Task Supportive and Personalized Human-Large Language Model Interaction: A User Study

Ben Wang
benw@ou.edu
The University of Oklahoma
Norman, Oklahoma, USA

Jiqun Liu
jiqunliu@ou.edu
The University of Oklahoma
Norman, Oklahoma, USA

Jamshed Karimnazarov
jamshed.k@ou.edu
The University of Oklahoma
Norman, Oklahoma, USA

Nicolas Thompson
nicolas.f.thompson-1@ou.edu
The University of Oklahoma
Norman, Oklahoma, USA

ABSTRACT

Large language model (LLM) applications, such as ChatGPT, are a powerful tool for online information-seeking (IS) and problem-solving tasks. However, users still face challenges initializing and re-fining prompts, and their cognitive barriers and biased perceptions further impede task completion. These issues reflect broader challenges identified within the fields of IS and interactive information retrieval (IIR). To address these, our approach integrates task context and user perceptions into human-ChatGPT interactions through prompt engineering. We developed a ChatGPT-like platform integrated with supportive functions, including perception articulation, prompt suggestion, and conversation explanation. Our findings of a user study demonstrate that the supportive functions help users manage expectations, reduce cognitive loads, better refine prompts, and increase user engagement. This research enhances our comprehension of designing proactive and user-centric systems with LLMs. It offers insights into evaluating human-LLM interactions and emphasizes potential challenges for under served users.

CCS CONCEPTS

• Information systems → Retrieval tasks and goals; Personalization; • Human-centered computing → Natural language interfaces.

KEYWORDS

Human-LLM Interaction, ChatGPT, Prompt Engineering, Information Seeking, Proactive System

ACM Reference Format:
Ben Wang, Jiqun Liu, Jamshed Karimnazarov, and Nicolas Thompson. 2024. Task Supportive and Personalized Human-Large Language Model Interaction: A User Study. In Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10– 14, 2024, Shefield, United Kingdom. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3627508.3638344

CHIIR '24, March 10–14, 2024, Shefield, United Kingdom

ACM ISBN 979-8-4007-0434-5/24/03
https://doi.org/10.1145/3627508.3638344

1 INTRODUCTION

The release of ChatGPT has sparked considerable interest in the interaction between humans and AI. This interest has led to a rising number of individuals employing large language models (LLMs) for various purposes such as task assistance, entertainment, education, and even as an alternative to traditional search engines [3, 5]. Despite the prevalence of ChatGPT, users still face challenges formulating prompts, and cognitive barriers and biased perceptions further impede task completion [26, 34]. These issues reflect broader challenges identified within the fields of information seeking (IS) and interactive information retrieval (IIR), particularly concerning task context and user perceptions, such as task topic and type, user intent, topic familiarity, and task expectations [2, 14, 16, 23, 30].

Previous IIR studies have underscored the complexity of integrating these fluctuating user perceptions into a predominantly static search system. Fortunately, the evolution of LLMs marks a transformative phase in information access (IA) paradigms, introducing a promising avenue for incorporating more nuanced interaction data through conversational context between the user and generative IA (GIA) system [24]. Therefore, it becomes crucial to explore methodologies for embedding task context and user perceptions into ChatGPT interactions, subsequently evaluating their impact on user experience and task completion, which are essential criteria for evaluation IIR and Human-AI Interaction [1, 17, 19, 26].

To achieve this, we have developed a task platform that emulates the oficial ChatGPT interface, incorporating the GPT-3.5-turbo model. Aiming to support users with the challenges mentioned above, we designed and implemented three supportive functions: 1. Perception Articulation: allows users to clarify their perceptions, including topic familiarity, and expected task complexity. This perception articulation will be then input to ChatGPT through prompt engineering to enrich the context information; 2. Prompt Suggestions: generates prompt revisions and follow-up questions, aiding users who struggle with prompt formulation; 3. Conversation Explanation: generates explanations for the ongoing conversation (i.e., the user's prompt and ChatGPT's response pair) for users to better comprehend ChatGPT's interpretation of the conversation.

To validate our approach, we conducted a naturalistic user study, involving 16 participant of college students and crowdsourced workers with self-defined tasks. These tasks spanned various lengths and cover diverse topics including creative writing, professional development, and specific programming questions.

Our analysis underscores the effectiveness of the supportive functions, illuminating their role in facilitating user experience and task completion. The findings reveal that these functions proved instrumental in managing user expectations, reducing cognitive load, guiding prompt refinement, and increasing user engagement. This research further enhances our understanding of designing proactive and user-centric systems with LLMs, offering insights into evaluating human-LLM interactions from both the system and user ends, and underscoring potential challenges for under served users in this new era of AI.

2 RELATED WORK

2.1 Capability of ChatGPT

LLMs have emerged as a groundbreaking development in the realm of artificial intelligence, leveraging sophisticated architectures trained on extensive data to understand and emulate human-like text generation [20]. One such application is ChatGPT, which has seen the most significant growth in its userbase. One key access to the versatility of LLMs is prompt engineering, a method that uses specific information and instruction in the input to optimize ChatGPT's output content and format [18]. Previous studies have examined using ChatGPT and other LLMs in educational settings where they can personalize content delivery and foster enhanced learning experiences [4, 8, 29, 35]. In addition, ChatGPT has demonstrated potential in providing emotional support, by playing therapeutic roles based on user sentiment and need [29].

2.2 LLMs as Information Access Systems

Incorporating LLMs in information access systems introduces transformative prospects from GIA, particularly through multi-turn interactions that resemble traditional IIR processes [24, 26]. However, harnessing LLMs in information access systems presents pronounced challenges. Users encounter dificulties in query formulation or interpreting search results, often due to cognitive barriers, which are common when initializing and refining prompts during interactions with LLMs [23, 34]. These barriers often stem from task context and user perceptions, such as lack of prior knowledge, low familiarity level with the topic, high complexity, and inappropriate expectations, leading to potential misconceptions about how LLMs interpret and respond [6, 7, 19, 23, 30, 31].

2.3 Evaluating Human-LLM-Interaction

In recent literature, the evaluation of human-LLM interaction has garnered significant attention, especially as these models become increasingly used in human tasks. A comprehensive approach to this evaluation has been proposed, encompassing aspects such as task performance, user experience, and general "Human-AI eXperience" (HAX) [1, 10]. With similar interaction process and challenges, there are also notable parallels between the evaluation of human-LLM interaction and the assessment of IIR. Both areas highlight the significance of user experience and perception. Another critical facet enhancing user trust and comprehension is explainability in AI, which merits more profound exploration [9, 33].

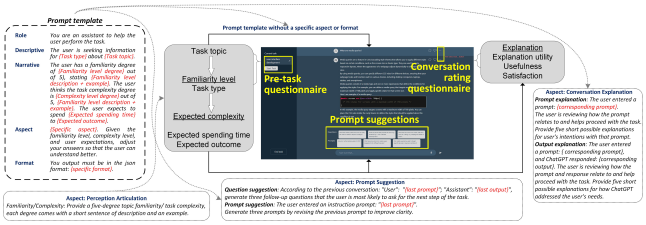

Figure 1: User study platform and prompt templates for supportive functions.

3 RESEARCH QUESTIONS

In respect to the challenges and opportunities in human-LLM interaction and the roles of task context and user perceptions in IS and IIR research, our study aims to investigate the research question RQ: How can we provide support with task context and user perception information to mitigate user challenges in tasks when interacting with ChatGPT?

To answer this question, we explore our methodology for collecting and integrating features related to task context and user perception, importing these features into the system through prompt engineering, and developing supportive functions to enhance user assistance in the task.

4 METHODOLOGY

4.1 System design and supportive functions

We developed an interface similar to the oficial ChatGPT, with GPT-3.5-turbo as the model, and integrated questionnaires and supportive functions. The interface and the prompt templates for the supprortive functions are shown in Figure 1. To interact with the system, users need to enter the pre-task questionnaire highlighted in upper left of the interface. The pre-task questionnaire collects features of task context and user perceptions, including task topic and type, familiarity level, and expectations (e.g., expected task complexity, spending time, and outcome). The detailed descriptions of these features are in Table 1. Within the pre-task questionnaire, perceptions articulation is implemented as the generative features: familiarity level and expected complexity. Unlike traditional surveys that use Likert scales to measure these two features, our approach utilizes ChatGPT to generate five degrees of familiarity or task complexity with descriptions and examples. This function aims to assist participants in better comprehending and selecting the option that most closely aligns with their perceptions. The chosen degree, along with its description and example, are then formatted in the prompt template to enrich the context of the main prompt template, which is demonstrated in the left dashed box in Figure 1. We developed this main prompt template with several components, including role, description, narrative, aspect, and format according to a prior study for designing effective prompts [28]. This template aims to provide ChatGPT with comprehensive background information about task context and user perceptions.

After this pre-task questionnaire, we implement the second supportive function, prompt suggestions, by using the main prompt template involved ongoing conversations. The suggestions are displayed in separate tabs at the bottom of the interface. Furthermore,

Table 1: Features in the pre-task questionnaire and the conversation rating questionnaire.

Feature	Description
Pre-task questionnaire	
Task topic	User input text of task topic
Task type	General task type options, such as Learning a new topic, generating text, casual conversation, replacing search engine, etc.
Familiarity level	Five-degree choices with descriptions and examples generated by a prompt involving {task topic}
Expected complexity	Five-degree choices with descriptions and examples generated by a prompt involving {Task topic} and {Familiarity level}
Expected spending time	Three options: less than 30 minutes, 1 to 2 hours, more than 3 hours
Expected outcome	Four options: get a broad idea of the task, partially complete the task, fully complete the task, other(text input)
Conversation rating questionnaire	
Explanation	Five potential explanations for corresponding {user prompt} and {ChatGPT response}
Explanation utility	Five-degree options with descriptions from "very poor" to "excellent"
Conversation usefulness	Five-degree options from not useful to extremely useful.
Conversation satisfaction	Five-degree options from very unsatisfied to very satisfied.

for each conversation, we implement the conversation explanation function in the rating questionnaire. This function generates five explanation options, allowing users to select the one that best aligns with their intent. We also include an explanation utility question, which allows users to rate the utility of the chosen explanation on a five-point scale. The purpose of the conversation explanation function is to present ChatGPT's interpretation of each prompt or response for users and investigate the potential of these explanations in enhancing user engagement and experience.

4.2 Participant recruitment

We targeted two distinct user groups for our research: college students at a research university and crowdsourced workers from Amazon mTurk. The recruitment process contains two steps: Step 1: participants were asked to complete a registration survey. This survey collected background information, including demographics and prior experience with ChatGPT. Step 2: We then inquired participants if they wished to proceed to the remote user study. Those who opted in then reported the tasks they planned to perform with ChatGPT. We required participants to report task plans with three anticipated task lengths: short (less than 30 minutes), medium (1 to 2 hours), and long (3 hours or more). According to the naturalistic study setting, participants were allowed to edit the task plan when they had new task ideas and complete planned tasks in five days. Before their own tasks, they would perform a warm-up task to get familiar with the platform interface and the study process. After the user study, participants could opt for an interview where we sought their feedback and insights on their experiences. In these interviews, we specifically explored their views on the task experience using our platform, the effectiveness of the supportive functions, and any suggestions or opinions they might have. Compensation for participants includes $5 for the step 1 registration survey and $50 for the step 2 user study. This compensation exceeds the minimum wage threshold, and our research has received approval from the Institutional Review Board (IRB).

Table 2: Average results of the tasks and user experience.

Participant group	College student			Crowd worker		
Expected length	Short	Median	Long	Short	Median	Long
Task count	7	6	3	3	7	3
# Prompt	5.9	5.2	22.7	12.5	12.5	25.3
# Suggestion	1.2	0.8	1	12.5	12	23
Duration (min)*	223.5	445.2	2488.6	558.3	31.5	32.9
Usefulness	3.8	4.2	3.2	4.3	4.2	4
Explanation utility	4	4.1	3.2	4.2	3.8	2.8
Satisfaction	4	4.2	3.3	4.3	4.5	4.2

*This duration reflects the gap between the start and end of the task, not the exact amount of time spent on the task.

The decision to choose two distinct participant groups aimed at broadening user diversity. While past studies focused on college students as early adopters of ChatGPT, they still highlighted the need for a more heterogeneous user group. Consequently, our participant pool includes college students from diverse fields such as Computer Science, Library and Information Science, and Public Health. Additionally, we incorporated crowd workers to ensure an even broader user spectrum. However, we set a qualification with an age range of 18-25 for crowd workers to facilitate a comparative analysis between the two groups.

4.3 Analysis

For this small-scale user study, we utilized a descriptive analysis by presenting the tasks users performed on our platform and explaining how the platform influenced user experience and assisted them in task completion. In addition, we delved into the interview data as case studies to illuminate users' experiences and insights.

5 RESULTS

5.1 Participants and tasks

As a result, 16 participants enrolled in the user study, comprising 8 college students and 8 crowd workers. The college students came from various academic backgrounds, including computer science, library/information science, and public health, ranging from sophomores to graduates. The crowd workers specialized in fields such as information technology and business, and were either pursuing or had already obtained their bachelor's degrees. Out of the 16 participants, six completed the tasks according to their task plans and participated in the interview (3 college students and 3 crowd workers), while the remainder finished at least the warm-up task.

Table 2 presents the average results of the tasks. Excluding the warm-up task, there were 29 tasks in total, comprising 10 short tasks, 13 medium tasks, and 6 long tasks. There were notable differences between the college student group and the crowd worker group, especially concerning the numbers of prompts and used prompt suggestions, and task duration. College students submitted approximately 5 to 6 prompts in short or medium tasks, though the duration for medium tasks was about double that of short tasks. They submitted over 20 prompts in the long task, completing the task in almost two days. In all task lengths, they adopted about one prompt from the suggestions in average. Conversely, crowd workers spent more time and prompts on short tasks than college students did, but they spent less time on medium and long tasks. This discrepancy could stem from their reliance on prompt suggestions, as nearly all their prompts were derived from these suggestions. Regarding the conversation ratings, college students

Table 3: Summary and examples of task topics and types performed by users.

Participant group	College student			Crowd worker		
Expected length	Short	Median	Long	Short	Median	Long
Topic	Learning a language, Grammar Checking, Basic programming	Specific programming problems	Specific programming problems	(writing techniques)	JK (Rowling), Company market	Learning (online learning platform), Developing (learning programming languages)
Type	Learning a new topic, Writing or generating text	Learning a new topic, Developing or programming	Developing or programming	Casual conversation (Learning anew topic)	Casual conversation (Learning a new topic)	Casual conversation, Learning a new topic

Task topics and types in "()" provide clarifications for ambiguous topics or types that users entered in the pre-task questionnaire, as further inferred from actual conversations.

had a positive experience (high usefulness, explanation utility, and satisfaction) in short and medium tasks but a moderate experience in long tasks. Crowd workers generally had a positive experience, except for the moderate explanation utility in long tasks.

To gain deeper insight into the participants' experiences, table 3 provides a summary and examples of tasks topics and types, grouped by users and expected task length. College students (especially computer science students) engaged in tasks that included learning new topics (in short and medium tasks) and solving specific programming problems (in medium and long tasks). In contrast, crowd workers did not specify clear task topics in the pre-task questionnaire. They used vague terms like "JK," "learning," and "developing," primarily intending to engage in casual conversations with ChatGPT. Consequently, based on their input and heavy reliance on the prompt suggestion function, those suggestions led the conversations towards topics such as "J.K. Rowling", "online learning platforms", and "learning programming languages".

5.2 Interview case study on task experience

We further present insights from the interview as case studies, examining the impact of supportive functions on user experience and task completion. Table 3 outlines participants' backgrounds, prior experiences with ChatGPT, task experiences during this study, and insights. We interviewed three computer science college students, P1, P2, and P3, all of whom showed considerable enthusiasm and engagement in both the tasks and subsequent interviews. Additionally, we interviewed crowd workers P4, P5, and P6. We delve into detailed feedback from P1, P2, and P3 and summarize the concerns raised by P4, P5, and P6.

P1, a self-identified expert of ChatGPT, considering ChatGPT as useful but not compatible with human experts. In this study, P1 recognized the value of perception articulation in "guiding expectations". In addition, the conversation explanations helped P1 "balance expectations" and satisfaction. For instance, P1 previously experienced ChatGPT's shortcomings in understand the specific rules of haikus, and they (a gender-neutral alternative to he/she) had a low expectation and started with some simple rules of haikus in this study. Surprisingly, P1 found the generated haikus followed the correct rules. However, inconsistencies arose when P1 increased the rule complexity. After reviewing the conversation explanations, P1 adjusted their expectations, acknowledging ChatGPT's limitations but remaining satisfied. P1 also noticed that explanations were more detailed for tasks with low familiarity but more concise and "straight to the point" for more familiar topics. However, P1 found the prompt suggestion function sometimes misaligned with their intended task direction, particularly in specific programming problems. Nonetheless, P1 appreciated the guidance from suggested prompts when exploring unfamiliar content, especially in the internship related question. They also highlighted the study platform's effectiveness in maintaining focus on specific tasks, thereby enhancing engagement and eficiency.

P2 previously found ChatGPT unexpectedly eficient in suggesting coding solutions. However, they did not anticipate fully completing tasks from ChatGPT's responses but expected to "get a broad idea" to guide their own solution development. In this study, P2 recognized how the platform interpreted task context and user perceptions in the pre-task questionnaire. For example, when encountering challenges that ChatGPT's response exceeded P2's knowledge, they realized overestimated familiarity of a topic and reassessed it. The conversation explanation function allowed P2 to understand where any miscommunication began and to refine prompts. For example, in the complex C++ coding task, P2 gained confidence and refined prompts through iterative interactions and could use ChatGPT to "build all remembered functions in one prompt", attributing this improvement to learning from the explanations. However, P2 became dissatisfied when the code generated by ChatGPT was incompatible with an updated game development package. They then realized that ChatGPT had limited knowledge on this development package and discontinued the conversation for alternative resources. Regarding the prompt suggestions, P2 found them useful in initializing short tasks, but overly broad and generalized in the longer or more complex tasks.

P3 self-identified as proficient with ChatGPT, but they experienced miscommunication with ChatGPT's interpretations. In this study, P3 provided unique insights when revising their resume. P3 observed redundancy and paraphrasing issues among the explanation options. For the prompt suggestions, P3 felt that those suggestions could sometimes distract from their initial intents. Nevertheless, P3 acknowledged the suggestions' potential to unveil new perspectives, such as unconsidered resume layouts. P3 also emphasized the effort required to modify ChatGPT's output to their preferred style and content.

Regarding the crowd workers (P4, P5, and P6) involved in IT-related fields, their previous interaction with ChatGPT was primarily work-related. In this study, as per the tasks summarized in Tables 1 and 2, they appeared hurried in their completion, heavily relying on prompt suggestions, possibly due to their focus on Human Intelligence Tasks (HITs). Their feedback during interviews remained nonspecific, centered more around task completion for compensation rather than the study's investigative purpose. It seems they misunderstood the task purpose of this study, or they were outliers of target users, which raised concerns discussed in the next section.

Table 4: Summary of user experience and insights from the interview.

Participant	P1	P2	P3	P4 P5 P6
Background	Computer science senior	Computer science sophomore	Computer science sophomore	Crowd workers in information technology fields
Previous experience	ChatGPT not compatible with human experts	ChatGPT with unexpected performance	Miscommunication between the user and ChatGPT	Using ChatGPT for text generation work
Task in this study	Generating Haikus, Internship specific question, Specific Coding problem	Learning a game dev tool, Using a game dev package, Specific coding problem	Resume revision	Casual conversation, learning a topic
Experience in this study	• Detailed pre-task perceptions for guiding expectations; • Prompt explanations for balancing expectations; • Prompt suggestions for curiosity or unfamiliarity; • Task-level conversation	• Overestimated familiarity recognition through retrospection; • Prompt suggestions for initializing the conversation; • Prompt suggestion not applicable for complex tasks; • Explanations for prompt refining; • Final prompt development or final solution guidance	• Paraphrasing issues in explanations; • Distracting issues in prompt suggestions; • Prompt suggestions for new and diverse information; • Efforts in refining desired output	• Prompt suggestion reliance; • Satisfaction
Insight	• Expectation management • Task reliability • Prompt refining willingness • Diverse prompt suggestions • Task-focused interface	• Dynamic perceptions • Expectation management • Task reliability • Explainability for prompt refining • Interactive prompt refining process	• Diverse explanation, • Diverse prompt suggestions • Interactive prompt refining process	• Misuse of prompt suggestions

6 DISCUSSION AND CONCLUSION

In closing discussion, we incorporate insights derived from our study, focusing on enhancing human-LLM interaction and identifying prevalent challenges during these interactions.

Insights and Implications for System Evaluation: The interview insights suggest distinct evaluation metrics at both the system and user ends. For the system, we can propose task reliability, which reflects the LLM's deep understanding and capability of task-specific knowledge beyond the text completion. This reliability could be evaluated through the LLM's ability to describe different levels of task familiarity and complexity for better recognizing the user's task states. Another evaluation aspect could be conversation explainability, reflecting the LLM's ability to interpret the conversation from a task-aware perspective. Enhancements of task reliability and explainability might include graph-based methods to represent the task in a knowledge structure with contextual information that enhances situational awareness [22, 32]. On the user end, metrics on task engagement could be measured by users' willingness to refine prompts and be influenced by their expectations [17, 34].

Proactive User Interface and System Design: Our study offers insights for proactive user interface and system design, which extends previous work on proactive IR [15]. Our approach involved utilizing questionnaires and specific interface tabs, yet the need for more seamless integration emerged. Following established AI interaction guidelines, a separate interface, such as a sidebar or secondary tab, could be deployed to facilitate supportive functions without interruptions on the ongoing conversation with ChatGPT. This proactive interface could also adopt a conversation-based mechanism for capturing user perceptions and provide suggestions, assisting the main ChatGPT interaction. This system holds the potential to leverage a fine-tuned LLM, for analyzing ongoing conversations and offering task-aware recommendations [11–13, 25]. This insight represents a step towards more intuitive, assistive AI, catering to diverse user needs without over-complicating the interaction.

Recognizing the Unique Position of Crowd Workers: This study revealed unexpected findings concerning crowd workers' interactions with ChatGPT. These anomalies may not solely be attributed to users' misunderstanding of the study's purpose but also possibly to the inherent challenges these users faced in formulating their own tasks. HITs for crowd workers are usually straightforward and repetitive tasks, like data labelling, even though they still require instructions and a gold standard to ensure accuracy [28]. Although they can produce text resembling task plans to meet the HIT criteria, this text does not necessarily mirror their real information needs [21, 27]. Formulating their "own" tasks or even identifying their information needs may be internally complex tasks. Additionally, the emphasis on task-centric LLM in this study might overshadow challenges with exploratory or open-ended user intentions. Therefore, it is crucial to comprehend the diverse information needs and usage contexts among more user categories, so that the proactive system should be intuitively informative and inclusive, rather than restrictively functional for certain user groups [24].

Limitations and Future Directions: Acknowledging our study's constraints in its limited scale, we advocate for expansive user studies with participants from diverse communities, populations, and backgrounds. Beginning for the prototype system in this study, future work could involve improving the interface and system design, conducting prompt template ablation studies, and exploring LLM fine-turning for developing task support affordances. Further discussions might explore task automation and copilot, with the challenge of balancing user engagement in information seeking and learning, while providing eficient, single-prompt task resolution.

In conclusion, our user study delved into supportive functions for ChatGPT, bolstering user experiences and task completion. The results indicated that these functions effectively guided users in managing expectations, reducing cognitive load, better refining prompts, and increasing user engagement. This investigation markedly advances our understanding of how to leverage LLMs for proactive IS systems. Moreover, it sheds light on the evaluation of human-LLM interactions and highlights the potential oversight of certain user groups' unique challenges in the era of Generative AI.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) Award IIS-2106152, a grant from the Seed Funding Program of the Data Institute for Societal Challenges, the University of Oklahoma, and a fund from Microsoft for Startups Founders Hub. Any opinions or findings here do not necessarily reflect those of the sponsors.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In Proceedings of the 2019 chi conference on human factors in computing systems. 1–13.

[2] Leif Azzopardi. 2021. Cognitive biases in search: a review and reflection of cognitive biases in Information Retrieval. In Proceedings of the 2021 conference on human information interaction and retrieval. 27–37.

[3] Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. 2023. ChatGPT: Applications, opportunities, and threats. In 2023 Systems and Information Engineering Design Symposium (SIEDS). IEEE, 274–279.

[4] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI 7, 1 (2023), 52–62.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems 33 (2020), 1877–1901.

[6] Barbara J Ford. 1995. Information literacy as a barrier. IFLA journal 21, 2 (1995), 99–101.

[7] Ahmed Hassan, Ryen W White, Susan T Dumais, and Yi-Min Wang. 2014. Struggling or exploring? Disambiguating long search sessions. In Proceedings of the 7th ACM international conference on Web search and data mining. 53–62.

[8] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual differences 103 (2023), 102274.

[9] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. " Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–17.

[10] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. arXiv preprint arXiv:2212.09746 (2022).

[11] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. 1244–1247.

[12] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive Conversational Agents in the Post-ChatGPT World. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 3452–3455.

[13] Jiqun Liu, Matthew Mitsui, Nicholas J Belkin, and Chirag Shah. 2019. Task, information seeking intentions, and user behavior: Toward a multi-level understanding of Web search. In Proceedings of the 2019 conference on human information interaction and retrieval. 123–132.

[14] Jiqun Liu, Shawon Sarkar, and Chirag Shah. 2020. Identifying and predicting the states of complex search tasks. In Proceedings of the 2020 conference on human information interaction and retrieval. 193–202.

[15] Jiqun Liu and Chirag Shah. 2019. Proactive identification of query failure. Proceedings of the Association for Information Science and Technology 56, 1 (2019), 176–185.

[16] Jiqun Liu and Chirag Shah. 2022. Leveraging user interaction signals and task state information in adaptively optimizing usefulness-oriented search sessions. In Proceedings of the 22nd ACM/IEEE joint conference on digital libraries. 1–11.

[17] 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. 1533–1542.

[18] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. Comput. Surveys 55, 9 (2023), 1–35.

[19] Daan Odijk, Ryen W White, Ahmed Hassan Awadallah, and Susan T Dumais. 2015. Struggling and success in web search. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 1551–1560.

[20] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35 (2022), 27730–27744.

[21] Joel Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? Shifting demographics in Mechanical Turk. In CHI'10 extended abstracts on Human factors in computing systems. 2863–2872.

[22] Shawon Sarkar, Maryam Amirizaniani, and Chirag Shah. 2023. Representing Tasks with a Graph-Based Method for Supporting Users in Complex Search Tasks. In Proceedings of the 2023 Conference on Human Information Interaction and Retrieval. 378–382.

[23] Reijo Savolainen. 2015. Cognitive barriers to information seeking: A conceptual analysis. Journal of Information Science 41, 5 (2015), 613–623.

[24] CHIRAG SHAH and EMILY M BENDER. 2023. Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? (2023).

[25] Chirag Shah, Ryen White, Paul Thomas, Bhaskar Mitra, Shawon Sarkar, and Nicholas Belkin. 2023. Taking search to task. In Proceedings of the 2023 Conference on Human Information Interaction and Retrieval. 1–13.

[26] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The user experience of ChatGPT: Findings from a questionnaire study of early users. In Proceedings of the 5th International Conference on Conversational User Interfaces. 1–10.

[27] Manuel Steiner, Damiano Spina, Falk Scholer, and Lawrence Cavedon. 2021. Crowdsourcing Backstories for Complex Task-Based Search. In Proceedings of the 25th Australasian Document Computing Symposium. 1–6.

[28] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large language models can accurately predict searcher preferences. arXiv preprint arXiv:2309.10621 (2023).

[29] Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments 10, 1 (2023), 15.

[30] Ben Wang and Jiqun Liu. 2022. Investigating the Relationship between In-Situ User Expectations and Web Search Behavior. Proceedings of the Association for Information Science and Technology 59, 1 (2022), 827–829.

[31] Ben Wang and Jiqun Liu. 2023. Investigating the role of in-situ user expectations in Web search. Information Processing & Management 60, 3 (2023), 103300.

[32] Xiang Wang, Dingxian Wang, Canran Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. 2019. Explainable reasoning over knowledge graphs for recommendation. In Proceedings of the AAAI conference on artificial intelligence, Vol. 33. 5329–5336.

[33] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8. Springer, 563–574.

[34] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–21.

[35] Xiaoming Zhai. 2022. ChatGPT user experience: Implications for education. Available at SSRN 4312418 (2022).