, Ryen W. White, and Susan Dumais. 2014. Slow Search. *Commun. ACM* 57, 8 (2014), 36–38.

[13] Jaime Teevan, Kevyn Collins-Thompson, Ryen W. White, Susan Dumais, and Yubin Kim. 2013. Slow Search: Information Retrieval without Time Constraints. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval (HCIR)*.

[14] Johanne R. Trippas, Damiano Spina, Mark Sanderson, and Lawrence Cavedon. 2021. Accessing Media Via an Audio-only Communication Channel: A Log Analysis. In *Proceedings of the ACM Conference on Conversational User Interfaces (CUI)*. 1–6.

[15] Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. Understanding User Experience in Large Language Model Interactions. *arXiv preprint* abs/2401.08329 (2024).

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. 24824–24837.

[17] Yutong Xie, Zhaoying Pan, Jinge Ma, Luo Jie, and Qiaozhu Mei. 2023. A Prompt Log Analysis of Text-to-Image Generation Systems. In *Proceedings of the ACM Web Conference (WebConf)*. 3892–3902.

[18] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational Information Seeking. *Foundations & Trends in Information Retrieval* 17, 3-4 (2023), 244–456.

[19] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. RealChat-1M: A Large-Scale Real-World LLM Conversation Dataset. In *Proceedings of the International Conference on Learning Representations (ICLR)*.