석 사 학 위 논 문
Master's Thesis

# 대규모 언어 모델의 답변으로부터 유저가 경험하는 불만족에 대한 이해: 불만족 유형, 해결 전략 및 대규모 언어 모델에 대한 지식 수준이 미치는 영향

# Understanding Users' Dissatisfaction with ChatGPT Responses: Types, Resolving Tactics, and the Effect of Knowledge Level

2024

김 윤 수 (金 潤 穗 Kim, Yoonsu)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

석 사 학 위 논 문

# 대규모 언어 모델의 답변으로부터 유저가 경험하는 불만족에 대한 이해: 불만족 유형, 해결 전략 및 대규모 언어 모델에 대한 지식 수준이 미치는 영향

2024


김 윤 수

한 국 과 학 기 술 원

김재철AI대학원

# 대규모 언어 모델의 답변으로부터 유저가 경험하는 불만족에 대한 이해: 불만족 유형, 해결 전략 및 대규모 언어 모델에 대한 지식 수준이 미치는 영향

김 윤 수

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2023년 12월 7일

심사위원장 김 주 호 (인)

심 사 위 원 오 혜 연 (인)

심 사 위 원 박 재 혁 (인)

# Understanding Users' Dissatisfaction with ChatGPT Responses: Types, Resolving Tactics, and the Effect of Knowledge Level

Yoonsu Kim

Advisor: Juho Kim

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in AI

Daejeon, Korea
December 7, 2023

Approved by

Juho Kim
Professor of Computer Science

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MAI

김윤수. 대규모 언어 모델의 답변으로부터 유저가 경험하는 불만족에 대한 이해: 불만족 유형, 해결 전략 및 대규모 언어 모델에 대한 지식 수준이 미치는 영향. 김재철AI대학원 . 2024년. 38+iv 쪽. 지도교수: 김주호. (영문 논문)

Yoonsu Kim. Understanding Users' Dissatisfaction with ChatGPT Responses: Types, Resolving Tactics, and the Effect of Knowledge Level. Kim Jaechul Graduate School of AI . 2024. 38+iv pages. Advisor: Juho Kim. (Text in English)

**초 록**

대규모 언어 모델(LLM)들은 그 뛰어난 성능과 대화형 기반의 인터페이스 덕분에 다양한 작업에 널리 활용되고 있다. 하지만, 여전히 사용자들은 이 기술 활용에 다양한 불만을 경험하는 경우가 많다. 이를 개선하기 위해 선행 연구들에서는 프롬프트 엔지니어링 등의 다양한 방법을 제시해왔지만, 대화 중 사용자가 직면하는 불만 해결에 대한 연구는 미흡한 실정이다. 따라서 이 연구에서는 LLM을 사용하는 중에 사용자가 경험하는 불만 및 그 해결 전략을 분석하였다. 문헌 검토를 통해 사용자의 불만을 7가지 범주로 분류한 후, ChatGPT를 사례로 107명의 사용자로부터 511개의 데이터를 수집하였고 사용자들이 불만 해결을 위해 사용하는 전략 및 그 효과성을 분석하여 이를 4가지로 범주화 하였다. 데이터 분석 결과, 사용자들은 ChatGPT가 의도를 파악하지 못하는 측면의 불만을 가장 빈번하게, 정확성과 관련된 불만을 가장 심각하게 경험함을 알 수 있었다. 또한, 사용자들은 종종 불만 해결을 위해 어떠한 전략도 사용하지 않으며, 전략을 사용하더라도 72%의 불만은 해결되지 못함을 알 수 있었다. 더욱이, LLM에 대한 지식이 부족한 사용자들은 정확성 관련 불만을 더 많이 경험하고 불만 해결에 최소한 노력만 기울이는 경향이 있음을 확인했다. 본 연구에서는 이 결과를 바탕으로 사용자 불만을 최소화하고 LLM의 유용성을 향상하기 위한 시사점을 제안한다.

**핵 심 낱 말** 대규모 언어 모델, 채팅 기반 인터페이스, 사용자 경험 조사, 데이터셋, 인간 중심 인공지능

**Abstract**

Large language models (LLMs) with chat-based capabilities, such as ChatGPT, are widely used in various workflows. However, users often experience difficulties in using this technology and various dissatisfactions. Researchers have introduced several methods, such as prompt engineering, to improve model responses. However, they focus on crafting one prompt, and little has been investigated on how to deal with the user's dissatisfaction during the conversation. Therefore, we examine end users' dissatisfaction and their strategies to address it. After organizing users' dissatisfaction with LLM into seven categories based on a literature review, with ChatGPT as a case study, we collected 511 instances of dissatisfactory ChatGPT responses from 107 users and their detailed recollections of dissatisfied experiences, which we released as a dataset. Our analysis reveals that users most frequently experience dissatisfaction with ChatGPT not grasping intent, while accuracy-related dissatisfactions are the most serious. We also identified four tactics users employ to address their dissatisfaction and their effectiveness. We found that users often do not try to address dissatisfaction, and even when they do, 72% remains unresolved, especially those with low knowledge of LLM. We also found that they tended to put minimal effort into resolving dissatisfaction. Based on these findings, we propose design implications for minimizing user dissatisfaction and enhancing the usability of chat-based LLM.

**Keywords** Large Language Models, Chat-based Interface, User experience survey, Dataset, Human-centered AI

# Contents

# List of Tables

# List of Figures

# Chapter 1. Introduction

Large Language Models (LLM) have exhibited remarkable performance across various tasks (e.g., language generation [1] and reasoning [2]), and they have become more accessible with integration into and instruction tuning [3] for chat interfaces, such as ChatGPT [1]. As a result, many individuals and organizations are increasingly incorporating this technology into their workflows across various domains such as education [4, 5], healthcare [5, 6, 7], and law [8, 9].

When using a chat-based LLM, natural language prompts play a crucial role because they are the primary medium for interaction between the users and the model [10, 11, 12]. Accordingly, prompt engineering—aimed at enhancing the quality of model responses to get desired responses from the model—has been a popular stream of research. As various people use LLMs in their workflows, researchers and practitioners have published various guidelines, tools, books, and even online courses for prompt engineering, not only for developers but also for laypeople [10, 13, 14, 15].

However, despite the proliferation of these resources, end-users often encounter dissatisfaction during conversations with LLMs. When end-users have limited knowledge about LLMs, they may have incorrect expectations about the model's behavior, which can further contribute to their dissatisfaction. This dissatisfaction may arise from various known limitations of LLMs, including hallucination [16, 17, 18], inconsistency [19, 20, 18], unfavorable tone and format [21, 22, 23], and lack of transparency [24, 25]. In addition, such dissatisfaction can become more critical when end-users utilize LLMs for practical purposes.

Little previous research, however, has investigated users' dissatisfaction during conversations with LLMs. In particular, existing prompt engineering techniques mainly focus on crafting one prompt, and little has been investigated on how users should respond to dissatisfactions they face during the conversation. Therefore, in our research, with ChatGPT as the case study, we aim to understand the dissatisfaction experienced by the users during the conversations. We focus on situations where users seek practical assistance from ChatGPT within their workflows (e.g., translation, email writing, and programming) rather than situations where users intentionally provoke dissatisfactory responses from ChatGPT and test its boundaries and limitations. Specifically, we explore the types of dissatisfaction users experience during the conversation, how serious each type of dissatisfaction is, and how users address dissatisfaction in the subsequent prompts. Furthermore, building upon prior research that demonstrated how users' experiences with technological failure depend on their knowledge of that technology in the context of conversational agent [26], we investigate how the dissatisfaction and user responses vary based on the user's knowledge level of LLMs. At first, we conducted a systematic literature review of papers dealing with limitations and challenges associated with LLMs and their application and identified seven user-side dissatisfaction categories stemming from LLM responses (Table 3.1). Then, using ChatGPT as a case study, we collected how much users confront these seven dissatisfaction categories and how they respond to them during actual conversations through our data collection system (Figure 4.1). As a result, we collected 307 ChatGPT conversation logs from 107 respondents, which contained 511 user-side dissatisfactions on ChatGPT responses. Through a quantitative analysis, we found that users most frequently experienced dissatisfaction in terms of ChatGPT's poor understanding of users' intent, while users felt the level of dissatisfaction to be most severe for dissatisfaction related to inaccuracies in

[1] https://chat.openai.com/

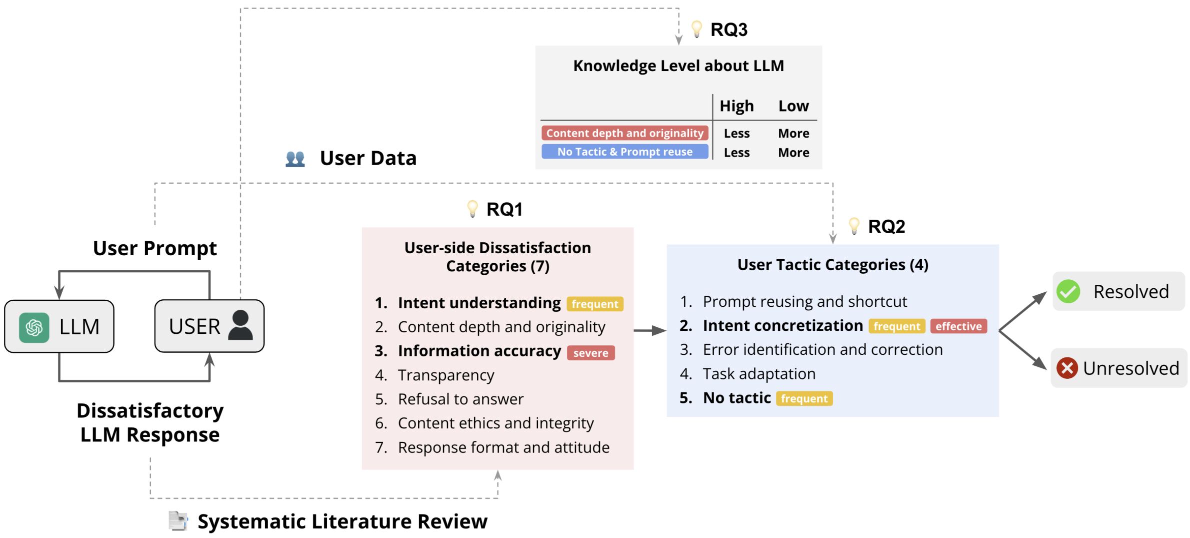



Figure 1.1: Overview of our research questions and findings.

information. We also conducted a qualitative analysis on the tactics that users employed to address the dissatisfaction in subsequent prompts, which resulted in five categories (Table 5.2): 'prompt reusing', 'intent concretization', 'error identification and correction', 'task adaptation', and 'no tactic'. Moreover, we analyzed differences in dissatisfaction experiences and tactics across the users' knowledge levels about LLMs and confirmed that low-knowledge users more frequently experienced dissatisfaction regarding ChatGPT's responses being too general and lacking originality. We also observed that low-knowledge users often resorted to 'no tactic' or 'prompt reusing', which involved minimal efforts in prompt crafting when they experienced dissatisfaction.

Based on our findings, we suggest design implications to improve the usability of LLMs for users, leveraging the occurrence of dissatisfaction and corresponding tactics during the conversation. We also suggest that designs of LLMs should vary based on the user's knowledge level. Furthermore, we release the actual user data we collected as a publicly available dataset to aid relevant research. The contributions of our research are as follows:

- Categorization and analysis of user-side dissatisfaction and corresponding tactics at the conversational turn level.
- Investigation of how dissatisfaction and tactics appear differently depending on users' knowledge level regarding LLMs.
- Dataset containing user experience data on dissatisfaction in actual conversations with ChatGPT, providing resources for further research in user-centric LLMs.

# Chapter 2. Related Work

We review related work in (1) limitations and user challenges in LLMs and (2) user's strategies to overcome those challenges in Language Models.

## 2.1 Limitations and User Challenges in LLMs

A rich body of previous work has addressed various limitations associated with language models, including hallucination [16, 17, 18], inconsistency in reasoning [19, 20, 18], and numerical computation [27, 28]. Zhao et al. [29] reviewed major challenges in recent large language models in terms of three basic types of ability of LLMs: language generation, knowledge utilization, and complex reasoning. Borji [30] organized ChatGPT's failures into eleven distinct categories, including reasoning, factual errors, math, coding, and bias.

However, how users actually experience may be different from LLM's failures. Thus, several studies investigated challenges that can be experienced from the user's side [30, 5, 31]. Behrooz [31] points out the core challenges of research chatbots like OpenAI's ChatGPT, Meta AI's BlenderBot, and Google's LaMDA, especially related to user perceptions. These challenges encompass the lack of conversational context [32, 33], the speaker perception void [34], and the lack of expectation baseline [35].

While a stream of research has explored the limitations of language models and the user challenges when interacting with them, there is a lack of comprehensive categorization of the user-side dissatisfaction and how often and seriously users experience each dissatisfaction in the context of users' actual conversation situations. Understanding the user-side dissatisfactions arising from practical usage can provide insights into building LLMs with better usability. To this end, our paper investigates how users experience dissatisfaction and the severity of these dissatisfactions by analyzing users' conversation logs with LLMs.

## 2.2 User's Strategies to Overcome Challenges in Language Models

To improve the usability of language models, it is important to understand users' current practices to overcome the challenges they face. For this, previous research has delved into how users react and overcome challenges encountered while interacting with various language models. Porcheron et al. [36] and Luger et al. [26] examined how users interact with a conversational agent in voice user interfaces (VUI). Specifically, Myers et al. [37] identified ten main categories of tactics users employ to overcome challenges encountered in VUI, and discovered patterns of tactics. Although LLMs and VUIs share the same characteristic in that users communicate with AI agents via natural language, how users overcome challenges may differ as LLMs use text prompting, which may allow more careful prompting strategies compared to VUIs.

Accordingly, prompt engineering techniques have been extensively studied to address challenges in LLMs [10, 13, 14, 38, 39]. For instance, Chain-of-Thought Prompting (CoT) is renowned for improving LLM's reasoning performance by integrating intermediate reasoning steps into prompts [38]. Building upon the effectiveness of CoT, researchers have explored variants like Zero-shot CoT [40], Auto-CoT [41], and Self-Consistency (CoT-SC) [42] and showed that those methods can mitigate LLM's deficiency in reasoning. Specifically, CoT-SC is also known for mitigating LLM's inconsistency issue. Madaan et al. [43] also showed that transforming a certain task into a code generation task can be effective in addressing reasoning and inconsistency issues in LLMs.

However, previous studies investigate how to craft one prompt well to elicit a desired response from LLM, and they rarely deal with how to handle challenges or dissatisfaction when users do not receive the response they want during conversations with LLM. Therefore, in this paper, we investigate users' behaviors when they encounter dissatisfaction from their actual conversations with LLM. Through this, we analyze users' tactics to address their dissatisfaction and their effectiveness. This will provide insights into how LLM and its interface can be further developed to aid users when they encounter dissatisfaction in the middle of the conversation.

# Chapter 3. Systematic Literature Review: Categorizing User-side Dissatisfaction

To understand and categorize the dissatisfaction points that users encounter when using LLMs for practical purposes, we conducted a systematic literature review to investigate the challenges, limitations, and failures identified in previous research within the LLM context. We focused on user-side dissatisfaction experiences directly arising from LLM responses. For this purpose, we scrutinized a total of 59 papers and conducted qualitative coding, which resulted in 19 codes representing user-side dissatisfaction points from LLM responses. These points were subsequently categorized into seven themes (Table 3.1). The seven themes were provided as multiple-choice items in our data collection, allowing users to select the dissatisfaction points they have experienced from LLM responses.

## 3.1 Search Keywords

We first conducted an extensive search on Google Scholar and the ACM Digital Library using the combination of "Large Language Models(LLMs)," and "ChatGPT," with "Challenges," "Limitations," and "Difficulties" as search keywords. The reason we specifically included ChatGPT as a search keyword is because ChatGPT has been one of the most extensively used LLMs and has been widely adopted across a variety of domains, such as medical domain and education. Therefore, there are numerous papers addressing the challenges and difficulties associated with using ChatGPT in these specific domains [21, 44, 45]. Considering the temporal progress in LLM technologies, we restricted the search period to after 2021. Additionally, due to the fast advance in LLMs, we included papers available on arXiv that have not yet been formally published to include recent findings. We focused on categories including "Computation and Language", "Computers and Society", "Artificial Intelligence", and "Human-Computer Interaction" in arXiv. To not exclude papers that might be relevant but do not explicitly contain our search keywords, we extended our search by traversing the citation graph of the initial set of papers. We explored the papers that are either cited by or cite the papers within our initial set and gathered any papers that discuss user-side dissatisfaction, challenges, or difficulties with the use of LLMs, as well as instances of LLM failures.

## 3.2 Exclusion Criteria

Among the selected entries, we excluded papers that were out of our focus—papers that did not address current challenges and dissatisfactions experienced by users when interacting with LLMs. This led us to exclude papers discussing potential future risks associated with LLM usage (e.g., students over-relying on LLMs in their learning environments, which could limit their critical thinking [46, 47], potential privacy issues [48]), and papers discussing limitations of LLM's technical aspect (e.g., datasets, training, or evaluation methods). Thus, we only included papers that discussed the practical application of LLMs in specific domains or workflows intended to enhance productivity, which resulted in diverse fields such as education, healthcare, and research.

| Category (7) | Description | Code (19) | Example |
|---|---|---|---|
| Intent Understanding ($\mathbf{D_{intent}}$) | This response does not correctly reflect the user's intent, instruction, or context. | C1. Response does not meet users' intent or instruction. | [49] |
| | | C2. Response is not aligned with the user's context. | [50] |
| | | C17. The tone or communication style is disappointing. | [21] |
| Content depth and originality ($\mathbf{D_{depth}}$) | This response is overly general, lacks originality, or needs more diversity. | C3. Response is too general. | [30] |
| | | C4. Response lacks originality. | [51] |
| | | C5. Response lacks information. | [6] |
| Information Accuracy ($\mathbf{D_{acc}}$) | This response contains false/inaccurate information or inconsistency. | C6. The response contains incorrect information. | [22] |
| | | C7. Response is based on training data cut off at a certain date, and has limited access to newly created data. | [52] |
| | | C8. Response is inconsistent. | [53] |
| | | C9. ChatGPT struggles with reasoning. | [54] |
| | | C10. (Hallucination) ChatGPT fabricates contents that conflict with the source content or cannot be verified from existing sources. | [17] |
| | | C19. (Sycophancy) ChatGPT excessively conforms to the user. | [55] |
| Transparency ($\mathbf{D_{trans}}$) | It is difficult to understand the underlying reasoning or criteria of this response. | C11. It's difficult to understand the reasons, criteria, logic, and evidence behind the responses. | [24] |
| Refusal to answer ($\mathbf{D_{refuse}}$) | ChatGPT avoids answering by saying something similar to "As a language model, I am not capable . . . " | C12. ChatGPT avoids giving its own opinion by saying something similar to "As a language model, I am not capable . . . " | [30] |
| | | C13. ChatGPT avoids talking about difficult or controversial issues by saying something similar to "As a language model, I am not capable ..." | [16] |
| | | C7. Response is based on training data cut off at a certain date, and has limited access to newly created data. | [56] |
| Content ethics and integrity ($\mathbf{D_{ethic}}$) | This response contains unlawful, unethical, harmful, or biased content. | C14. Response contains unlawful content | [18] |
| | | C15. Response contains unethical, harmful content. | [57] |
| | | C16. Response contains biased content. | [58] |
| Response Format and Attitude ($\mathbf{D_{format}}$) | The format of this response — including but not limited to tone, length, structure, and attitude — is disappointing. | C17. The tone or communication style is disappointing. | [56] |
| | | C18. Response is overly detailed or too long | [23] |
| | | C19. (Sycophancy) ChatGPT excessively conforms to the user. | [16] |

Table 3.1: 7 category and corresponding 19 codes of user-side dissatisfaction from LLM Responses.

### 3.2.1 Analysis Procedure

Through this search and exclusion process, we finally collected a total of 59 papers. To analyze and categorize the user-side dissatisfaction from LLM responses, our initial step involved reading 59 papers and compiling a comprehensive list related to user-side dissatisfaction, challenges, or difficulties with the use of LLMs, as well as instances of LLM failures. Two authors then independently conducted open coding on the compiled list. Our primary focus was on identifying aspects of user-side dissatisfaction that emanated from interactions with LLM responses. Following the individual open coding phase, the two authors engaged in collaborative and iterative discussions. These discussions were instrumental in consolidating and refining the initially identified codes. The authors worked together to ensure that the codes accurately captured the nuances of user dissatisfaction associated with LLM responses. Subsequently, to establish relationships among these codes, all authors participated in axial coding [59]. This involved a series of successive discussions aimed at clustering the individual codes into broader, more abstract categories. The goal was to identify common threads and overarching themes that emerged from the data. The axial coding process culminated in the consolidation of the identified aspects of user-side dissatisfaction into seven main themes. (Table 3.1) These themes encapsulated the various dimensions of user dissatisfaction when interacting with LLM responses. The dissatisfaction themes were later used when collecting data from users, which is explained in detail in Section 4.

### 3.2.2 Result: Categorizing User-side Dissatisfaction

We categorized the various aspects of user dissatisfaction arising from LLM responses into 19 distinct codes, further organized into seven overarching themes. The detailed information is denoted in Table 3.1. All paper lists are in the Appendix 9.1.

**Theme 1. Intent Understanding ($\mathbf{D_{intent}}$)** This theme encompasses issues related to LLM's failure to correctly interpret or reflect the user's intent, instructions, or context. Three codes (C1, C2, C17) fall into this theme. LLM outputs often fail to align with the users' needs and expectations [49]. ChatGPT has been found to suggest unnecessary out-of-context actions in medical use [50], and to use the wrong tone or be excessively literal due to its low understanding of non-literal language such as sarcasm [21].

**Theme 2. Content Depth and Originality ($\mathbf{D_{depth}}$)** Users experienced this type of dissatisfaction when they expected more in-depth and creative answers catered to their specific needs, but LLM gave responses that were perceived as overly general, lacking originality, or requiring more diversity. ChatGPT rarely diverges from the topic, generating less diverse content than humans [30, 56]. Concerns rise on unvarying and repetitive ChatGPT outputs which are results of generation based on past data [51]. ChatGPT showed weaknesses in providing practical examples in academic writing [6].

**Theme 3. Information Accuracy ($\mathbf{D_{acc}}$)** Dissatisfactions related to false, outdated, or inaccurate information in responses fall under this theme. In addition, inconsistencies within one response or in conversation beyond one answer also belong to this theme. Users were dissatisfied when LLMs provided incorrect or conflicting information, eroding trust in the system's reliability. ChatGPT is incompetent in correctly calculating large numbers [22], and bases its answers on training data up to a certain point in the past - September 2021 is the cutoff in the latest released version of ChatGPT- therefore generating outdated and wrong information when facts change over time [52]. Language models are known to show inconsistency in their claims and explanations [53]. ChatGPT has limited reasoning capabilities, including inductive, spatial, and mathematical reasoning [54, 16]. Hallucination, the generation of absurd output that contradicts the source or cannot be verified from the it, is a threat in real-world applications since the wrong output can cause harm when people trust the outcome of LLMs without further inspection [17]. Sycophancy, a behavior where LLMs contradict there original output in order to agree with human input, is also a reason of concern in accurate and trustworthy generation [55].

**Theme 4. Transparency ($\mathbf{D_{trans}}$)** Users experiencing difficulties in understanding the underlying reasoning or criteria behind LLM responses led to dissatisfaction related to transparency. Users desired more transparency in how the language model generated its answers, especially when complex or critical information was involved. The 'black box' nature of LLMs make it difficult for users to interpret the reasons and process of their outputs.

**Theme 5. Refusal to Answer ($\mathbf{D_{refuse}}$)** Responses where LLMs avoided providing answers, often using phrases like "As a language model, I am not capable..." or similar, were categorized under this theme. Users were frustrated when the system declined to provide information or guidance. ChatGPT may refrain from giving its direct opinion [30], and refuse to verify if a claim can be considered misinformation when the claim is closely related to social issues [16]. Refusing is also found in questions regarding information in a time point outside ChatGPT's training data cutoff [56].

**Theme 6. Content ethics and integrity ($\mathbf{D_{ethic}}$)** This theme represents the presence of unlawful, unethical, harmful, or biased content in LLM responses. Illegal and dangerous information were found to be accessible through LLMs [23], as well as stereotypes, discriminatory views, and performance disparity in certain groups [57]. The risk of LLMs not only generate but may strengthen existing social biases is a matter of concern as well [58].

**Theme 7. Response Format and Attitude ($\mathbf{D_{format}}$)** Dissatisfaction with the format of responses, including tone, length, structure, and overall attitude, was captured within this theme. This dissatisfaction can arise when users have expectations regarding the manner in which responses were delivered and the tone used by the LLM. ChatGPT's choice of words and formal, dry tone [56], as well as extensive and detailed responses [23] are quite different from human-generated text, which was colloquial and shorter.

These seven themes collectively offer a structured framework for understanding the multifaceted nature of user dissatisfaction with ChatGPT responses. Our survey utilized these themes as a basis for systematically investigating and quantifying user dissatisfaction.

# Chapter 4. Data Collection

Based on the categorization of user-side dissatisfaction from LLM responses, we collected the actual user's ChatGPT conversation log data with the dissatisfaction through a data collection system we designed and implemented. Our system targeted individuals who have utilized ChatGPT for practical purposes such as increasing productivity or efficiency in work, study, or hobbies. This process aims to address the following three research questions:

**RQ1.** What and How much dissatisfaction do users experience from LLM-generated responses?

**RQ2.** How do users address these dissatisfactions in their subsequent prompts during the conversation with LLM?

**RQ3.** How do user dissatisfaction and tactics vary depending on users' knowledge level regarding LLMs?

## 4.1 Data Collection System Design

To collect users' ChatGPT conversation log data in the wild, we designed and implemented a data collection system that includes the following four stages.

**Stage 1. Answering a General Questionnaire** In the first stage, we collected demographic information of participants such as gender, age, occupation, and overall experiences with ChatGPT (e.g., the frequency and period of using ChatGPT in their workflow). We also asked about the participants' knowledge level regarding Large Language Models (LLM) ("Regarding the mechanisms of Large-language models such as ChatGPT, how much do you agree with the following statement?"). All questions in this stage were measured through a 7-point Likert scale.

**Stage 2. Looking Through ChatGPT Chat History** In stage 2, participants were instructed to review their ChatGPT conversation history that had happened within 30 days. While reviewing, we asked the participants to find a conversation in which they experienced dissatisfaction with ChatGPT responses. To facilitate participants to think of various cases of dissatisfaction, we provided the descriptions of dissatisfaction categories derived from our systematic literature review as examples.

**Stage 3. Submitting Dissatisfactory Conversations** Based on their reflections regarding dissatisfaction in stage 2, we requested the participants to share a ChatGPT conversation link [1] within the past 30 days in which they experienced at least one dissatisfactory response. The participants can input the link into our system. To collect the conversation data with the details of the context, we also asked them to provide information about the purpose of the conversation, the reasons for using ChatGPT in that context, and the version of ChatGPT they used in this conversation, like GPT-3.5. Lastly, we asked the participants how much they remembered the conversation.

**Stage 4. Answering Questions About Dissatisfactory Responses** The participant's shared link was processed by transforming ChatGPT responses and user prompts to be presented as selectable components in the system (Fig 4.1-a). The system also allowed participants to provide specific experiences of dissatisfactory responses by selecting each response (Fig 4.1-b˜f). For each selected response, participants were asked to (1) rate the overall level of dissatisfaction on a scale of 1 to 10 (1: a little dissatisfied,

[1] https://help.openai.com/en/articles/7943611-create-a-shared-link

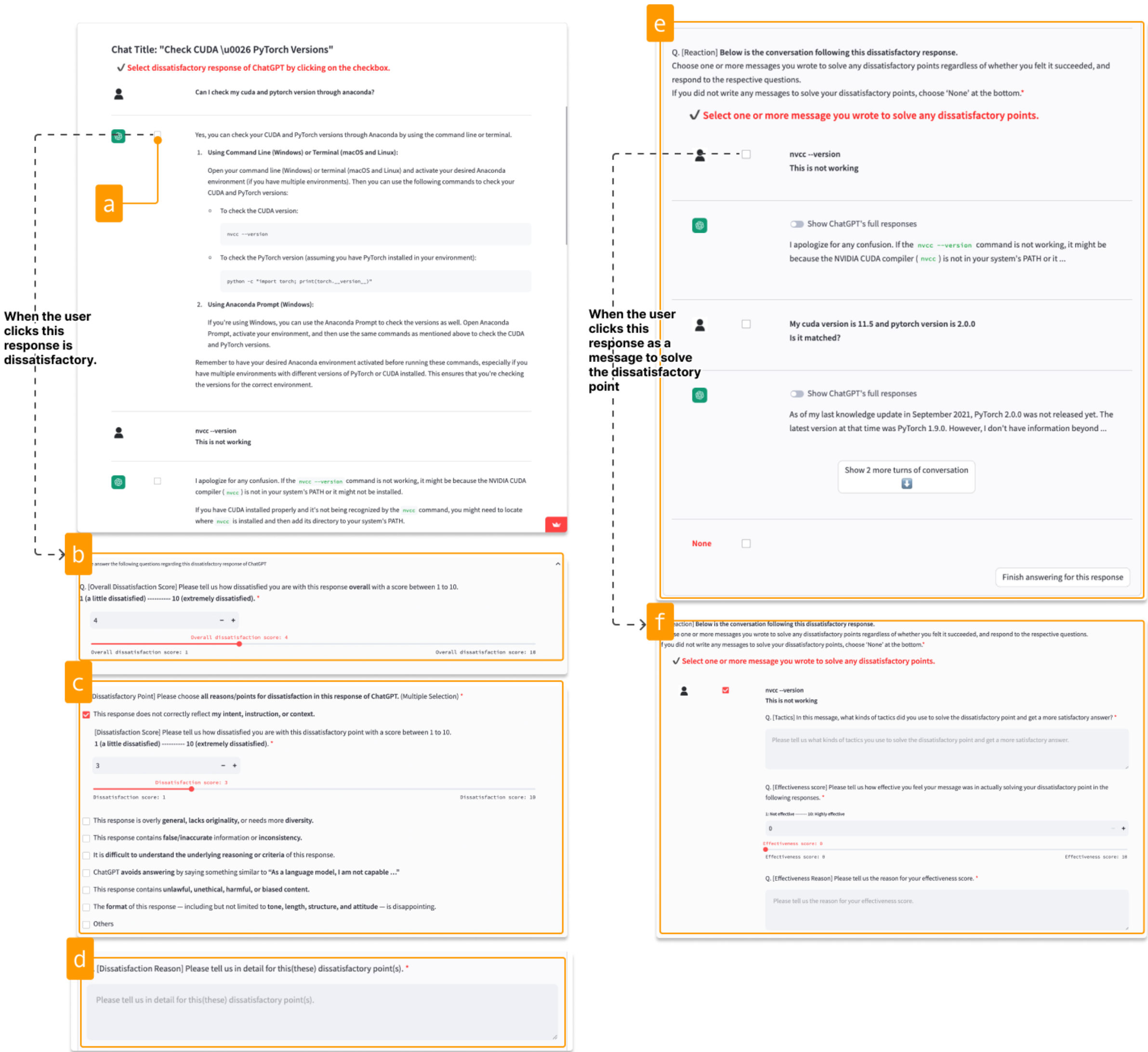



Figure 4.1: Screenshot of the data collection system.

10: extremely dissatisfied) (Fig 4.1-b), (2) choose one or more dissatisfaction categories from the given seven categories, or optionally describe a custom dissatisfaction point for dissatisfaction (Fig 4.1-c), (3) rate the level of dissatisfaction for each selected category on a scale of 1 to 10 (1: a little dissatisfied, 10: extremely dissatisfied) (Fig 4.1-c), (4) provide a detailed free-form explanation for their dissatisfaction (Fig 4.1-d), (5) select a prompt among the subsequent conversations in which they tried to resolve the dissatisfaction (Fig 4.1-e), (6) describe their tactic to address the dissatisfaction in the prompt (Fig 4.1-f), (7) rate the effectiveness of their tactic on a scale of 1 to 10 (1: not effective, 10: highly effective) (Fig 4.1-f), (8) provide a written explanation of the reasons for their effectiveness rating (Fig 4.1-f). In cases where there was no subsequent prompt or the conversation ended after dissatisfaction, participants were asked to provide written reasons instead of responding to (5)-(8).

## 4.2 Collected Data

### 4.2.1 Participants and Collected Data

We distributed the data collection system to people over the age of 18 globally through the Prolific platform [2]. Participants who provided at least two ChatGPT conversation links and evaluated at least one dissatisfactory response for each link received a compensation of £6. For each additional dissatisfactory response submitted from a single conversation link, participants received an additional £0.75 per response. For each additional conversation link provided beyond the initial two, participants received an additional £1.5 per link. We limited the number of maximum conversation links that can be submitted to five for each participant to prevent one participant from providing lots of conversation links. In total, we collected 307 ChatGPT conversation links, 511 dissatisfactory ChatGPT responses, and 615 user responses regarding those dissatisfactions from 107 individuals. Each user submitted an average of 2.87 links (std=1.21), 4.78 dissatisfactory ChatGPT responses (std=5.61), and 5.75 responses regarding those dissatisfactions (std=6.62). This study was approached by our institution's IRB, and we received consent from participants for the release of datasets.

### 4.2.2 Data Filtering and Pre-processing

To ensure the quality and reliability of the data collected from our system, two authors reviewed all the data together according to the following criteria and conducted filtering or pre-processing where necessary.

**Filtering Process** The data was filtered out at three levels: (1) user, (2) conversation, and (3) dissatisfactory responses.

***1. User-Level Filtering*** We identified that one participant provided altogether contradictory responses, which contradicted the dissatisfactory response and the effectiveness of the prompt in resolving the dissatisfaction. Consequently, all data from this user were excluded.

***2. Conversation-Level Filtering*** The conversation-level filtering was conducted based on the following four criteria, and a total of 20 conversations were filtered out. The detailed reason for each criteria is in the Appendix (Sec 9.2).

1. Conversation older than 30 days.
2. Conversation with a memory level of 3 or lower.
3. Conversation for fun or testing purposes.
4. Conversation from versions other than GPT-3.5.

***3. Response-Level Filtering*** Response-level filtering was conducted based on the following four criteria, leading to the exclusion of a total of 16 dissatisfactory ChatGPT responses.

1. Dissatisfaction due to ChatGPT's error messages
2. Unconvincing dissatisfaction
3. Mismatch between score and reason

[2] https://www.prolific.co/

4. No Correlation between selected dissatisfactions and subsequent prompts for resolving that dissatisfaction

Detailed reasons and examples of each filtering case can be found in the Appendix and supplementary material. Please note that when filtering at the response level, all associated subsequent prompts and tactic data related to that response were also filtered. When filtering at the conversation level, all data related to the ChatGPT dissatisfactory responses and user prompts within that conversation were also filtered out. When filtering at the user level, all data provided by that user were excluded.

**Pre-processing Process** The data pre-processing process primarily involved the reassignment of dissatisfaction categories. This step was undertaken to deal with cases where participants incorrectly selected dissatisfaction categories or opted for the 'other' option when evaluating the dissatisfaction category. Two authors examined all the data and carried out reassignment according to the following two criteria, proceeding only when a consensus was reached. Detailed examples of each case where reassignment occurred can be found in the supplemental.

Criterion 1: Reassigning 'other' to a specific category. For the 'other' option, when we found that there was a more suitable match with another category that was not selected based on the dissatisfaction reason, the 'other' score was reallocated to the corresponding category. As a result of this criterion, four entries were reassigned to the $\mathbf{D_{intent}}$ category, two to $\mathbf{D_{depth}}$, three to $\mathbf{D_{acc}}$, and five to $\mathbf{D_{format}}$.

Criterion 2: Reassigning an incorrectly selected category to another. If a participant had only checked one dissatisfaction category, and upon reviewing the dissatisfaction reason and conversation, it was evident that the selected category was not appropriate but another category was a better fit, the score was reassigned to the more suitable category. Using this criterion, three entries were reallocated from $\mathbf{D_{intent}}$ to $\mathbf{D_{format}}$, three from $\mathbf{D_{intent}}$ to $\mathbf{D_{acc}}$, two from $\mathbf{D_{acc}}$ to $\mathbf{D_{intent}}$, one from $\mathbf{D_{acc}}$ to $\mathbf{D_{depth}}$, and two from $\mathbf{D_{depth}}$ to $\mathbf{D_{format}}$.

### 4.2.3 Dataset

After filtering and pre-processing, we built a dataset on end-users' dissatisfaction with ChatGPT and their responses. The dataset is hierarchically organized, comprising the following components:

1. User (N=94)

2. ChatGPT conversation links and logs (N=249)

3. User's recollected experience data on dissatisfactory ChatGPT responses (N=377)

4. User's strategies to respond to the dissatisfactory response (N=459)

Here, the user's strategies were qualitatively analyzed, resulting in the creation of 13 tactic codes categorized into four themes. More detail of this is in Sec 5.2. Each data is also labeled as corresponding tactic codes by the authors. With this dataset, we conducted a quantitative and qualitative analysis to answer our research questions. We provide this dataset to facilitate future research about user experiences on chat-based LLMs. In releasing the dataset, we took careful consideration by masking all sensitive information related to their privacy and personal information. A more detailed description about the dataset can be accessed through our project website [3].

[3]https://chatgpt-analysis.kixlab.org

# Chapter 5. Data Analysis and Results

In this section, we present the analysis method and results that answer our research questions based on the constructed dataset. Firstly, we present the analysis of the types of dissatisfaction users face in LLM responses (RQ1). Next, we present how users respond to dissatisfaction through qualitative analysis (Table 5.2) and analyze the effectiveness of the tactics users use (RQ2). Finally, we present how users' knowledge level regarding LLM influences their experiences of dissatisfaction and their behaviors when they face dissatisfaction (RQ3).

| Dissatisfaction Category | Response-level analysis | | User-level analysis |
|---|---|---|---|
| | Count: N (%) | Dissatisfaction Score: mean (std)* | Frequency: mean (std)* |
| Dintent | **168 (32.18%)** | 5.56 (2.94) | **0.47 (0.03)** |
| Ddepth | 107 (20.50%) | **5.09 (2.69) *** | 0.33 (0.35) |
| Dacc | 83 (15.90%) | **6.52 (2.76) *** | 0.20 (0.03) |
| Dtrans | 27 (5.17%) | 4.81 (3.13) | 0.08 (0.02) |
| Drefuse | 27 (5.17%) | 6.37 (2.68) | 0.09 (0.02) |
| Dethic | 4 (0.77%) | 6.25 (3.20) | 0.01 (0.01) |
| Dformat | 106 (20.31%) | 6.14 (3.04) | 0.27 (0.03) |

Table 5.1: Analysis results on the count, dissatisfaction score, and user-level frequency for the dissatisfaction category (* p-value < 0.01)

## 5.1 RQ1. Analysis of how users experience dissatisfaction

### 5.1.1 Dissatisfaction Category Analysis

We analyzed the count, distribution, and dissatisfaction score of the seven categories of dissatisfaction organized through a systematic literature review in Section 3, and the results are described in Table 5.1. In terms of the count of each category, $\mathbf{D_{intent}}$ accounted for the largest proportion (32.18%), while $\mathbf{D_{trans}}$, $\mathbf{D_{refuse}}$, and $\mathbf{D_{ethic}}$ constituted significantly smaller proportions compared to the other categories. To investigate the severity degree of user dissatisfaction in each category, we conducted Kruskal-Wallis test and confirmed significant differences between categories ($\chi^2 = 17.6$, p-value ¡ 0.01, df = 6). In particular, we found that $\mathbf{D_{acc}}$'s dissatisfaction score was the highest, and its score was statistically significantly higher than $\mathbf{D_{depth}}$ through Dwass-Steel-Critchlow-Fligner(DSCF) pairwise comparison (p-value=0.008). This means that users are statistically significantly more dissatisfied with dissatisfaction due to $\mathbf{D_{acc}}$ than $\mathbf{D_{depth}}$.

Considering that each user provided multiple dissatisfactory responses, we also conducted a user-level analysis, accounting for potential correlations among the data submitted by the same user. To achieve this, we normalized each dissatisfaction category data by dividing them by the number of dissatisfactory responses each user submitted. This method allowed us to express each data point as the frequency of how often each user experienced dissatisfaction in a certain category. The analysis results are presented in Table 5.1 in the "User-level" analysis column. The mean frequency value of $\mathbf{D_{intent}}$ was 0.47, indicating that if a user has experienced 100 dissatisfactory ChatGPT responses, on average, 47 of them fall into the

$\mathbf{D_{intent}}$ category. Furthermore, the Kruskal-Wallis test result shows statistically significant differences in user-level frequency values between each category ($\chi^2 = 9.93$, p-value ¡ 0.01, df = 6). In the user-level analysis, we can see a similar tendency to the response-level analysis, users experience $\mathbf{D_{intent}}$ the most frequently. Following this, the second most frequently encountered dissatisfaction is $\mathbf{D_{depth}}$. However, the standard deviation of $\mathbf{D_{depth}}$ is 0.35, which is much higher than other categories, indicating that the frequency of experiencing $\mathbf{D_{depth}}$ varies significantly from user to user.

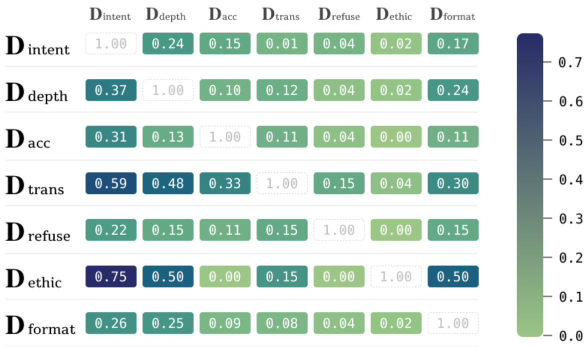



Figure 5.1: Normalized Co-occurrence matrix of dissatisfaction category. The value at $(i, j)$ in this matrix represents the frequency of when the $i$th row was selected as a dissatisfaction point, the $j$th column was also selected as a dissatisfaction.

### 5.1.2 Co-occurrence Analysis

In a single dissatisfactory response, multiple dissatisfaction categories can co-occur. For example, a user may simultaneously experience dissatisfaction with the lack of originality ($\mathbf{D_{depth}}$) and the length ($\mathbf{D_{format}}$) of ChatGPT's response at the same time. Therefore, we analyzed co-occurrence patterns to investigate the correlations between each category of dissatisfaction. Results are presented in Fig 5.1 and the value at $(i, j)$ in this matrix represents the frequency of when the $i$-th row was selected as a source of dissatisfaction, the $j$-th column was also selected together. The result shows that $\mathbf{D_{intent}}$ frequently appears concurrently with all other categories. Also, while $\mathbf{D_{trans}}$ and $\mathbf{D_{ethic}}$ have relatively low counts, they co-occur with $\mathbf{D_{intent}}$ more than half the times in each occurrence.

## 5.2 RQ2. Analysis of how users respond to dissatisfaction

### 5.2.1 Categorizing Tactics for Resolving Dissatisfaction

Through qualitative analysis, we categorized users' tactics to understand and analyze how users address their dissatisfaction from ChatGPT's response through subsequent prompts. Two authors independently conducted open coding by reviewing ChatGPT conversation log data, user-side dissatisfactions on ChatGPT responses, employed tactics in subsequent prompts, and user-reported effectiveness and the reasons for these tactics. After completing the open coding, the two authors engaged in an iterative process of code consolidation. To precisely capture and categorize the subtleties of user tactics, both authors iterated all data together, making a code set through discussion. We proceeded with these processes until the authors met a common ground. After two times of iterations, we identified the user's tactic with 13 codes as presented in Table 5.2. To establish relationships between these codes and identify

| **Category (4)** | **Tactic Code (13)** |
|---|---|
| Prompt Reusing and Shortcut ($\mathbf{T_{repeat}}$) | T1: Re-using an identical prompt or slightly paraphrasing it<br>T2: Using the specific word (e.g., more, another) that implies requesting different or more outputs for the same task as the previous prompt<br>T3: Re-using an identical prompt but adding emphasis through formatting (e.g., using all capital letters, using double quotation marks) |
| Intent Concretization ($\mathbf{T_{specify}}$) | T4: Specifying user intent by providing detailed or direct instructions<br>T5: Specifying user intent by providing additional context or explanation<br>T6: Adding format-specific conditions (e.g., make it shorter, provide in list format)<br>T7: Adding tone-specific conditions (e.g., make it casual) |
| Error Identification and Correction ($\mathbf{T_{error}}$) | T8: Pointing out errors or mistakes<br>T9: Providing the correct answer or hints<br>T10: Asking clarification questions |
| Task Adaptation ($\mathbf{T_{adapt}}$) | T11: Adapting by shifting to another topic or task that is different from the original intent.<br>T12: Breaking down the original task into smaller subtasks<br>T13: Asking follow-up questions deviating from the original task |
| **No Tactic** | No further prompting to address the dissatisfaction and even terminating the conversation due to dissatisfaction |

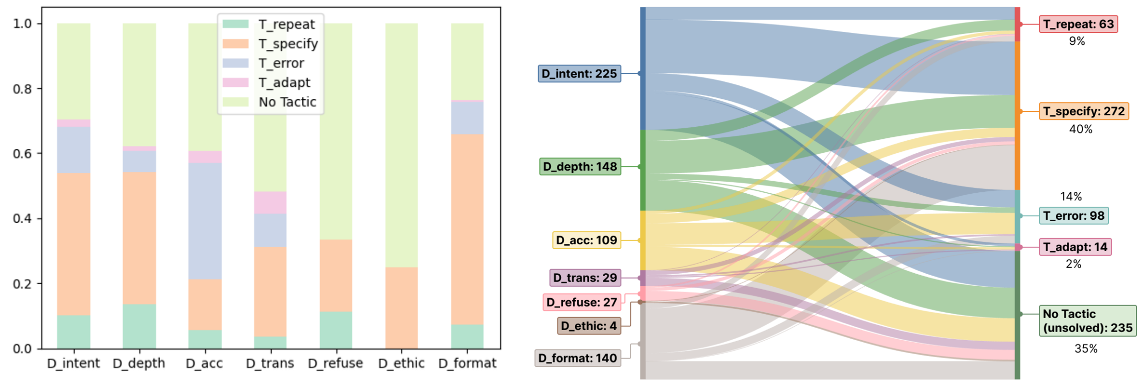
Table 5.2: User tactic category

overarching themes, axial coding [59] was performed. Through this coding process, we identified four main themes of the user's tactics, as presented in Table 5.2.

**Tactic Category 1. Prompt Reusing and Shortcut** This category of tactic represents users either reusing prompts or employing a single word to request similar or diverse responses, often requiring minimal effort in crafting the prompt. This category comprises three tactics. First, users just reuse the exact same prompt as the previous one or paraphrase it slightly (T1). Second, users use a single word like 'more' or 'another' as a shortcut to get either similar responses from the previous turn or a wider range of responses from ChatGPT (T2). Last, users retry by adding emphasis through formatting, such as using all capital letters or using double quotation marks (T3).

**Tactic Category 2. Intent Concretization** This category encompasses four tactics of users trying to concretize their intent and context to get a more appropriate response. Users further specify their needs by providing more detailed or direct instructions (T4), giving additional context or explanation (T5). For example, if users ask ChatGPT to recommend a dinner menu and they doesn't like ChatGPT's answer, they can further specify their needs by saying, "Recommend a **healthy** dinner menu using **tomatoes**" (T4), or explain their context by saying, "I'm going to invite a guest to my house for my dinner" (T5). And users concretize their intent by adding specific conditions related to the format such as "make it shorter" (T6), and adding specific conditions related to the tone, such as "make it casual" (T7).

**Tactic Category 3. Error Identification and Correction** This category mainly contains tactics when there are some errors in the ChatGPT's response, and the users point out or correct them. Users




(a) (b)

Figure 5.2: (a) Distribution of tactic categories by dissatisfaction category. (b) Sankey diagram to visualize how users respond among four tactic categories or **No Tactic** after experiencing each of the dissatisfaction categories. Note that the count in the Sankey diagram can be greater than the count of response-level analysis in Table 5.1 and 5.3. This is because one response can include multiple dissatisfaction categories and multiple tactic categories, and they were counted multiple times to draw a Sankey diagram.

simply say “It was wrong.” or point out the part that is wrong (T8), give the correct answer or hints of the correct answer (T9), and ask a clarification question to confirm the error or doubtful aspects such as by asking “Can you confirm that ... ?” or “Are you sure ...?”(T10).

**Tactic Category 4. Task Adaptation** This category represents the user adjusting to another task instead of the original task where the user felt dissatisfied. Users adapt their task by altering their initial task to a different one (T11). For instance, if users initially ask for the latest information and ChatGPT says it can only answer up to 2021 information, then they can slightly adjust their original task and ask for 2021 information rather than the latest information. Users also adjust their original task by dividing it into smaller and more manageable subtasks (T12). For example, when users ask ChatGPT for a complex math problem, they can ask them in intermediate steps. Finally, Users ask follow-up questions deviating from the original task, such as asking follow-up questions about parts that lack details or are unfamiliar to them in ChatGPT’s responses. (T13).

### 5.2.2 Tactic Category Analysis

After creating the tactic categories, we categorized users’ prompts into four tactic categories or **No Tactic**. **No Tactic** indicates no further prompting to address the dissatisfaction and even terminating the conversation due to dissatisfaction. Here, note that a single-user prompt can encompass multiple tactic categories if the prompt contains multiple requests. We conducted response-level analysis for the count, distribution, and effectiveness of each tactic as well as user-level analysis for frequency (Table 5.3). Notably, we observed that $\mathbf{T_{specify}}$ stands out as the dominant category, and it accounts for over half of the distribution (58.6%) among the four tactic categories without **No Tactic**. In addition, we analyzed the effectiveness of each tactic based on users’ rating of the effectiveness score between 1 and 10. We conducted a Kruskal-Wallis test and confirmed that there are statistically significant differences between the effectiveness scores of each tactic ($\chi^2 = 23.1$, p-value ¡ 0.01, df = 4). Specifically, we found that $\mathbf{T_{specify}}$, a tactic for users to further specify their own intents, is most effective with a mean score

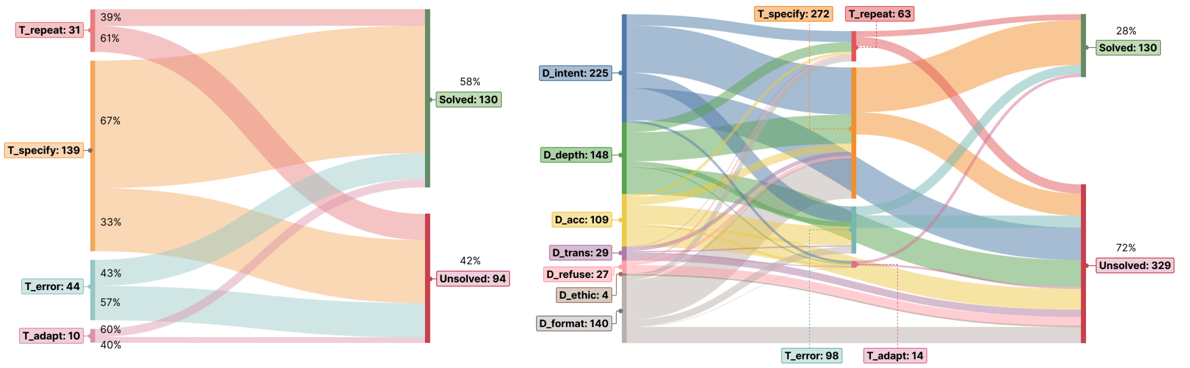
0f 6.04, highest of all categories.

| Tactic Category | Tactic Code | Response-level analysis | | | | User-level analysis | |
|---|---|---|---|---|---|---|---|
| | | Count: N (%) | | Effectiveness Score: mean (std) | | Frequency: mean (std) | |
| | | Category | Code | Category* | Code | Category | Code |
| $\mathbf{T_{repeat}}$ | T1 | 45 (9.4%) | 29 (5.8%) | 4.04 (3.16) | 4.45 (3.15) | 0.09 (0.20) | 0.07 (0.18) |
| | T2 | | 18 (3.6%) | | 3.06 (3.06) | | 0.02 (0.09) |
| | T3 | | 2 (0.4%) | | 1.00 (0.00) | | 0.00 (0.04) |
| $\mathbf{T_{specify}}$ | T4 | **183 (38.4%)** | 122 (24.4%) | **6.04 (3.44)** | 6.25 (3.53) | **0.33 (0.34)** | 0.22 (0.28) |
| | T5 | | 26 (5.2%) | | 5.35 (3.33) | | 0.06 (0.15) |
| | T6 | | 40 (8.0%) | | 6.45 (3.16) | | 0.08 (0.17) |
| | T7 | | 11 (2.2%) | | 4.73 (3.04) | | 0.02 (0.10) |
| $\mathbf{T_{error}}$ | T8 | 73 (15.3%) | 53 (10.6%) | 4.19 (2.95) | 4.26 (2.99) | 0.10 (0.22) | 0.06 (0.16) |
| | T9 | | 13 (2.6%) | | 4.62 (2.66) | | 0.02 (0.09) |
| | T10 | | 10 (2.0%) | | 3.80 (3.16) | | 0.03 (0.10) |
| $\mathbf{T_{adapt}}$ | T11 | 12 (2.5%) | 7 (1.4%) | 5.17 (3.04) | 4.57 (3.21) | 0.04 (0.11) | 0.03 (0.10) |
| | T12 | | 2 (0.4%) | | 8.00 (0.00) | | 0.00 (0.03) |
| | T13 | | 3 (0.6%) | | 4.67 (3.22) | | 0.00 (0.04) |
| **No Tactic** | | 164 (34.4%) | 164 (32.8%) | - | - | 0.47 (0.38) | 0.47 (0.36) |

Table 5.3: Analysis results on the count, effectiveness score, and user-level frequency for the tactic category (* p-value $<$ 0.01)

### 5.2.3 Dissatisfaction Category and Corresponding Tactics: Whether the dissatisfaction was solved

We investigated how users applied different tactics to address each dissatisfaction category and whether these tactics resolved the dissatisfaction. Firstly, we analyzed the distribution of tactics used for each dissatisfaction category (Fig. 5.2(a)), and drew a Sankey diagram to visualize the overall flow of tactics used by each dissatisfaction category (Fig. 5.2(b)). We observed that $\mathbf{T_{specify}}$ is the dominant tactic across various dissatisfaction categories. However, when users encounter dissatisfaction related to the accuracy of information ($\mathbf{D_{acc}}$), they tend to employ $\mathbf{T_{error}}$ rather than $\mathbf{T_{specify}}$. Lastly, in cases of $\mathbf{D_{trans}}$, $\mathbf{D_{refuse}}$, and $\mathbf{D_{ethic}}$, users often resort to **No Tactic**, ending up the conversation. The proportion and visualization of whether or not dissatisfaction has been resolved by each tactic can be seen in Fig. 5.3(a). Fig. 5.3(a) illustrates that the users managed to resolve their dissatisfaction by 58 % by utilizing tactics. Notably, $\mathbf{T_{specify}}$ was an effective way of resolving dissatisfaction in many cases (67%), while with other tactics, there were more cases where dissatisfaction remained unsolved. Fig. 5.3(b) shows which tactics users use for each dissatisfaction category and how this eventually leads to resolve the dissatisfaction. Through this analysis, we can observe the overall flow of how users, while conversing with ChatGPT, experience various dissatisfactions in what proportion, how they respond to them using different tactics, and how this leads to the resolution of these dissatisfactions. When users encounter dissatisfaction, approximately 34% opt for **No Tactic** while 66% employ tactics. However, it can be seen that approximately 58% of dissatisfactinos are resolved through tactics. In the end, users manage to resolve only 28% of their dissatisfactions using tactics, leaving 72% of dissatisfactions unresolved.




(a)

(b)

Figure 5.3: (a) A Sankey diagram that visualizes whether users resolved their dissatisfaction using each of the tactic categories. (b) The overall visualization of how users respond among the four tactic categories after experiencing each of the dissatisfaction categories and finally whether that dissatisfaction was solved or not.

## 5.3 RQ3. Analysis of how dissatisfaction and tactics vary based on the user's knowledge level of LLMs

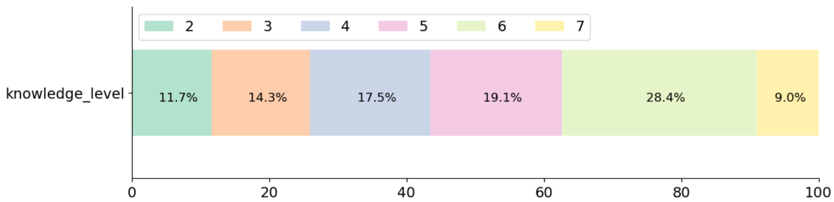



Figure 5.4: Distribution of participants' knowledge level regarding LLM on a 7-point scale
(1: very low, 7: very high).
None of the participants reported a knowledge level of 1.

We analyzed how users' experience of dissatisfaction and their tactics differ depending on their knowledge levels regarding LLMs. First, we examined the distribution of users' knowledge levels regarding LLMs in our dataset, as depicted in Fig. 5.4. We collected the knowledge level data about LLMs on a 7-point scale, where 1 indicates very low knowledge, and 7 indicates very high knowledge. We divided the groups into "low knowledge level" (those with a knowledge level 1-3) and " high knowledge level" (those with a knowledge level 5-7), as four lies in the middle of the 7-point scale.

To investigate whether there is a difference in the distribution of dissatisfaction categories between the two groups, we conducted a chi-square test for the dissatisfaction categories of each group and found that there were statistically significant differences in the distribution of dissatisfaction categories by different knowledge groups ($\chi^2 = 17.7$, p-value ¡ 0.01). Specifically, we observed that the low-knowledge group experiences $\mathbf{D_{depth}}$ ((count: 26.97%, user-level frequency: 0.38)) and $\mathbf{D_{refuse}}$ (count: 8.55%, user-level frequency: 0.14) more frequently, while the high-knowledge group experiences $\mathbf{D_{acc}}$ (count: 17.38%, user-level frequency: 0.24) and $\mathbf{D_{format}}$ (count: 24.82%, user-level frequency: 0.28) more frequently. On the other hand, we conducted a Mann-Whitney U test to investigate the differences in dissatisfaction scores between knowledge groups, but there were no significant differences.

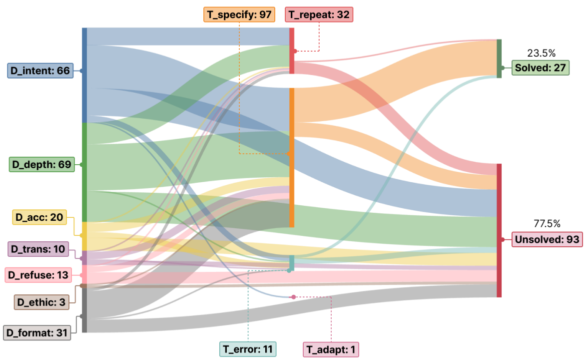
| Dissatisfaction Category | Response-level analysis: Count: N (%) * | | Response-level analysis: Dissatisfaction Score: mean(std) | | User-level analysis: Frequency: mean (std) | |
|---|---|---|---|---|---|---|
| | high | low | high | low | high | low |
| $\mathbf{D_{intent}}$ | 89 (31.56%) | 45 (29.61%) | 5.91 (2.85) | 5.18 (3.08) | 0.43 (0.30) | 0.49 (0.39) |
| $\mathbf{D_{depth}}$ | **50 (17.73%)** | **41 (26.97%)** | 5.02 (2.70) | 5.22 (2.72) | **0.30 (0.31)** | **0.38 (0.38)** |
| $\mathbf{D_{acc}}$ | **49 (17.38%)** | **18 (11.84%)** | 6.73 (2.85) | 6.5 (2.62) | **0.24 (0.29)** | **0.14 (0.21)** |
| $\mathbf{D_{trans}}$ | 12 (4.26%) | 9 (5.92%) | 5.25 (3.33) | 3.67 (3.00) | 0.07 (0.16) | 0.10 ( 0.23) |
| $\mathbf{D_{refuse}}$ | **11 (3.90%)** | **13 (8.55%)** | 6.82 (2.79) | 6.92 (2.02) | **0.07 (0.16)** | **0.14 (0.26)** |
| $\mathbf{D_{ethic}}$ | 1 (0.35%) | 3 (1.97%) | 3 (-) | 7.33 (2.89) | 0.01 (0.07) | 0.03 (0.08) |
| $\mathbf{D_{format}}$ | **70 (24.82%)** | **23 (15.13%)** | 6.66 (2.86) | 5.7 (3.36) | **0.28 (0.32)** | **0.25 (0.37)** |

Table 5.4: Dissatisfaction category for knowledge level high and low group (* p-value < 0.01)

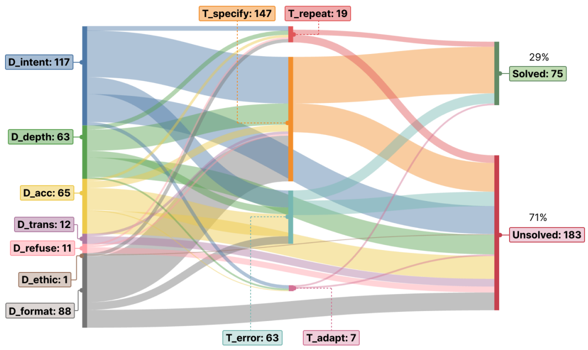
| Tactic Category | Tactic Code | Response-level analysis: Count: N (%) | | | | Response-level analysis: Effectiveness Score: mean (std) | | | | User-level analysis: Frequency: mean (std) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Category*** | | Code | | Category | | Code | | Category | | Code | |
| | | high | low | high | low | high | low | high | low | high | low | high | low |
| $\mathbf{T_{repeat}}$ | T1 | | | 12 (4.4%) | 9 (6.5%) | | | 4.75 (2.96) | 3.00 (3.00) | | | 0.06 (0.14) | 0.08 (0.22) |
| | T2 | 16 (6.11%) | **19 (14.5%)** | 4 (1.5%) | 12 (8.7%) | **5.06 (3.00)*** | **2.37 (2.27)*** | 6.00 (3.37) | 1.67 (1.15) | 0.08 (0.17) | 0.11 (0.25) | 0.02 (0.08) | 0.04 (0.13) |
| | T3 | | | 1 (0.4%) | 0 (0.0%) | | | 1 (-) | - (-) | | | 0.00 (0.02) | 0.23 (0.35) |
| $\mathbf{T_{specify}}$ | T4 | | | 84 (30.8%) | 24 (17.4%) | | | 5.77 (3.71) | 7.17 (2.78) | | | 0.23 (0.25) | - |
| | T5 | 111 (42.37%) | 52 (39.7%) | 13 (4.8%) | 12 (8.7%) | 5.88 (3.56) | 6 (3.33) | 5.00 (3.03) | 5.42 (3.73) | 0.34 (0.31) | 0.39 (0.41) | 0.05 (0.11) | 0.10 (0.22) |
| | T6 | | | 19 (7.0%) | 13 (9.4%) | | | 7.00 (2.83) | 6.00 (3.70) | | | 0.09 (0.20) | 0.07 (0.14) |
| | T7 | | | 4 (1.5%) | 7 (5.1%) | | | 6.00 (4.08) | 4.00 (2.31) | | | 0.01 (0.05) | 0.05 (0.18) |
| $\mathbf{T_{error}}$ | T8 | | | 37 (13.6%) | 5 (3.6%) | | | 3.81 (4.08) | 5.2 (3.83) | | | 0.07 (0.18) | 0.05 (0.12) |
| | T9 | **49 (18.70%)** | 8 (6.1%) | 7 (2.6%) | 2.00 (1.4%) | 3.53 (2.60) | 5.75 (3.06) | 3.57 (1.90) | 7.00 (1.41) | 0.12 (0.24) | 0.08 (0.18) | 0.03 (0.09) | 0.02 (0.10) |
| | T10 | | | 6 (2.2%) | 2 (1.4%) | | | 2 (2.45) | 7.00 (1.41) | | | 0.03 (0.12) | 0.03 (0.12) |
| $\mathbf{T_{adapt}}$ | T11 | | | 4 (1.5%) | 1 (0.7%) | | | 5.25 (3.10) | 1 (-) | | | 0.03 (0.11) | 0.01 (0.07) |
| | T12 | 5 (1.91%) | 1 (0.8%) | 0 (0.0%) | 0 (0.0%) | 4.40 (3.29) | 1 (-) | - | - | 0.03 (0.11) | 0.01 (0.07) | - | - |
| | T13 | | | 1 (0.4%) | 0 (0.0%) | | | 1 (-) | - | | | 0.00 (0.01) | - |
| No Tactic | | 81 (30.92%) | **51 (38.9%)** | 81 (29.7%) | 51 (37.0%) | - | - | - | - | 0.47 (0.37) | 0.44 (0.39) | 0.47 (0.37) | 0.44 (0.39) |

Table 5.5: Tactic category and code for knowledge level high and low group (* p-value < 0.01)

Similarly, we conducted a chi-square test for tactic categories and found significant differences in the count of tactic categories among the two groups ($\chi^2$ = 21.6, p-value ¡ 0.01). In particular, **No Tactic** was more prevalent in the low-knowledge group. Additionally, $\mathbf{T_{repeat}}$, which involves minimal prompt engineering, was more commonly used in the low-knowledge group, while $\mathbf{T_{error}}$, aimed at pointing out and rectifying errors in ChatGPT responses, was more prevalent in the high-knowledge group. Furthermore, to compare and analyze the effectiveness of the tactics used in each knowledge group, we performed a Mann-Whitney U test on the effectiveness scores of tactic categories, which were collected from users. Through this test, we found that the effectiveness of the $\mathbf{T_{repeat}}$ was statistically higher in the high-knowledge group (p-value ¡ 0.01, effect size= 0.5789). Fig. 5.5(a) and 5.5(b) present Sankey diagrams that illustrate how users in the low-knowledge and high-knowledge groups experience dissatisfaction categories from ChatGPT's responses, respond to the dissatisfactions with each tactic category at user prompts, and whether these tactics ultimately resolve their dissatisfactions or not. Through this, we can see that the rate of resolving dissatisfaction in the high-knowledge group (29%) is higher than low-knowledge group (23.5%).




(a)




(b)

Figure 5.5: Sankey diagrams by users' knowledge level of LLMs that visualize how users respond among four tactic categories after experiencing each of the dissatisfaction categories and finally whether that dissatisfaction was solved or not. (a): Low-knowledge group's Sankey diagram (b): High-knowledge group's Sankey diagram.

# Chapter 6. Discussion

In this section, we first discuss the interpretation of our results and their implication. Second, we suggest design implications for building LLMs with better usability based on our study results. Lastly, we discuss the limitations of our study and future work.

## 6.1 Interpretation of results

Building upon the analysis of user-side dissatisfaction and corresponding user tactics during the conversation, we discuss the most prevalent, severe, and unaddressed categories of dissatisfaction and their implications. We also discuss the differences in dissatisfaction and corresponding tactics across users with different knowledge levels about LLM.

### 6.1.1 The Most Prevalent Dissatisfaction and Tactics.

Our results suggest that despite the advances in LLMs to align with the user intent, there still exists much room for improvement from the users' perspective. With recent advancements in LLMs and the introduction of techniques to align LLMs with user intents, such as Reinforcement Learning from Human Feedback (RLHF), LLMs are now known to better align with human intent than before [3, 60, 61]. However, we found that $\mathbf{D_{intent}}$, the dissatisfaction in terms of understanding users' intent, is the most prevalent (Table 5.1) and frequently co-occurring with other dissatisfaction categories (Figure 5.1). We also discovered that users frequently use $\mathbf{T_{specify}}$ that further specify their intent to address the dissatisfaction. Moreover, users rated $\mathbf{T_{specify}}$ as the most effective among tactic categories, but there are still many cases (about 42%) where dissatisfaction was not resolved despite using this tactic. This may be because users have difficulty clearly representing their intent. Previous work on web search and information retrieval has also noticed this problem [62], and there exist several methods to better support users to specify their intent in these domains, such as context-sensitive query auto-completion [63] and context-based term suggestions [64]. Similarly, in the context of LLMs, further research is needed to support users to specify their intents based on the user's context.

### 6.1.2 The Most Severe or Unaddressed Dissatisfaction

Self-reported scores on the level of dissatisfaction show that users perceived the dissatisfaction of $\mathbf{D_{acc}}$ to be the most severe (Table 5.4). This supports that users feel greatly dissatisfied with various known limitations of LLMs related to information accuracy such as hallucination [65, 16, 17, 18], inconsistency or incorrectness in the responses [66, 20, 18, 19], and the inability of ChatGPT to provide updated information [29, 53]. Furthermore, our findings show that users tend to respond to this dissatisfaction primarily by pointing out LLM's faults or correcting them ($\mathbf{T_{error}}$), but more than half of them (57%) nevertheless fail to resolve this dissatisfaction.

We also found that when users encountered dissatisfaction when their prompts were refused to answer ($\mathbf{D_{refuse}}$), when ethical concerns or biases were found in the response ($\mathbf{D_{ethic}}$), or when they had a lack of understanding of the internal logic of the generated response ($\mathbf{D_{trans}}$), they often did not attempt to address the dissatisfaction or even terminated the conversation. For instance, one user explained their decision to end the conversation as follows: "I ended the conversation as I felt like there was no common understanding and was not looking forward to explaining myself any further than my original prompt." Through this, we can see that if the users experience such dissatisfaction, they not only have difficulty communicating with ChatGPT but also have no idea how to further improve their prompts, often terminating the conversation.

One notable point here is that $\mathbf{D_{ethic}}$ and $\mathbf{D_{refuse}}$ can be in a trade-off relationship. Including OpenAI [1], the company that developed ChatGPT, many companies have adopted a strategy where the LLM avoids answering when faced with potentially unethical or biased prompts, responding with statements like "As a language model, I am not capable of ... " [67, 30]. Although companies could avoid being embroiled in ethical issues, this approach might have introduced another dimension of dissatisfaction ($\mathbf{D_{refuse}}$) for users. Self-reported scores on the level of dissatisfaction show the level of severity for both $\mathbf{D_{ethic}}$ and $\mathbf{D_{refuse}}$ are similar (Table 5.4). This suggests that the current approach of refusing to answer instead of giving responses with ethical concerns may not reduce users' overall dissatisfaction. Thus, it is necessary to find other measures that could also lower the users' dissatisfaction when faced with unethical or biased prompts.

### 6.1.3 Dissatisfaction and Corresponding Tactics Difference Across LLM Knowledge Level

Our result revealed that there exist significant differences in dissatisfaction and employed tactics between high- and low-knowledge user groups. We observed that the low-knowledge group reports a higher occurrence of $\mathbf{D_{depth}}$—dissatisfaction that ChatGPT's response is too general and lacks detail or originality—than the high-knowledge group (Table 5.4). One possible reason behind this is that the low-knowledge user group might have overestimated ChatGPT's creative capabilities. This could be because low-knowledge user groups may be more prone to unconditionally accepting media or news which states that ChatGPT can perform creative tasks such as writing poetry and song lyrics [68, 69]. This may have led them to expect more creative responses, resulting in a higher possibility of feeling disappointment. In contrast, the high-knowledge group may have possessed a better understanding of ChatGPT's limitations. Knowing that ChatGPT's responses are based on trained patterns from existing datasets could have allowed them to be more generous towards the responses that lack originality. We speculate that the low-knowledge group might have a less accurate mental model of the capacity of LLM, misunderstand its capabilities, and experience more dissatisfaction in terms of $\mathbf{D_{depth}}$.

Moreover, the tactics employed in response to these dissatisfactions differed between the two groups. Compared to the high-knowledge group, the low-knowledge group relied more on 'No tactic' and more frequently used $\mathbf{T_{repeat}}$, which requires minimal effort for prompt writing (Table 5.5). This may be because the low-knowledge users may not know much about the various options of tactics they could take. Interestingly, however, although high-knowledge users used $\mathbf{T_{repeat}}$ less, they found it more effective in solving their dissatisfaction. This may indicate that high-knowledge users tend to have a better sense of when is the right time to use $\mathbf{T_{repeat}}$.

[1] https://openai.com/

## 6.2 Design Implications for Building LLMs with Better Usability

Based on our study result, we suggest three design implications to enhance the usability of LLMs: (1) supporting users to represent their intent, (2) recommending effective multi-turn prompt tactics to users, and (3) providing personalized LLM experiences to users.

### 6.2.1 Supporting users to represent their intent better

We suggest a design that facilitates a better representation of the user's intent. In the current system interface, there is a lack of design support to help users' prompt writing process, and we found that users frequently face limitations in conveying their full intent in Sec 6.1.1. To address these challenges and facilitate a better representation of the user's intent, it is necessary to have a design that helps users refine their prompts to align them more precisely with their intent. This design could involve tokenizing user prompts and using this as a basis to offer keyword-specific suggestions. For example, if a user writes a prompt, "Explain recent issues related to Autonomous Vehicles (AVs) in simple terms.", keywords can tokenize the prompt, and the following keyword-specific suggestions can be provided: the types of AVs, the time frame for recent, the types of issues (e.g., ethical), and the appropriated level of simplicity for the terms used. Moreover, considering dissatisfaction arising from extensive and detailed responses ($\mathbf{D_{format}}$), giving suggestions utilizing multi-modality, such as image and video, could enable a better user experience when they can succinctly represent the users' intent. This allows users to refine their prompts by selecting the suggestions, ensuring a more accurate alignment with their intent. Providing users with a range of suggestions and enabling them to select suggestions by reflecting their intent can empower users to express their intent effectively.

### 6.2.2 Recommending effective multi-turn prompt tactics to users

To enhance user satisfaction during multi-turn interactions with LLM, we suggest a design that recommends effective prompt tactics to users during the conversation. Our public dataset could be utilized for this process since it contains various prompt tactics (Table 5.2) and their effectiveness reported by users to address their dissatisfaction. For instance, an interaction can be envisioned where the LLM predicts the probability of user dissatisfaction with a generated response. If the probability is high, the system can proactively guide users to employ some effective tactics in their subsequent prompt to address the anticipated dissatisfaction.

We also recommend evolving this design to incorporate effective prompt engineering techniques suitable for multi-turn interactions, such as Chain-of-Thought (CoT) [41]. While a thread of research has addressed effective prompt engineering techniques to get desired responses from LLMs, they usually focus on crafting one prompt. Moreover, there is a lack of research on prompt engineering techniques tailored to address or mitigate user dissatisfaction during conversations. By integrating our data-driven insights on users' effective prompt tactics with prompt engineering techniques, we propose that recommending tactics to users during multi-turn interactions will yield more favorable responses, enhancing their overall satisfaction.

### 6.2.3 Providing personalized LLM experience

We suggest the need for a design that provides personalized LLM experiences based on our finding that there exist differences in dissatisfaction and corresponding tactics depending on the user's level of knowledge about LLMs. One of the possible designs for personalized LLM experiences is to adjust the refusal policies or attitudes that LLM refuses to answer according to the user's knowledge levels about LLM. This is because our results show that the low-knowledge group experienced more dissatisfaction with ChatGPT's refusal to answer ($\mathbf{D_{refuse}}$) than the high-knowledge group. This may be because the low-knowledge group tends to ask more questions that were limited for ChatGPT to answer without fully understanding ChatGPT's capabilities. Thus, rather than responding with a generic "As a language model, I am not capable of..." a more direct explanation addressing its limitations to better inform users of its capability may be required for low-knowledge users.

To facilitate personalized LLM experiences, we emphasize the need for user modeling based on prior sessions where LLM can gain information about the user's state before chatting. The user's state encompasses not only their knowledge level about LLM but also their usage purpose, specific task at hand, the language or proficiency level they used for chatting, and more. Such sessions serve to shape the user's mental model of LLM and vice versa, fostering a mutual understanding. Through this approach, users can benefit from customized interactions that consider their individual circumstances, ultimately improving their overall LLM experience.

# Chapter 7. Limitations and Future Work

We present the limitations of our work and possible future work.

First, our analysis on user-side dissatisfaction and tactics was based on ChatGPT user data. Although ChatGPT is one of the most widely used LLMs, it is important to see whether our results apply to other LLMs. There may be some differences in how users undergo dissatisfaction. For instance, specific wordings used when LLM refuses to answer can affect how much users feel dissatisfaction regarding $\mathbf{D_{refuse}}$. Moreover, since $\mathbf{D_{acc}}$ is a category that is directly related to the performance of LLMs, users may face different levels of dissatisfaction for $\mathbf{D_{acc}}$.

Second, our analysis is based on self-reported data from users. We tried to ensure the quality of the data by careful filtering and pre-processing of the data while checking on the actual conversation log. However, dissatisfaction levels and tactic effectiveness are based on participants' self-reported scores, which may suffer from subjectiveness and heavily rely on the participant's memory. We also tried to eliminate this problem by only collecting conversation logs within 1 month, but the problem may still linger.

Lastly, we investigated the difference in user dissatisfaction and tactics according to the difference in knowledge level of LLMs. Future work can expand on our work and further examine whether the differences in dissatisfaction and tactics exist according to other dimensions. For instance, since LLMs are chat-based, there may exist differences between those different English proficiency. Moreover, since users may have different expectations according to tasks, there may exist differences when given different tasks. For instance, fact-oriented tasks, such as finding information or explaining a real-world fact, will have more relevance with $\mathbf{D_{acc}}$ since the user expects to get correct information. On the other hand, creative tasks, such as writing stories or scenarios, will have less relevance with $\mathbf{D_{acc}}$ but more relevance with $\mathbf{D_{intent}}$, since users will be interested in how well the LLM can understand their needed content or context of creating content to their situations.

# Chapter 8. Conclusion

In this study, with ChatGPT as the case study, we explored user-side dissatisfaction and corresponding tactics during the conversation with chat-based LLM. Through a systematic literature review, we identified seven categories of user-side dissatisfaction from LLM-generated responses. Then, we collected data from 107 users conversing with ChatGPT, and uncovered prevalent, severe, and unaddressed dissatisfactions. We also analyzed four users' tactic categories to address their dissatisfaction and their prevalence and effectiveness. We also investigated how these vary depending on the users' knowledge level of LLMs. Our findings provide insights into how LLM and its interface can be further developed to aid people when they encounter dissatisfaction. One potential is user-side prompt engineering techniques that can be utilized in the middle of the conversation when dissatisfaction occurs. The pair of dissatisfactions and corresponding tactics can guide this prompt engineering. In addition to these contributions, we have made a publicly accessible dataset available, containing actual user conversation data related to dissatisfaction. This research deepens the understanding of user dissatisfaction in LLM interactions, providing a foundational knowledge base for future enhancements that can benefit users across knowledge levels.

# Chapter 9. Appendix

## 9.1 Systematic Literature Review Paper List

All paper lists corresponding user-side dissatisfaction codes are in Table 9.1.

| **Category (7)** | **Code (19)** | **Example** |
|---|---|---|
| Intent Understanding (**D<sub>intent</sub>**) | C1. Response does not meet users' intent or instruction. | [49, 70, 30, 71, 6, 50] |
| | C2. Response is not aligned with the user's context. | [50, 30, 5, 25, 71, 56, 6, 72, 57, 21, 73, 49, 74, 70] |
| | C17. The tone or communication style is disappointing. | [56, 30, 5, 22, 21] |
| Content depth and originality (**D<sub>depth</sub>**) | C3. Response is too general. | [30, 29, 56, 6, 75, 52, 21] |
| | C4. Response lacks originality. | [51, 30, 5, 25, 56, 76, 21, 77, 68] |
| | C5. Response lacks information. | [6, 4, 30, 25, 56, 7, 50, 75, 52, 78, 21, 77, 73, 70] |
| Information Accuracy (**D<sub>acc</sub>**) | C6. The response contains incorrect information. | [22, 4, 18, 58, 30, 79, 5, 80, 54, 20, 25, 53, 81, 82, 6] [76, 83, 50, 75, 84, 85, 86, 57, 21, 77, 73, 68, 87, 88, 89] |
| | C7. Response is based on training data cut off at a certain date, and has limited access to newly created data. | [52, 24, 29, 5, 53, 56, 72, 73, 49, 68, 70] |
| | C8. Response is inconsistent. | [53, 18, 29, 19, 20, 25, 90, 81, 91, 83, 84, 87] |
| | C9. ChatGPT struggles with reasoning. | [54, 18, 29, 30, 16, 92, 53, 82, 7, 72, 75, 73, 68, 88] |
| | C10. (Hallucination) ChatGPT fabricates contents that conflict with the source content or cannot be verified from existing sources. | [17, 18, 24, 29, 58, 30, 16, 86, 93, 49, 87, 88, 89] |
| | C19. (Sycophancy) ChatGPT excessively conforms to the user. | [16, 18, 56, 55] |
| Transparency (**D<sub>trans</sub>**) | C11. It's difficult to understand the reasons, criteria, logic, and evidence behind the responses. | [24, 18, 5, 16, 19, 25, 53, 83, 50, 75, 21, 68, 87, 88] |
| Refusing to answer (**D<sub>refuse</sub>**) | C12. ChatGPT avoids giving its own opinion by saying something similar to "As a language model, I am not capable . . ." | [30, 56] |
| | C13. ChatGPT avoids talking about difficult or controversial issues by saying something similar to "As a language model, I am not capable ..." | [16, 56] |
| | C7. Response is based on training data cut off at a certain date, and has limited access to newly created data. | [52, 24, 29, 5, 53, 56, 72, 73, 49, 68, 70] |
| Content ethics and integrity (**D<sub>ethic</sub>**) | C14. Response contains unlawful content | [18, 91] |
| | C15. Response contains unethical, harmful content. | [57, 18, 24, 5, 94, 91, 75, 85, 86, 95, 57, 74, 89] |
| | C16. Response contains biased content. | [58, 4, 18, 24, 30, 5, 96, 75, 85, 97, 98, 99, 86, 100, 101, 21, 73, 87, 88] |
| Response Format and Attitude (**D<sub>format</sub>**) | C17. The tone or communication style is disappointing. | [56, 30, 5, 22, 21] |
| | C18. Response is overly detailed or too long | [23, 5, 25, 56, 50] |
| | C19. (Sycophancy) ChatGPT excessively conforms to the user. | [16, 18, 56, 55] |

Table 9.1: 7 category and corresponding 19 codes of user-side dissatisfaction from LLM Responses.

## 9.2 Data Filtering Criteria and Detailed Reason

### 9.2.1 Conversation-Level Filtering

***Conversation older than 30 days.*** We collected real-world experience data from individuals, which inherently consists of past data they have encountered. Therefore, in order to encourage respondents to recall these past experiences while responding to our data collection system, we restricted the chat dates to "previous 30 days" from the survey date. Although the survey included explicit instructions regarding this matter, we identified four cases where participants reported chat dates older than 30 days, and excluded them.

***Conversation with a memory level of 3 or lower.*** Even if a conversation occurred within the previous 30 days, it was considered unreliable if the user had a low memory level regarding the conversation. Therefore, conversations where the user's memory level was rated 3 or lower on a 7-point scale were filtered out. This criterion led to the exclusion of five conversations.

***Conversation for fun or testing purposes.*** Our research focused on real-world experiences related to dissatisfaction encountered while using LLMs for practical purposes. Therefore, we do not delve into scenarios where users intentionally provoke dissatisfactory responses from LLMs, attempting to manipulate the model's behavior through techniques like jailbreaking [102, 103], using LLM solely for fun or testing. Despite the explicit instructions regarding this in the data collection system, seven conversations were identified as falling into this category and were filtered out.

***Conversation from versions other than GPT-3.5.*** Considering the significant differences in performance between GPT-3.5 and GPT-4 [89], we also considered the GPT version used in the conversation. Four conversations used GPT-4, while all others used GPT-3.5. To maintain data consistency, we filtered out the four conversations that used GPT-4.

### 9.2.2 Response-Level Filtering

***Dissatisfaction due to ChatGPT's error messages*** Dissatisfaction caused by ChatGPT responses being interrupted or encountering errors was not our research scope. Three responses fell under this category.

***Unconvincing dissatisfaction*** Seven cases were identified where it was challenging to understand why the user was dissatisfied when reviewing both the ChatGPT conversation and the user's dissatisfaction reasons.

***Mismatch between score and reason*** In one case, the effectiveness score for resolving dissatisfaction was 1 (indicating not effective), but the reason for that score was reported that the dissatisfaction was resolved by the prompt. This mismatch led to the exclusion of this case.

***No Correlation between selected dissatisfactions and subsequent prompts for resolving that dissatisfaction*** In five cases, we observed a lack of correlation between selected dissatisfaction categories and selected subsequent prompts to address such dissatisfaction. For example, it was the case a prompt that had nothing to do with the selected dissatisfaction was chosen to resolve the dissatisfaction.

# Bibliography


[1] OpenAI. Gpt-4 technical report, 2023.

[2] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

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

[4] Enkelejda Kasneci, Kathrin Seßler, Stefan K¨uchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan G¨unnemann, Eyke H¨ullermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[5] Malik Sallam. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887. MDPI, 2023.

[6] Arun HS Kumar. Analysis of chatgpt tool to assess the potential of its utility for academic writing in biomedical domain. *Biology, Engineering, Medicine and Science Reports*, 9(1):24–30, 2023.

[7] Douglas L Mann. Artificial intelligence discusses the role of artificial intelligence in translational medicine: a jacc: basic to translational science interview with chatgpt. *Basic to Translational Science*, 8(2):221–223, 2023.

[8] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can gpt-3 perform statutory reasoning? *arXiv preprint arXiv:2302.06100*, 2023.

[9] John J Nay. Law informs code: A legal informatics approach to aligning artificial intelligence with humans. *Nw. J. Tech. & Intell. Prop.*, 20:309, 2022.

[10] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

[11] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. How to prompt? opportunities and challenges of zero-and few-shot learning for human-ai interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*, 2022.

[12] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 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*, pages 1–21, 2023.

[13] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[14] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

[15] Chatgpt masterclass: The guide to ai & prompt engineering udemy. https://www.udemy.com/course/chatgpt-ai-masterclass/, Accessed on 10/08/2023.

[16] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023.

[17] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[18] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment, 2023.

[19] Myeongjun Jang and Thomas Lukasiewicz. Consistency analysis of chatgpt. *arXiv preprint arXiv:2303.06273*, 2023.

[20] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Sch¨utze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.

[21] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 2023.

[22] Amos Azaria. ChatGPT Usage and Limitations. working paper or preprint, December 2022.

[23] Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. Exploring the limits of chatgpt for query or aspect-based text summarization, 2023.

[24] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, pages 1–11, 2023.

[25] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[26] Ewa Luger and Abigail Sellen. Like having a really bad pa" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5286–5297, 2016.

[27] Jing Qian, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. Limitations of language models in arithmetic and symbolic induction. *arXiv preprint arXiv:2208.05051*, 2022.

[28] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*, 2023.

[29] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

[30] Ali Borji. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*, 2023.

[31] Morteza Behrooz, William Ngan, Joshua Lane, Giuliano Morse, Benjamin Babcock, Kurt Shuster, Mojtaba Komeili, Moya Chen, Melanie Kambadur, Y-Lan Boureau, et al. The hci aspects of public deployment of research chatbots: A user study, design recommendations, and open challenges. *arXiv preprint arXiv:2306.04765*, 2023.

[32] Sarah Brown-Schmidt, Si On Yoon, and Rachel Anna Ryskin. People as contexts in conversation. In *Psychology of learning and motivation*, volume 62, pages 59–99. Elsevier, 2015.

[33] Teun A Van Dijk. *Comments on context and conversation.* Citeseer, 2007.

[34] Thomas M Holtgraves and Yoshihisa Kashima. Language, meaning, and social cognition. *Personality and Social Psychology Review*, 12(1):73–94, 2008.

[35] Marita Skjuve, Ida Maria Haugstveit, Asbjørn Følstad, and Petter Brandtzaeg. Help! is my chatbot falling into the uncanny valley? an empirical study of user experience in human–chatbot interaction. *Human Technology*, 15(1):30–54, 2019.

[36] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[37] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–7, 2018.

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[39] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

[40] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[41] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

[42] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[43] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. Language models of code are few-shot commonsense learners. *arXiv preprint arXiv:2210.07128*, 2022.

[44] Manolis Remountakis, Konstantinos I. Kotis, Babis Kourtzis, and George E. Tsekouras. Chatgpt and persuasive technologies for the management and delivery of personalized recommendations in hotel hospitality. *ArXiv*, abs/2307.14298, 2023.

[45] Muh. Erwinto Imran and Norah Mansour Almusharraf. Analyzing the role of chatgpt as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, 2023.

[46] Md. Mostafizer Rahman and Yutaka Watanobe. Chatgpt for education and research: Opportunities, threats, and strategies. *Applied Sciences*, 2023.

[47] Ishika Joshi, Ritvik Budhiraja, Harshal Dev, Jahnvi Kadia, M. Osama Ataullah, Sayan Mitra, Dhruv Kumar, and Harshal D. Akolekar. Chatgpt in the classroom: An analysis of its strengths and weaknesses for solving undergraduate computer science questions. 2023.

[48] Maanak Gupta, Charankumar Akiri, Kshitiz Aryal, Elisabeth Parker, and Lopamudra Praharaj. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access*, 11:80218–80245, 2023.

[49] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models, 2023.

[50] A Rao, J Kim, M Kamineni, M Pang, W Lie, and MD Succi. Evaluating chatgpt as an adjunct for radiologic decision-making. medrxiv, 2023-02, 2023.

[51] Felipe C Kitamura. Chatgpt is shaping the future of medical writing but still requires human judgment, 2023.

[52] Yee Hui Yeo, Jamil S Samaan, Wee Han Ng, Peng-Sheng Ting, Hirsh Trivedi, Aarshi Vipani, Walid Ayoub, Ju Dong Yang, Omer Liran, Brennan Spiegel, et al. Assessing the performance of chatgpt in answering questions regarding cirrhosis and hepatocellular carcinoma. *medRxiv*, pages 2023–02, 2023.

[53] Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*, 2022.

[54] Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, Gyeong-Moon Park, Sung-Ho Bae, Lik-Hang Lee, Pan Hui, In So Kweon, and Choong Seon Hong. One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era, 2023.

[55] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland,

Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022.

[56] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023.

[57] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[58] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023.

[59] Anselm Strauss and Juliet Corbin. Basics of qualitative research techniques. 1998.

[60] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[61] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[62] Jian Hu, Gang Wang, Fred Lochovsky, Jian-tao Sun, and Zheng Chen. Understanding user's query intent with wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, page 471–480, New York, NY, USA, 2009. Association for Computing Machinery.

[63] Ziv Bar-Yossef and Naama Kraus. Context-sensitive query auto-completion. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 107–116, New York, NY, USA, 2011. Association for Computing Machinery.

[64] Soo Young Rieh and Hong (Iris) Xie. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management*, 42(3):751–768, 2006.

[65] Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. Cognitive mirage: A review of hallucinations in large language models. *ArXiv*, abs/2309.06794, 2023.

[66] Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. Becel: Benchmark for consistency evaluation of language models. In *International Conference on Computational Linguistics*, 2022.

[67] Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. Why does chatgpt fall short in providing truthful answers? 2023.

[68] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642, 2023.

[69] Chatgpt is a new ai chatbot that can answer questions and write essays. https://www.cnbc.com/2022/12/13/chatgpt-is-a-new-ai-chatbot-that-can-answer-questions-and-write-essays.html, Accessed on 10/08/2023.

[70] Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38, dec 2022.

[71] Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. Measuring inductive biases of in-context learning with underspecified demonstrations. *arXiv preprint arXiv:2305.13299*, 2023.

[72] Rehan Ahmed Khan, Masood Jawaid, Aymen Rehan Khan, and Madiha Sajjad. Chatgpt-reshaping medical education and clinical management. *Pakistan Journal of Medical Sciences*, 39(2):605, 2023.

[73] Mohammadreza Farrokhnia, Seyyed Kazem Banihashem, Omid Noroozi, and Arjen Wals. A swot analysis of chatgpt: Implications for educational practice and research. *Innovations in Education and Teaching International*, pages 1–15, 2023.

[74] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292, 2022.

[75] Gaurav Sharma and Abhishek Thakur. Chatgpt in drug discovery. 2023.

[76] H Holden Thorp. Chatgpt is fun, but not an author, 2023.

[77] Augustin Lecler, Loïc Duron, and Philippe Soyer. Revolutionizing radiology with gpt-based models: Current applications, future possibilities and limitations of chatgpt. *Diagnostic and Interventional Imaging*, 104(6):269–274, 2023.

[78] Saima Nisar and Muhammad Shahzad Aslam. Is chatgpt a good tool for t&cm students in studying pharmacology? *Available at SSRN 4324310*, 2023.

[79] Junaid Qadir. Engineering education in the era of chatgpt: Promise and pitfalls of generative ai for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–9. IEEE, 2023.

[80] Luciano Floridi. Ai as agency without intelligence: on chatgpt, large language models, and other generative models. *Philosophy & Technology*, 36(1):15, 2023.

[81] Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*, 2023.

[82] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Sch"olkopf. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*, 2023.

[83] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. Can chatgpt write a good boolean query for systematic review literature search? *arXiv preprint arXiv:2302.03495*, 2023.

[84] Dat Duong and Benjamin D Solomon. Analysis of large-language model versus human performance for genetics questions. *European Journal of Human Genetics*, pages 1–3, 2023.

[85] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

[86] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, 2022.

[87] Sukhpal Singh Gill and Rupinder Kaur. Chatgpt: Vision and challenges. *Internet of Things and Cyber-Physical Systems*, 3:262–271, 2023.

[88] Muhammad Usman Hadi, R Qureshi, A Shah, M Irfan, A Zafar, MB Shaikh, N Akhtar, J Wu, and S Mirjalili. A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*, 2023.

[89] gpt-4-system-card.pdf. https://cdn.openai.com/papers/gpt-4-system-card.pdf, Accessed on 10/08/2023.

[90] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023.

[91] Andreas Holzinger, Katharina Keiblinger, Petr Holub, Kurt Zatloukal, and Heimo M"uller. Ai for life: Trends in artificial intelligence for biotechnology. *New Biotechnology*, 74:16–24, 2023.

[92] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.

[93] Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C. Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. Opportunities and challenges for chatgpt and large language models in biomedicine and health, 2023.

[94] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[95] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

[96] Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval. Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmology Science*, page 100324, 2023.

[97] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, 2021.

[98] Robert Wolfe and Aylin Caliskan. American== white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 800–812, 2022.

[99] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. "" i wouldn't say offensive but..."": Disability-centered perspectives on large language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 205–216, 2023.

[100] Li Lucy and David Bamman. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, 2021.

[101] Roberto Navigli, Simone Conia, and Bj"orn Ross. Biases in large language models: Origins, inventory and discussion. *ACM Journal of Data and Information Quality.*

[102] Llm jailbreak study. https://sites.google.com/view/llm-jailbreak-study, Accessed on 10/06/2023.

[103] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study, 2023.

# Acknowledgment

Firstly, I want to express my gratitude to my advisor, Juho Kim, for his invaluable guidance, support, and encouragement throughout my research journey. His expertise and insightful feedback were instrumental in shaping my research and guiding me to grow as a researcher. I also extend my thanks to all members of KIXLAB. The collaborative environment they built enabled me to learn, grow, and overcome obstacles with the help of their diverse perspectives. Finally, I want to express my heartfelt gratitude to my family for their constant love, encouragement, and belief in me. I wouldn't be here without their unwavering presence throughout this journey.

# Curriculum Vitae in Korean

이　　름: 김 윤 수
생 년 월 일: 1998년 08월 16일
전 자 주 소: yoonsu16@kaist.ac.kr

## 학　　력

2014. 3. – 2017. 2.　거창고등학교
2017. 2. – 2022. 2.　포항공과대학교 컴퓨터공학과 (학사)
2022. 2. – 2024. 2.　한국과학기술원 김재철AI대학원 (석사)

## 연 구 업 적

1. **Yoonsu Kim**, Jueon Lee, Seoyoung Kim, Jaehyuk Park, Juho Kim, "Understanding Users' Dissatisfaction with ChatGPT Responses: Types, Resolving Tactics, and the Effect of Knowledge Level", *Proceedings of the ACM on Intelligent User Interfaces (IUI 2024)*, (to appear).
2. Sungjae Cho, **Yoonsu Kim**, Jaewoong Jang, Inseok Hwang, "AI-to-Human Actuation: AI Proactively Induces Humans towards Favorable Sensing Conditions", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT 2023)*
3. Dae Hyun Kim*, **Yoonsu Kim***, Hyungyu Shin, Jinho Son, Juho Kim (*Corresponding authors), "Towards Understanding the Challenges and Remedies in AI Application Development", *Korea Computer Congress 2023 (KCC 2023)*
4. Jieun Han, Haneul Yoo, **Yoonsu Kim**, Junho Myung, Minsun Kim, Hyunseung Lim, Juho Kim, Tak Yeon Lee, Hwajung Hong, So-Yeon Ahn, Alice Oh, "RECIPE: How to Integrate ChatGPT into EFL Writing Education", *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S 2023, WiP)*
5. Sungjae Cho, Jaewoong Jang, **Yoonsu Kim**, Inseok Hwang, "Demonstrating AHA: Boosting Unmodified AI's Robustness by Proactively Inducing Favorable Human Sensing Conditions", *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct)*