Influence of ChatGPT Expertise Topical Expertise and Topical Interest on User Behavior and Engagement in Informational Search Sessions in ChatGPT

T. Mo

TUDelft

Influence of ChatGPT Expertise Topical Expertise and Topical Interest on User Behavior and Engagement in Informational Search Sessions in ChatGPT

by

T. Mo

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday December 11, 2023 at 3:00 PM.

Student number:	5283515	
Project duration:	April, 2023 – December, 2023	
Thesis committee:	Dr. Maria Soledad Pera,	TU Delft, Chair & Thesis Advisor
	Dr. Ujwal Gadiraju,	TU Delft, Thesis Co-Advisor
	Dr. Myrthe Tielman,	TU Delft, External Committee Member
	Alisa Rieger	TU Delft, Daily Supervisor

This thesis is confidential and cannot be made public until December 11, 2023.

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.
Cover image: Generated by ChatGPT

Preface

As my time as a student at TU Delft draws to a close, I am filled with a profound sense of gratitude. The journey of accomplishing this thesis was challenging yet worthwhile. I am really appreciative of Dr. Maria Soledad Pera and Dr. Ujwal Gadiraju for their great advice and guidance. Their insights shape my research direction and content. Sole provided me with valuable insights regarding the writing and research research questions. Ujwal provided me with a wealth of knowledge regarding the experimental setup and analysis. Also special thanks to Alisa Rieger, Alisa gave me invaluable assistance while I was struggling with the research. Finally, I would like to express my gratitude to Dr. Myrthe Tielman for being the committee member for my thesis. Additionally, support from my parents and my friends have been my backbone throughout this journey and their encouragement and patience are the source of my motivation and strength.

T. Mo
Delft, December 2023

Abstract

ChatGPT, a cutting-edge technology based on LLM, demonstrated great potential in search tasks. While the importance and potential of ChatGPT are growing, the gap in the understanding of how users interact and engage in ChatGPT search remains open. Past research has extensively examined traditional information search, but there is a need for investigation into user behaviour and engagement in LLM contexts like ChatGPT. To address this gap, our study aims to examine the impact of ChatGPT expertise, topical expertise, and topical interest on user behaviour and user engagement in ChatGPT search as we assume these factors play an important role in shaping user interactions with ChatGPT. We conducted an experiment to investigate the answer by inviting users (N=198) via the crowdsourcing platform to communicate with the mock ChatGPT application and their interactions were recorded for subsequent analysis. Prior to and after their interaction with the mock ChatGPT application, users are requested to complete a questionnaire to gather information about their user profiles and quantify their engagement. Our finding indicates that ChatGPT expertise has a partial influence on user engagement in ChatGPT search, which may highlight the importance of AI expertise in shaping user interactions. Furthermore, our research also indicates that the Affinity for Technology Interaction (ATI) has an impact on user engagement, which underscores the importance of understanding the psychological aspects in the age of artificial intelligence. The results of this study will not only address the current knowledge gap in ChatGPT search but also provide valuable information on how to improve ChatGPT to enhance user interaction experience.

Contents

1 Introduction

Information searching is a popular activity on the Internet, which is also considered one of the most frequent online activities [22]. Furthermore, it is also related to the learning activity. Due to the importance of learning in daily life, it is valuable to dive into the information search tasks. ChatGPT, a brand-new tool based on the Large language model (LLM), showed excellent performance on language-related tasks. Recently, ChatGPT was also tested on searching tasks and it showed outstanding performance than supervised methods on popular IR benchmarks [57]. However, even if its importance and potential as a search tool are growing, there is still a gap in the understanding of how users interact and engage with this novel technology. Traditional information search has been extensively studied, but the influence of factors on user behaviour and engagement in LLM contexts like ChatGPT needs to be thoroughly investigated.

Xu et al. [65] analyzed user behaviour across Google and ChatGPT and they found that search tasks were completed more rapidly on ChatGPT than in Google. Similarly, Spatharioti et al. [56] compared the user behaviour of traditional search engines versus LLM-based search tools and they also found that the LLM-based search tool led to quicker task completion. Additionally, the prompts from LLM users are shorter and more complicated prompts. However, even if these studies provide a comprehensive investigation of user behaviour in search sessions in ChatGPT, the research regarding what factors may influence user behaviour and engagement in search sessions in ChatGPT remains open. This research gap is especially important given the potential of ChatGPT as a search tool.

To fill this need, our study aims to investigate how ChatGPT expertise, topical expertise, and topical interest influence user behaviour and user engagement in ChatGPT search as we assume these factors play an important role in shaping user interactions with ChatGPT. We define ChatGPT expertise as knowledge and interest about the functionality, mechanism, strengths, and limitations of the ChatGPT. Research [68, 52] highlight the importance of prompt formulation in the ChatGPT and they found if the prompts are well-formulated then the answer accuracy will be increased. Otherwise, ChatGPT may even generate unexpected answers. To be able to formulate good prompts, the users are expected to be able to have sufficient knowledge about ChatGPT. For instance, the users are expected to know prompt engineering [59] and knowledge about limitation ChatGPT have such as bias [50], illusion [12, 50, 64]. As mentioned earlier, ChatGPT-related knowledge facilitates the use of ChatGPT. Therefore, it is intriguing to investigate the impact of ChatGPT expertise on user behaviour and engagement in ChatGPT search.

Additionally, we explored the role of topical expertise on user behaviour and engagement, which is defined as ”Knowledge of the topic of the information need” [60] and research indicates that users' level of topical expertise influences their search behaviour. For instance, [38, 61] found that the users in different topical expertise levels behave differently in web searches. White et al. [61] found that users with higher topical expertise spent a longer time in search sessions and used longer queries in web searches. And there is also a similar finding from [23]. Previous research showed [47] the reason why users with different topical expertise levels is that high topical expertise users are more likely to provide more controls in the search.

Our study also considers the impact of topical interest. We assume that topical interest could also play an important role in the behaviour and engagement in the informational search sessions since sometimes users have topical expertise not because of the interest. Edward et al. [15] found different levels of topical interest lead to different levels of user engagement in Interactive information retrieval (IIR). Fox et al. [19] found there users with higher topical interest will also present higher page activity in the meanwhile. Furthermore, [53] intrinsic motivation could be the reason why users with different topical interest levels behave differently.

In light of this background, we propose the following research questions:

RQ1: Do ChatGPT expertise, topical expertise, and topical interest influence user behaviour in informational search sessions in ChatGPT?

- **H1a:** People with relatively higher ChatGPT expertise behave differently compared to people with relatively lower ChatGPT expertise
- **H1b:** People with relatively higher topical expertise behave differently compared to people with relatively lower topical expertise
- **H1c:** People with relatively higher topical interest behave differently compared to people with relatively lower topical interest

RQ2: Do ChatGPT expertise, topical expertise, and topical interest influence user engagement in informational search sessions in ChatGPT?

- **H2a:** People with relatively higher ChatGPT expertise exhibit different levels of engagement compared to people with relatively lower ChatGPT expertise
- **H2b:** People with relatively higher topical expertise exhibit different levels of engagement compared to people with relatively lower topical expertise
- **H2c:** People with relatively higher topical interest exhibit different levels of engagement compared to people with relatively lower topical interest

To address the research questions, we designed an experiment wherein participants were assigned a topic from the dataset at random. Subsequently, we will request them to complete a questionnaire to assess their ChatGPT expertise, topical expertise, and interest in that topic. Afterwards, the participants will be directed to use the mock ChatGPT application to collect information regarding the allocated topic. Their usage behaviours within the mock ChatGPT, such as prompt information and session length, will be documented during the experiment. And the end of the experiment, we will ask participants to answer a questionnaire to measure their engagement during the task and also their Affinity for Technology (ATI). With a series of structured tasks in the experiment and subsequent analyses, we aim to explore how different user profiles lead to diverse interaction patterns. The findings of this research will not only fill the current knowledge gap but also yield insights into optimizing LLM-based search tools for enhancing user interaction.

1.1. Contribution

- Findings of the influence of factors (ChatGPT expertise, Topical expertise, and topical interest) on the user behaviour and engagement in search sessions in ChatGPT.
- A data set containing the participant's interactions with ChatGPT-based ChatBot and their profile information.
- Mock ChatGPT source code and required analysis script.

1.2. Outline

This thesis follows the structure: In Chapter 1 we present the motivation and contribution of this research. In Chapter 2 we provide the related research including existing research about the influence of the factors on the user behaviour and engagement in web search. In Chapter 3 we elaborate on how the experiment is designed for reaching the research goal. In Chapter 4 we cover quality control, how the data is processed, and how the ChatGPT expertise, topical expertise, and topical interest of users are classified into different levels. In Chapter 5 we share the statistical test results, and we also present our findings about how other interesting factors may affect user behaviour and engagement like ATI, usage frequency, and usage experience of chatbots. In Chapter 6 we discuss our findings and compare our results with existing research. In Chapter 7, the last chapter, we draw a conclusion based on our findings and explorations and share potential research directions in the future.

2 Related work

This chapter covers the literature on the topic of search. Initially, we present the development of user behaviour and engagement in the search field. Next, we depict the development of research regarding expertise, and topical interest in this area.

2.1. User behaviour in informational search sessions

Numerous studies have demonstrated that understanding user behaviour in search contexts plays an important role in enhancing search support and efficiency [25, 38, 22, 11]. As an example, Hölscher et al. [25] how web expertise influences user behaviour in the web search and how they found the web experts are more likely to use query formatting in the search as compared to non-web experts. Mao et al. [38] investigated the influence of topical expertise on user behaviour in web search and they found users with topical expertise tend to spend less time on search tasks. However, even if it is important to figure out user behaviour in the search, the research regarding behaviour in the ChatGPT context is still limited. Hence, one of our goals is to figure out how users with different features behave in the ChatGPT search. Metrics such as query length, number of inquiries, and session duration are often used to gauge the user behavior [22, 25]. Our research employs similar metrics in the experiment, facilitating comparisons with the findings of other studies and enabling additional exploration.

2.2. User engagement in informational search sessions

This research also examined how user engagement may be influenced in ChatGPT search sessions. Lalmas et al. [34] pointed out user engagement is one of the important metrics for the success of online service. Additionally, Hwang et al. [26] also conducted a meta-analysis to investigate the role of user engagement in system success and they found a correlation between user engagement and system success. Furthermore, Masrek et al. [40] investigate the relationship between satisfaction and user engagement in the web and digital library environment and found that user engagement is a strong predictor of user engagement.

Currently, there are various methods for quantifying user engagement. Lalmas et al. [35] proposed to use of mouse movement to detect user engagement. However, due to the difference between the SERP page and ChatGPT, it is not easy to adapt the existing method to the ChatGPT interface. Due to this feature of ChatGPT, we used the user engagement scale short form (UES-SF) proposed by [46], which is widely used for user engagement measurement. And it measures user engagement from four different dimensions namely Aesthetic Appeal (AE), Perceived Usability (PU), Focused Attention (FA), and Reward (RW). For this questionnaire, we also introduce the details about it, which can be found in Chapter 3. Zhuang et al. [66] investigated how to leverage user behaviour to predict user engagement as it is possible that the questionnaire is obtrusive in the experiments. In this research, they explored 37 different user behaviour features like the number of queries, task length, and number of numbers and investigated their relationship with the dimensions of user engagement mentioned before. And they found that the query-related features are most suitable for predicting Perceived Usability. The time-related features and query-related features performed best in user engagement as compared to other types of features. However, due to the difference between web search and ChatGPT search, the relationship between implicit features and user engagement is still an open question in the ChatGPT contexts. In addition, eye tracking is often utilised for measuring user engagement [6, 4, 31]. Nevertheless, the eye-tracking measurements may result in a reduction in the sample size due to increased expenses for employment and time allocation. After comparing different methods, we would like to propose using the user engagement scale short form (UES-SF) to measure user engagement in our experiment.

2.3. ChatGPT

ChatGPT is a nascent application built on LLM technology. Even though there are also available LLM-based applications like Bart from Google. However, due to the popularity of ChatGPT, we chose ChatGPT as our research object to look into user behaviour and engagement while searching for the LLM-based application. Sun et al. [57] investigated the search performance of ChatGPT and GPT-4 on popular information retrieval (IR) benchmarks in relevant ranking tasks and the experiment results indicated that these two LLM models showed better performance than the popular supervised model on these benchmarks. Similarly, Askari et al. [7] explored the capability of training data generation of LLM for cross-encoder re-rankers. They found that the response is more effective than the data generated by humans while evaluating the popular benchmarks and this result shows the strong potential of LLM to be used for training data generation in the web search field. Spatharioti et al. [56] compared the user behaviour in the traditional search and LLM-based model search and they found that the LLM-based searching tool users finished tasks in a shorter time as compared to the users using a traditional search engine. Furthermore, the prompts from LLM users are also shorter and more complicated. Chen et al. [11] experimented to monitor the performance of ChatGPT-3.5 and ChatGPT-4 in four different tasks namely math problems, sensitive question answering, code generation, and visual reasoning. And they the performance of ChatGPTs differs over time. For instance, ChatGPT-4 performed well on math problem tasks with an accuracy of 97.6% on March 2023. However, the accuracy dropped to 2.4% on June 2023. Due to this property of the ChatGPT, we also specify what version of ChatGPT was utilized in Chapter3 and it may facilitate other researchers to conduct evaluation results comparison in the future.

2.4. Expertise

As discussed in Chapter 1, our study examines the impact of ChatGPT expertise and topical expertise. In this section, we would like to introduce some related expertise in this area.

2.4.1. User expertise in ChatGPT

Current research regarding the influence of expertise on user behaviour in ChatGPT is still limited. In this case, we looked into research about the user behaviour in the web search as what we are focusing on is the search behaviour and web search is one of the most popular search ways in real life [8]. After looking into the search behaviour in the web search. We found there are different kinds of search behaviours. Abhishek et al. [32] looked into user behaviour in the conversational search and they used four different metrics to measure search behaviour namely No. of interactions, Average time per search, No. of documents per interaction, No. of Search tasks. They classified the behaviour of participants into four different categories based on the number of queries and number of opened documents in the search session. Liu et al. [36] compared the search behaviours between conversational search and traditional search and the number of queries, the number of cases, task time (s), dwell time per case (s) are used to measure the search behaviour. Schneider et al. [54] looked into the Search Behavior in the conversational search for the domain exploration and used in this research. However, not all the behavioural measurements are suitable for searching in the ChatGPT as the ChatGPT does not behave similarly to the search engine. In this case, we tried to adapt the existing behavioural measurement into the ChatGPT.

2.4.2. Topical expertise

White et al. [60] looked into how topical expertise may affect the interaction in the web search. They use log information from the internet and classify the users into expertise/non-expertise based on the website they visit. For instance, if a user visits medical-relevant website frequently then they will be considered as having medical expertise. They investigated four different topics and they found that topical expertise will send more queries and spend more time on the tasks. They think it is the information being sought is more important for the experts so they put more effort into the search. Mao et al. [38] conducted research on how the search interaction is affected by topical expertise and they found there is a relationship between topical expertise and search performance. They found that the topical expertise may spend less time on searching and have shorter queries. This is because participants with relevant topical expertise are able to conduct a search with less effort due to the knowledge. However, this finding contradicts the results reported by [60] and they assume that is because of different experimental setups because the experiment of [60] is based on the existing daily activity log information on the Internet and there was no specific tasks assigned and the users may put more effort to ensure information is well-round. As compared to [38]'s experiment, participants have specific tasks then they may tend to stop searching while the information is enough for answering the questions. Freund et al. [21] looked into the difference in web search behaviour between expertise and newbie and they found that the expertise may have longer queries. Hembrooke et al. [23] examined the impact of domain expertise on the choice of keywords in the web search. It was shown that domain expertise may send longer queries and more complicated queries. Vakkari et al. [58] conducted a study on the change of queries over the knowledge gained in web search and they found the query gets more complex as they gained more knowledge about this field. OB̀rien et al. [45] conducted a study on the influence of interest on user engagement and the research shows that there is a correlation between user engagement and topical interest. However, the reasons why topical interest could correlated with user engagement are not mentioned in the paper. Kelly et al. [33] examined the relationship between familiarity and the information search behaviour and they found that as the theme familiarity increases, the reading time will decrease. Xu et al. [65] investigated the difference in user performance while using ChatGPT and Google. The participant tends to spend less time while conducting the search in ChatGPT generating wrong information but it significantly enhances the performance of users disregarding the educational level. In this case, it is worth investing in the ChatGPT model due to its great potential.

As topical expertise can affect user behaviour and engagement by offering more control in search, we would like to look into how the topical expertise affects them in the ChatGPT search session. Furthermore, we also summarise the result of different user behaviour metrics as shown in the Table 2.1 and there is no existing summation regarding topical expertise in the existing literature. With this table, the researchers may be able to compare their results with existing works in terms of topical expertise.

	Query				Session	
Authors		Number	Length	Complexity		Length
Duggan et al. [14]			↓			↓
White et al. [60]		↑				↑
Mao et al. [38]		↓				↓
Freund et al. [21]			↑			
Hembrooke et al. [23]			↑	↑		
Vakkari et al. [58]				↑		
Kelly et al. [33]						↓

Table 2.1: Research regarding influence of topical expertise in search and corresponding results of user behaviour metrics

2.5. Topical Interest

In our research, we investigate how topical interest may influence user behaviour and engagement in the ChatGPT search. Research [24] indicates that topical interest is considered an important motiva-

Type	Topical interest modelling	Paper
Implicit feature	Click history record	[48, 2, 41]
Implicit feature	Datasets with topical interest information	[1]
Self-reported	Self-reported questionnaire	[45]
Implicit feature	Topic ranking	[15]
Implicit feature	Reading time	[44]
Implicit feature	Clickthrough rate	[29]
Implicit feature	Search queries, Rank of clicked documents	[3]
Implicit feature	Annotations	[27, 9]

Table 2.2: Measurement of topical interest in topical interest related papers

tional variable. Currently, motivation has been used to predict user behaviour in mobile applications [5]. In our research, we assume that users with different levels behave and engage in different levels. Besides, topical interest also plays an important role in web search modelling and has been used in several research [3, 43, 62].

Currently, Research investigating the relationship between topical interest and user behaviour and engagement in the search field is limited. O'Brien et al. [45] investigated how the topical interest, topical complexity, and user behaviour in search tasks could affect user engagement in web search. In this research, the topic interest is measured by a self-reported questionnaire and the participant is asked to fill in how interested they are in a series of topics based on a 5-likert scale the result showed that the topical interest will affect the user engagement in web search. Edward et al. [15] found that different levels of topical interest may lead to different levels of focused attention in user engagement. In this research, they investigate the relationship between user engagement and topical interest in Interactive information retrieval (IIR). By manipulating the topical interest, they found that user engagement will become higher. Furthermore, they also found that the participants show longer stroll times and longer query intervals, which indicate there are more interactions between high topical interest participants and the system. Besides, they also found that the participants showed lower heart rates and stronger electrodermal activity in the topic that they were interested in.

To investigate how to model the topical interest properly, we also look into different manners for topical interest modelling. Fox et al. [19] investigated whether the implicit metrics can measure user interest effectively since the explicit metrics like questionnaires sometimes be time-consuming or have some influence on the use pattern in the search. They found that there is a positive correlation between the results of implicit metrics and explicit metrics, which uncovered the possibility of using explicit metrics to measure topical interest. Qiu et al. [48] leveraged the user interest to personalize the search result. And the preference can be learned for the click-history data. In this research, they found that there is relation between user interest and search results. Based on this finding, they build a model to predict user interest based on the click history.

Furthermore, we also summarized the popular methods to model the topic interest in Table 2.2. Even if the advantages of using the implicit feature to measure the topical interest were discussed in the [19], in our research, we selected a 7-likert scale to measure the topical expertise since some of the existing implicit features are probably not capable of the accusatives in ChatGPT or are investment consuming. Besides, the topical interest questionnaire used in the experiment is also concise and it could probably reduce the influence to the users.

2.6. Usage Experience

The research regarding the influence of experience in ChatGPT or conversational agents is limited. However, due to the overlap between web search and conversational search, there is a possibility of gaining insight into this field by looking into the development of experience in traditional search. There is some research regarding the web experience in traditional search so far. Most works classified the novice users and experienced users according to how long they have used the laptop. For instance, Jenkins et al.[28] distinguish novice users and experienced users according to the experience of the computer. If the participants have used more than five years, then they will be classified as experienced users. However, Aula et al. [8] thought the experience is not a valid method to distinguish between the expert and non-expert as The level of expertise is not always influenced by experience and whether the users are able to use an efficient search method also played an important role. However, since the influence of usage experience is still a disrupted topic, the usage frequency is still used as an exploratory variable in our research.

2.7. Prompt formulation

There is a correlation between the prompt quality and answer quality in ChatGPT. Because of this, there are many researchers looking into the query formulation in ChatGPT. There is a popular term called prompt engineering in the LLM field and it is defined as ”the means by which LLMs are programmed via prompts” [59] and it plays an important role in the response quality. Zuccon et al. [68] looked into the influence of knowledge involved in the prompt on the correctness of the answers and they found if users provide prompts with wrong information then the correctness of the answer will be degraded because the ChatGPT may overturn the initial answers due to the knowledge in the prompt, which reflect the prompt knowledge can impact the response quality and reflect the importance role of prompt quality in the ChatGPT. White et al. [59] proposed to use a fixed pattern to frame the query and there are 16 patterns mentioned. It is likely to ensure the answers in good quality. However, it is not handy to remember so many different patterns and due to the rapid development of ChatGPT, the pattern could get outdated soon. Salle et al. [52] also found that ambiguous terms will have negative impacts on conversational search performance, which leads to worsening ranking performance. However, there are still no common definitions of good formation of prompts in ChatGPT since ChatGPT is also being enhanced over time. In our research, we will look into how the participants formulate the prompts since using different kinds of formulation ways could also contribute to a good search experience instead of sticking to the existing patterns.

3

Experimental Setup

As the literature introduced in the previous chapter, to understand how user behaviour and user engagement influence search in ChatGPT, we designed an experiment and, in this Chapter, we present the setup of the experiment. At first, we introduce the whole procedure of the experiment including the introduction regarding worker selection criteria, instruction, and questionnaires. Next, we introduce all variables measured in this experiment. In the end, we motivate the selection of the statistical tests and the decision on the sample size.

3.1. Experiment procedures

To investigate the user behaviour and user engagement in the ChatGPT search sessions, we designed an experiment and it procedures mainly comprise of five steps (worker selection, instruction, pre-task questionnaire, informational search session, post-task questionnaire) as shown in Figure 3.1. The details of each steps are provided as followed:

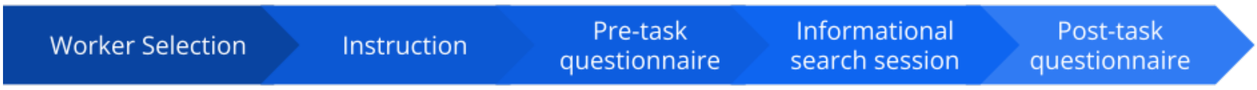


Figure 3.1: Data collection procedure

3.1.1. Worker selection

Worker selection are based on the criteria below:

- Speaking English fluently
- Acceptance rate above 90% in the Prolific
- Age above 18 years

Workers are expected to speak English fluently as the experiment will be conducted in English. By ensuring that the workers are less likely to misunderstand the contents of the experiment, and there is the probability to increase the credibility of the outcome. Furthermore, only workers who performed well were hired for the same purpose. Additionally, the participants must be older than 18 as required by law.

Regarding the worker salary, they were compensated by the UK minimum wage if their results are taken into account when the experiments are completed.

3.1.2. Instruction

Qualified workers was invited to participate in this experiment. After getting into the experiment, the first page to be shown is the consent form, which introduces the aim of the research and how the data will be processed after the experiment. Furthermore, participants can also find the contact emails from this page if they have any questions or concerns regarding this research.

3.1.3. Pre-task questionnaire

The subsequent step is to complete the pre-task questionnaire if agree with the content of the consent form, which aims to gather the following data:

Demographic information

To verify the representativeness of data, we collected data regarding the gender, age, and highest level of education completed and it can be realized by comparing their distributions to the distributions of populations.

ChatGPT expertise

Due to the limitation of existing research, we adapted the technical expertise questionnaire to capture the ChatGPT expertise [67]. Furthermore, to be able to obtain balanced data, we intend to control the number of data points of each ChatGPT expertise category to reach this goal. Specifically, when we have collected enough data points with specific ChatGPT expertise, the users classified as the same level of ChatGPT expertise were stopped to conduct the subsequent experiment and receive compensation.

Topical interest

In most research [30, 13, 45], the topical interest was gauged by asking the level of interest to a specific topic directly and we would like to follow the same way to measure the topical interest. Specifically, the topical interest is measured by the question ”How much interest do you have in the [Topic]” [30].

Topical expertise

In this research, we leveraged knowledge test designed by experts to measure the topical expertise [22]. There are 5 available topics in this research, which are listed in the Table 3.1. Each topic contains about 10 to 20 questions since content range of each topic differs intrinsically and, for each question in the knowledge test, participants are expected to select answer from ”Yes/No/I dont know” if they are not sure about what the answer is [22].

The topic selection follows this requirement below:

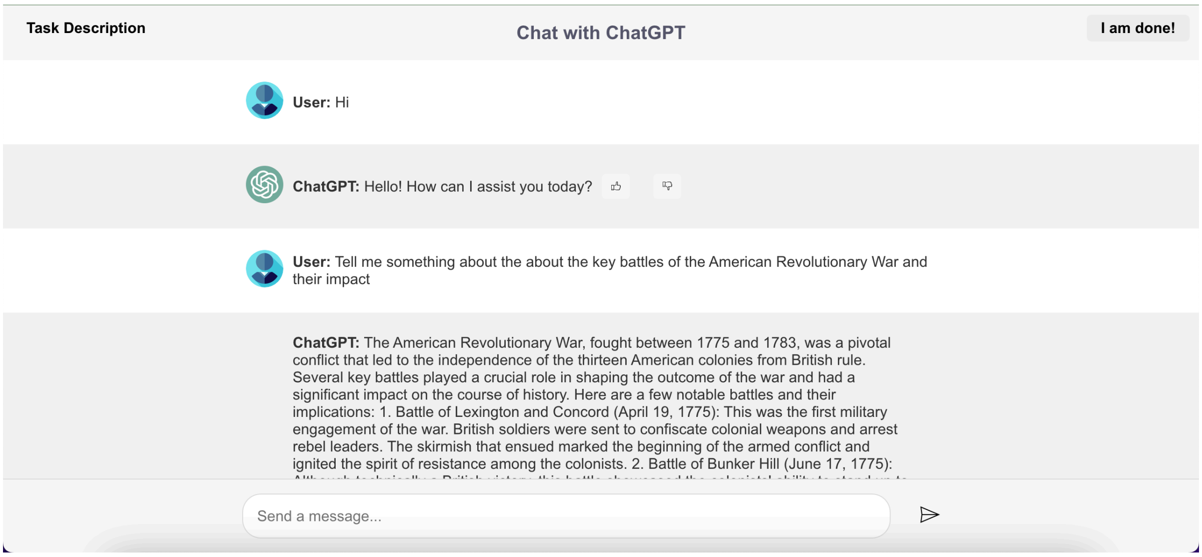
- Minimal overlap between topic and knowledge test
- Universally recognized topics

Regarding the first requirement, we would like to make sure the knowledge test does not impact the search pattern or the words they use during the interaction. In this case, we want the topic to have limited overlap between the topic and the knowledge test. For the second topic, if we use some regional topic like ’Orcas Island’, then it could be challenging to differentiate between high topical expertise and low topical expertise due to the topic popularity. Hence, topic selection is critical in differentiating users with different levels of topical expertise. For more information about topic, details can be found from Table 3.1.

3.1.4. Informational session

After completing the pre-task questionnaire, participants were directed to the task description. First, the worker will see the instruction once entering the task. Afterwards, they will be directed to the mock ChatGPT application. In this application, the workers are asked to collect information to satisfy the information need. If finishing the task earlier, then they can click the finish button at the upper right corner and the interface of this application is showed as Fig 3.2. During the main task, information regarding user behaviour will be collected.

Interface design
As our research goal is how to investigate how users behave in the ChatGPT, we try to make the interface look like the UI of the ChatGPT to flatten the learning curve. As shown in Fig 3.2, the user can input their prompts in the input area at the bottom of the interface. And the input is visible in real-time in the chatbox. Due to the limitation of the ChatGPT API, the response was be shown word by word in real-time as shown in the Figure 3.5. To mitigate the influence of the latency of the ChatGPT especially for the long response to ensure the users are less likely to lose concentration while interacting with the application. The response was shown as ”Generating answer...” so that the users can clearly know that the application is working instead of being crashed. Due to the importance of feedback for improving the performance of the system Once the answer is generated, the user can click the feedback button to share their idea for the response. If the user forgot what task they have, they can easily click the button on the top left corner. Then the description would be shown. Once the user was done with the searching, they clicked the button on the top right then they were directed to another page. To ensure that the user actually interacts with the application. We generated the unique verification code according to their prolific id and they were asked to input them when they jump back to the survey.



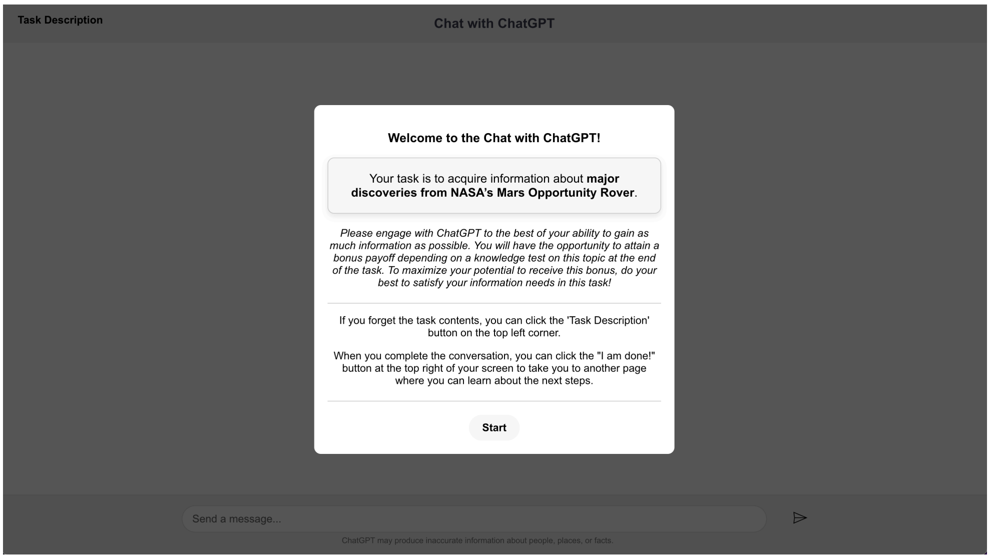
User: Hi

ChatGPT: Hello! How can I assist you today?

User: Tell me something about the about the key battles of the American Revolutionary War and their impact

ChatGPT: The American Revolutionary War, fought between 1775 and 1783, was a pivotal conflict that led to the independence of the thirteen American colonies from British rule. Several key battles played a crucial role in shaping the outcome of the war and had a significant impact on the course of history. Here are a few notable battles and their implications: 1. Battle of Lexington and Concord (April 19, 1775): This was the first military engagement of the war. British soldiers were sent to confiscate colonial weapons and arrest rebel leaders. The skirmish that ensued marked the beginning of the armed conflict and ignited the spirit of resistance among the colonists. 2. Battle of Bunker Hill (June 17, 1775):

Figure 3.2: Interface of the mock ChatGPT used in the experiment



Welcome to the Chat with ChatGPT!

Your task is to acquire information about **major discoveries from NASA’s Mars Opportunity Rover**.

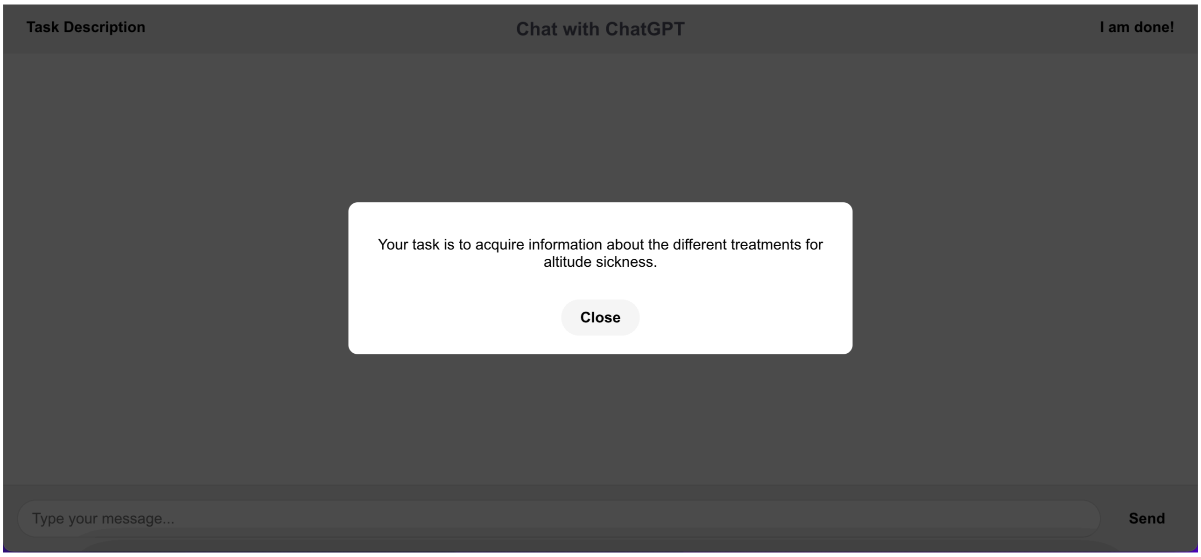
Please engage with ChatGPT to the best of your ability to gain as much information as possible. You will have the opportunity to attain a bonus payoff depending on a knowledge test on this topic at the end of the task. To maximize your potential to receive this bonus, do your best to satisfy your information needs in this task!

If you forget the task contents, you can click the 'Task Description' button on the top left corner.

When you complete the conversation, you can click the "I am done!" button at the top right of your screen to take you to another page where you can learn about the next steps.

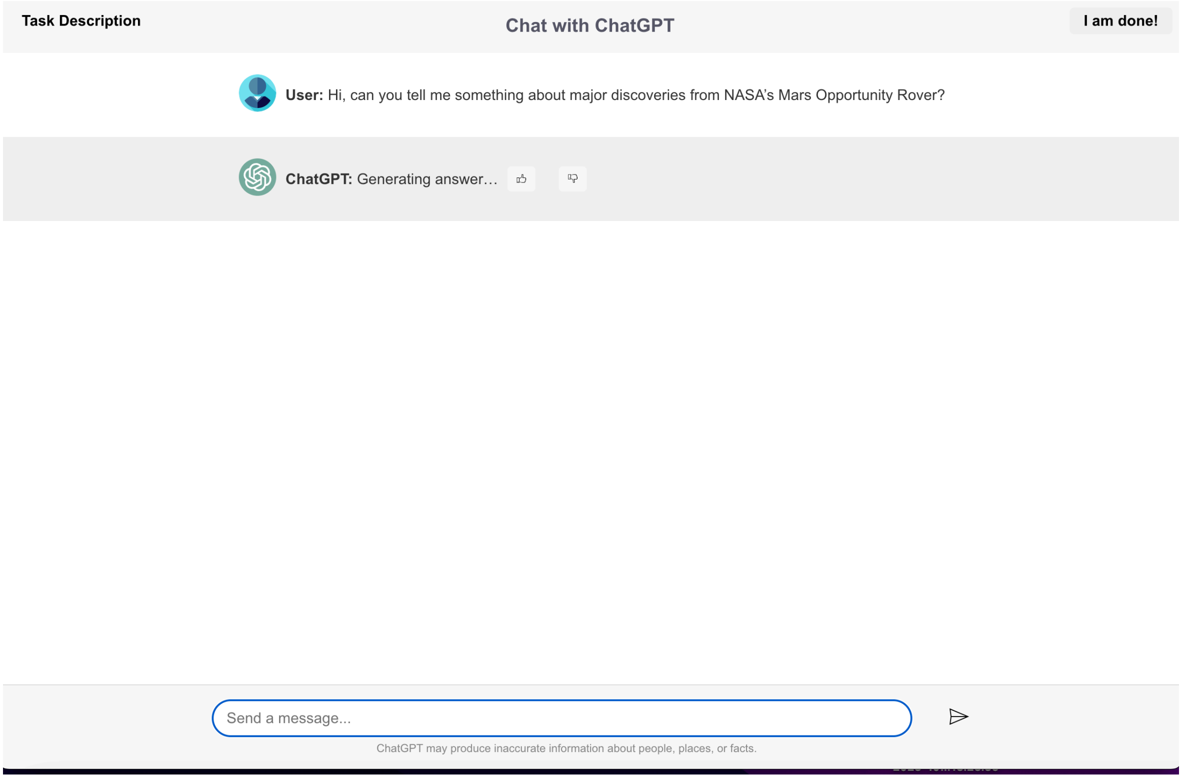
ChatGPT may produce inaccurate information about people, places, or facts.

Figure 3.3: Instruction of the mock ChatGPT used in the experiment



Your task is to acquire information about the different treatments for altitude sickness.

Figure 3.4: Task Description of the mock ChatGPT used in the experiment



User: Hi, can you tell me something about major discoveries from NASA's Mars Opportunity Rover?

ChatGPT: Generating answer...

ChatGPT may produce inaccurate information about people, places, or facts.

Figure 3.5: Generating answers of the mock ChatGPT used in the experiment

Implementation

Fig 3.6 shows the architecture of the experiment. Once workers qualify, they first joined the experiment via the link on the Prolific, after clicking the link, they was taken straight to the Qualtrics survey. In the middle of the survey, they were nudged into the application based on React and Express.js and deployed on the SURF Research. When the user gets into the application, the activity will be recorded by LogUI [42] and the data will be stored in the MongoSQL database. Due to the regulation of data management, the database was deployed on the secure cloud platform and the data will be stored in the TU Delft data storage. When the user input their information in the input part, The input was sent to the ChatGPT API based on version of GPT-3.5 on 11/9/2023.

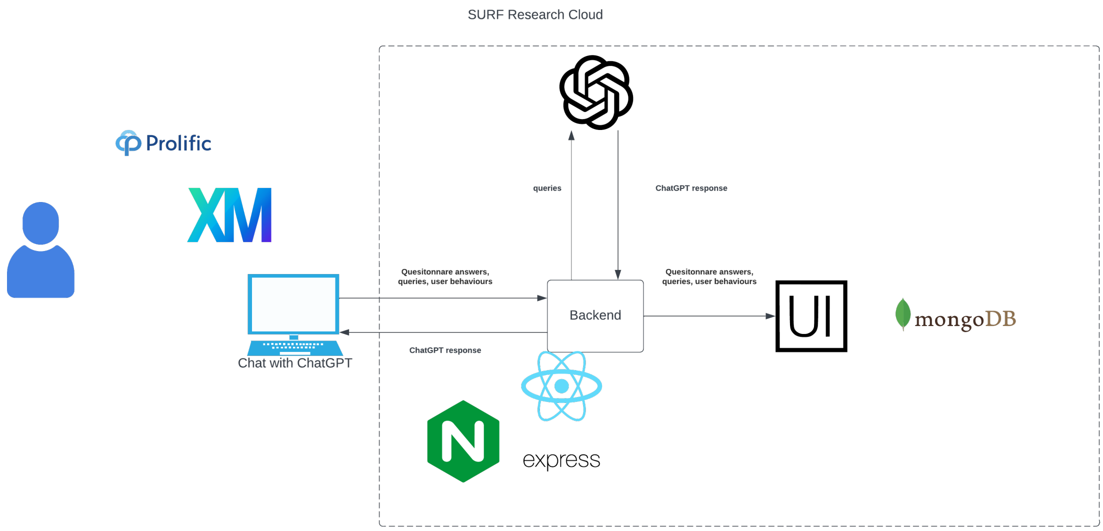
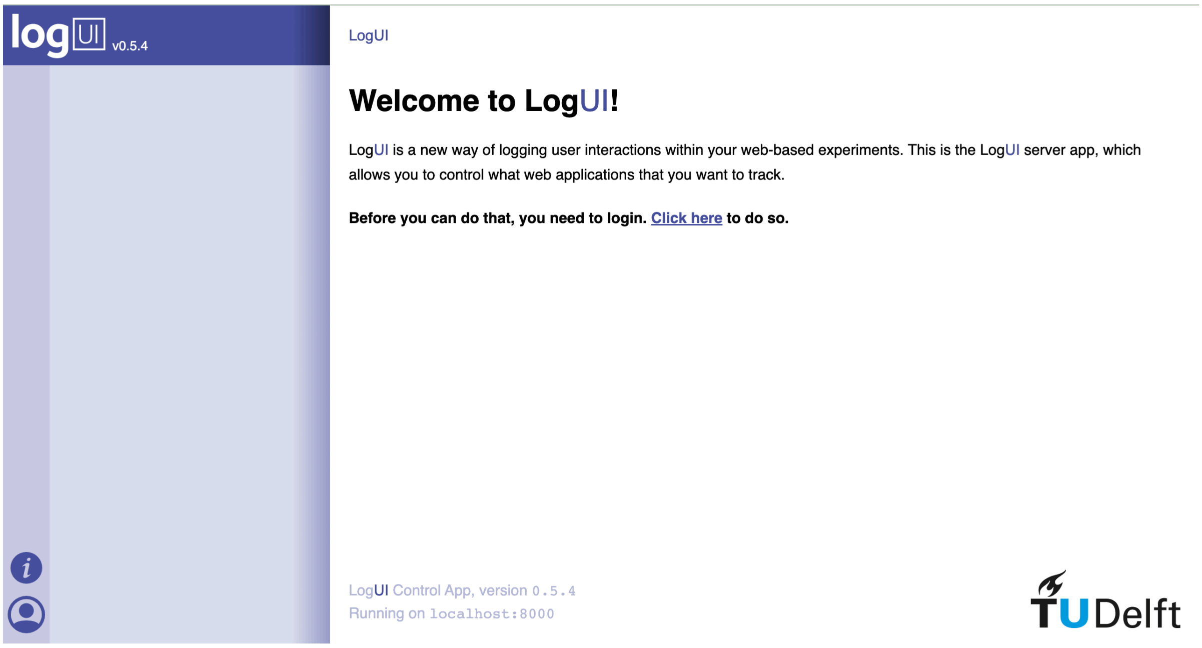


Figure 3.6: Architecture of the software



LogUI

Welcome to LogUI!

LogUI is a new way of logging user interactions within your web-based experiments. This is the LogUI server app, which allows you to control what web applications that you want to track.

Before you can do that, you need to login. Click here to do so.

LogUI Control App, version 0.5.4
Running on localhost:8000

Figure 3.7: LogUI interface

3.1.5. Post-task questionnaire

User engagement questionnaire

In this research, we intend to use the User Engagement Scale Short Form (UES-SF) to measure user engagement, which is widely used in user engagement measurement in HCI field [49]. This questionnaire measures user engagement with 12 questions that are divided into 4 different categories: focused attention, perceived usefulness, aesthetic appeal, and reward factor, respectively [46]. We used 7 point Likert scale for these questions and the user engagement level can be calculated by the average score of these 12 questions.

3.2. Variables

3.2.1. Independent variable

As discussed in the Chapter 1, we are investigating how ChatGPT Expertise, Topical Expertise, and Topical Interest influence user behaviour and user engagement in search sessions in ChatGPT. Hence, in our research, we have three independent variables as shown below:

- ChatGPT Expertise: Knowledge and interest about the functionality, mechanism, strengths, and limitations of the ChatGPT.
- Topical Expertise: ”Knowledge of the topic of the information need” [60].
- Topical Interest: Interest in a particular topic.

3.2.2. Dependent variables

Before introducing the dependent variables, we would like to point out that the search session in ChatGPT is defined as the period of time between clicking the ”Start” and ”Finish” buttons or of the longest duration possible.

User behaviour (RQ1)
As discussed in the Chapter 1, for RQ1, we investigate how user behaviour is affected in the To measure the user behaviour in ChatGPT search session. We propose the following metrics. These metrics are also widely used in the search field, which may facilitate to comparison of the result with existing work. And there are the metrics and the definitions below:

- **Average of prompt length.** The average length of prompts during a search session. (Continuous)
- **Number of prompts.** Number of prompts during a search session. (Continuous)
- **Number of unique prompts.** Number of unique prompts during a search session. (Continuous)
- **Average prompt duration.** The average amount of time that user spends in formulating a prompt and reading the response during a search session. (Continuous)
- **Prompt complexity.** Reading level which is represented by the average age of word acquisition of all query words during a search session [22, 16]. (Continuous)
- **Session length.** The duration of a search session (Continuous)

User engagement (RQ2)
As mentioned before, we selected to user engagement form (UES-SF) because of its popularity [46]. User engagement is measured from 4 different dimensions and more details are presented below:

- **User engagement level.** A feature of the user experience determined by the level of the user’s involvement with a digital system, which calculated by the average score of all the questions from UES-SF, which is measured by following 4 different dimensions and each dimension is measured by 3 questions [46, 45]. (Continuous)

  Aesthetic Appeal (AE). Level of interface appeal and visual appeal

  Perceived Usability (PU). Level of negative impact caused by interaction efforts.

  Focused Attention (FA). Level of feeling absorbed during the interaction.

  Reward (RW). Level of feeling rewarding during the interaction.

3.2.3. Descriptive and exploratory variables

Descriptive variables
To verify the representativeness of data, we introduce three descriptive variables into our experiment namely gender, age, and highest level of education completed. After collecting data, we will compare the distributions between the sample and the population. For gender [63] and age [10], we found their distribution in the world. For the educational level, due to the missing data on the education level of distribution in the world, we OECD country data to conduct a comparison.

Nr.	Topic	Task
1	Altitude Sickness	In this task, you are required to acquire information about the different treatments for altitude sickness. (20 items)
2	American Revolutionary War	In this task, you are required to acquire information about key battles of the American Revolutionary War and their impact. (10 items)
3	Carpenter Bees	In this task, you are required to acquire information about types of habitats carpenter bees prefer. (10 items)
4	Evolution	In this task, you are required to acquire information about main criticisms of the theory of evolution. (12 items)
5	NASA Interplanetary Missions	In this task, you are required to acquire information about major discoveries from NASA's Mars Opportunity Rover. (20 items)

Table 3.1: Topics for topic expertise measurement and corresponding tasks [22]

Exploratory variables
Excepts for the independent variables presented before, we are also interested in investigating how some other variables influence user behaviour and engagement. And there are affinity for technology interaction (ATI), frequency of use of ChatGPT, and Previous experience with chatbots or digital assistants. In the following part, we are going to introduce the each of them and corresponding motivation.

- **Affinity for technology interaction (ATI).** ”The tendency to actively engage in intensive technology interaction.”[20]

  Affinity for technology interaction (ATI) is defined as ”the tendency to actively engage in intensive technology interaction” [20] and, in other words, it is used to measure whether a person is willing to intensively interact with technology. Additionally, ATI can be easily measured by questionnaire and this questionnaire can be found in Appendix **??**. Since it is possible that the participants with high ATI and limited expertise may behave and engage differently due to the willingness to interact with technology even if expertise is limited, it is interesting to investigate how ATI influence the behaviour and engagement in ChatGPT.

- **Frequency of use of ChatGPT.** The frequency that the user interacted with ChatGPT in the past.
  According to the definition of ChatGPT expertise mentioned in the Section **??**, the frequency of use of ChatGPT is not taking into account as we assume that frequent usage does not make user be an expert [8]. However, it is also interesting to investigate the influence of the frequency of use of ChatGPT as some research consider the frequency of use also plays an important role in information seeking [25]. As the its definition is straightforward, it is measured by a question with 7-point Likert Scale. If result is larger than 3.5 then the frequency of use of ChatGPT is considered a high one.

- **Previous experience with chatbots or digital assistants.** The frequency that the user interacted with chatbots or digital assistants.
  Since we assume that the previous experience of interacting with chatbots could also influence how users behave and engage in ChatGPT due to the similarity between the chatbots and ChatGPT. So we also would like to look into how this experience may affect the behaviour and engagement in the interaction with ChatGPT, which is also captured by one question with a 7-point Likert Scale.

3.3. Sample size and analysis plan

We planned to use ANOVAs for hypothesis testing if some specific requirements are satisfied. However, because the data distribution does not satisfy the parametric assumption [18], the Kruskal-Wallis tests was used alternatively, which is also called one-way ANOVA on ranks [18]. The analysis was conducted in Python, and we used Holm–Bonferroni method to control the familywise error rates (FWER) so that, in this instance, the p-value will be $\frac{0.05}{6} = 0.0083$ [18]. Furthermore, we assume that effect size, alpha error probability, and power are equal to 0.25 (moderate effect). Due to the fix budget, the power is set to $1 - \beta = 0.6$ ([18]. If the hypothesis result is significant then we will conduct the post hoc test at the end of hypothesis testing. As we have 2 different groups in our research, according to the values of these parameters, the sample size is 156 participants based on the result of $G^{*}Power$ [17].

4

Data collection and processing

In this chapter, we outline the specific criteria for qualified submissions and also provide an explanation of the methodology employed in collecting and processing data to ensure a comprehensive understanding of our approach.

4.1. Quality control

To ensure the accuracy of our result, we involve quality control measures in the experiment and these measures are crucial components in our data collection process. If any participant submission did not meet one of criteria, it was not considered for analysis. And these criteria are presented below:

- Passing attention check tests
- Recaptcha score is larger than 0.5 in Qualtrics [51]
- ChatGPT experiment verification code is valid
- Using related prompts while interacting with ChatGPT

4.2. Data exclusion

Duration of the data collection, several workers were rejected or got compensation due to failed quality control or did not show Excluding the data points did not pass the quality control, we also realize there are some limitations in our instruction: Even if our survey and experiment are both in English, we did not explicitly mention that the users are supposed to interact with English. The reason why it is important to interact with ChatGPT in English is that different languages may have different habits while formulating the prompts and in this case, the result of user behaviour may be distorted. Furthermore, the research also shows that the ChatGPT may return different answers for the same topic given inputs in different languages, which may affect user behaviour in the conversation. In this case, experts for submissions violating quality control requirements, the submissions that are not fully in English are also disregarded in the analysis. Here are the data excluded while processing the data:

- 1 workers did not passed robot (Recaptcha) test.
- 20 workers did not use related prompts.
- 4 workers interacted with ChatGPT in non-English languages such as French, Spanish, or Polish.
- 3 workers were excluded due to technical issues.

After excluding invalid submissions, 191 data points are taken into account in the analysis.

Independent Variable	Low	High	Total
ChatGPT Expertise	106	85	191
Topical Expertise	92	95	187
Topical Interest	54	54	108

Table 4.1: Data distribution for each independent variables

4.3. Data progressing

4.3.1. Data management

To protect the privacy of the participant, the prolific ID is hashed to ensure our workers are not traceable based on the data and the Qualtrics survey data and log information from LogUI were stored in the TUDelft safe drive.

4.3.2. Variables calculation

To investigate the influence of the ChatGPT expertise, topical expertise, and topical interest, we first need to tag qualified participants with Low/High for each of the independent variables. Afterwards, we calculate the value of each metric for each participant.

Independent variables calculation

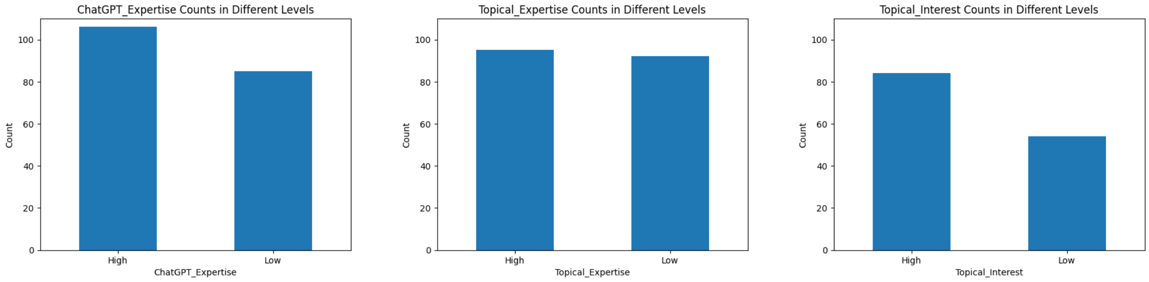
To investigate the answers to research questions, the independent variables are calculated in the manner below:

To enhance the generalizability of results, we aim to collect balanced ChatGPT expertise during the data collection. Ideally, we collect balanced data points for all 3 independent variables. However, due to the fixed funding, this goal is less realistic since it is required to pay for the unqualified workers based on the estimated experiment time. To reduce the effect of noise in the analysis, the 37.5th percentile and 67.5th percentile are used to classify participants. Specifically, if the user with ChatGPT expertise is above the 67.5th percentile□then the user will be tagged as 'High' for ChatGPT. If the topical expertise is above the 67.5th percentile, then the user will be tagged as 'High' for topical expertise. To obtain the balanced data for ChatGPT expertise with less waste of data points. We first keep collecting data until it is close to normal distribution. But if the distribution is skewed then we would think about another way to acquire balanced ChatGPT expertise. Fortunately, the distribution seems like a normal one after collecting around 100 data points and the statistical test result also shows that the distribution is likely to be a normal distribution. Based on the received data, the threshold for low ChatGPT expertise is set as 4.2 and the threshold for high ChatGPT expertise is set as 5.2. Afterwards, the rest of the data collection is based on these two thresholds and it was stopped when the budget was run out. There were 212 submissions at the end of data collection. For topical expertise and interest, 37.5th and 67.5th percentile thresholds were also selected to reduce the noise in the data. The distribution of topical expertise and topical interest for each topic can be found from the bar charts above and The number of data points for each level of expertise and interest are presented.

Dependent variables calculation

Since the raw data consist of prompt from users, and timestamps of the start and the end of the interaction, etc, the required dependent variables for user behaviour have to be calculated and Here is how these dependent variables are calculated shown below:

- **Average of prompt length.** The average number of words used in the prompts for each user.
- **Number of prompts.** The number of prompts of the user in the session.
- **Number of distinct prompts.** The number of distinct prompts of the user in the session and distinct prompts are identified by using cosine similarity.



(a) ChatGPT expertise distribution

(b) Topical expertise distribution

(c) Topical interest distribution

Figure 4.1: Independent variables distributions

- **Average prompt duration.** The average amount of time that the user spends for prompt formulation and response reading.
- **Prompt complexity.** It is represented by the maximal age of word acquisition of all query words of a user during the search session [22, 16].
- **Session length.** The duration of the search session of a user.

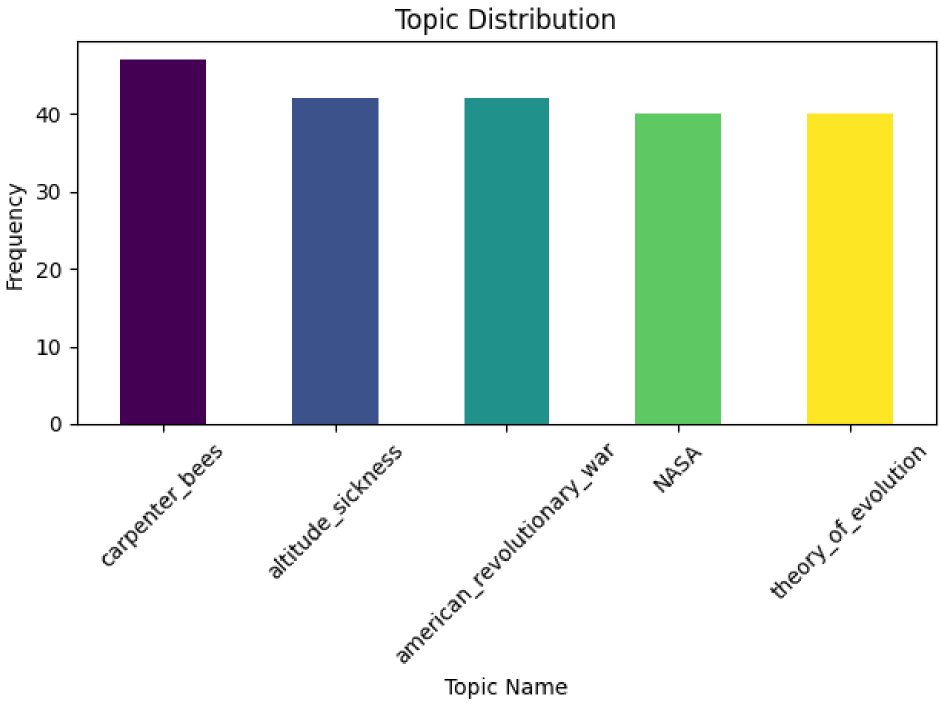
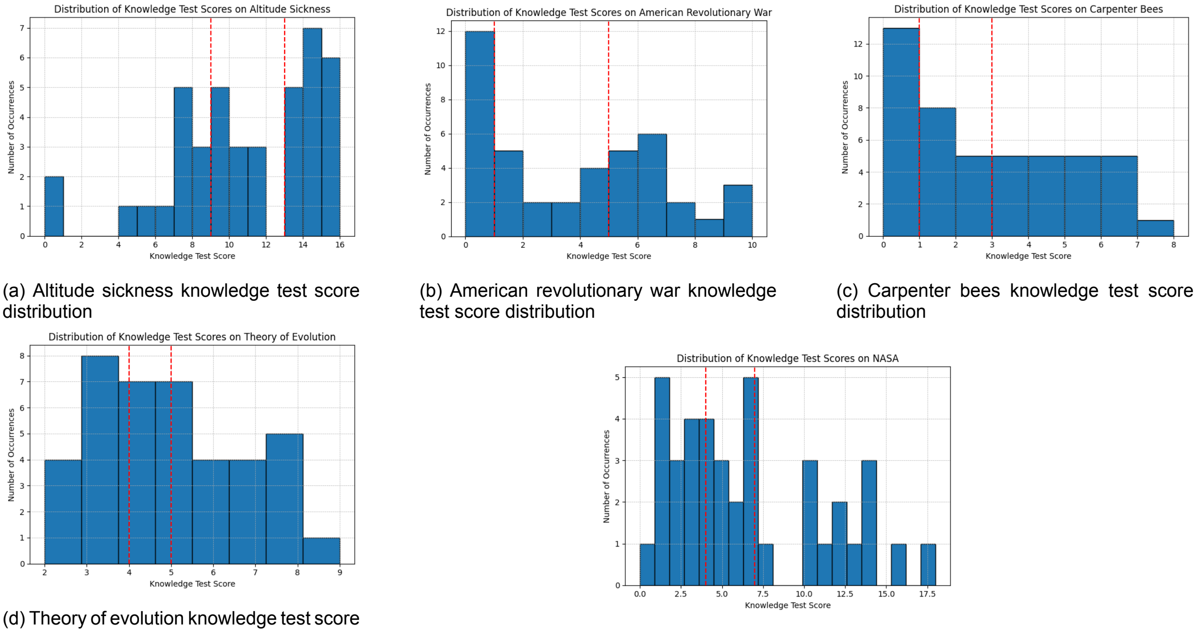


Figure 4.2: Topic distribution in the experiment

4.4. Statistical test selection

To be able to use parametric test like ANOVA, there are four assumptions [18] are supposed to be satisfied and they are:

- Variable are normally distributed.
- Variances of the different groups are supposed to be approximately equal.
- Data can be measured at the interval level.



(a) Altitude sickness knowledge test score distribution

(b) American revolutionary war knowledge test score distribution

(c) Carpenter bees knowledge test score distribution

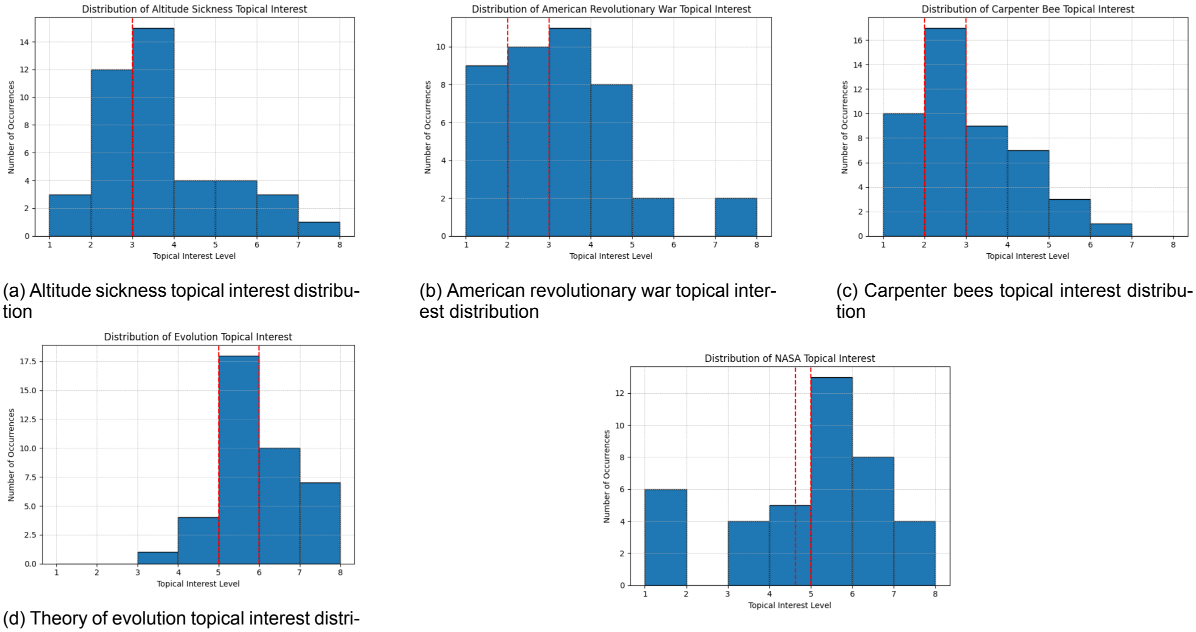
(d) Theory of evolution knowledge test score distribution

(e) NASA knowledge test score distribution

Figure 4.3: Knowledge test score distribution per topic

- Data are independent.

These four assumptions are supposed to be satisfied for paramedic testing [18] and one of the important assumptions is that the data assumption is supposed to be a normal distribution. To verify the distribution of dependent variables. The Levene test and Q-Q plot were conducted and drawn. Since the original data are not in normal distribution, data transformation was also conducted. However, the distributions are still skewed. Because of this reason, the Kruskal–Wallis tests were selected instead of ANOVA.



(a) Altitude sickness topical interest distribution

(b) American revolutionary war topical interest distribution

(c) Carpenter bees topical interest distribution

(d) Theory of evolution topical interest distribution

(e) NASA topical interest distribution

Figure 4.4: Topical interest distribution per topic

5 Results and analysis

In this chapter, we present our findings based on the processed data. First, we demonstrate the descriptive statistics to show the representativeness of our data. Next, statistical test results for both independent and exploratory variables and corresponding analysis are presented.

5.1. Descriptive Statistics

5.1.1. Gender

Figure 5.1 illustrates the gender distribution of the participants. By leveraging the balanced distribution function in the Prolific while collecting data, the gender ratio is closed to 1:1. which is similar to the poPerceived Usability dimensionlation male/female ratio 1.01:1 [63], which means the sample is demographically representative in terms of gender.

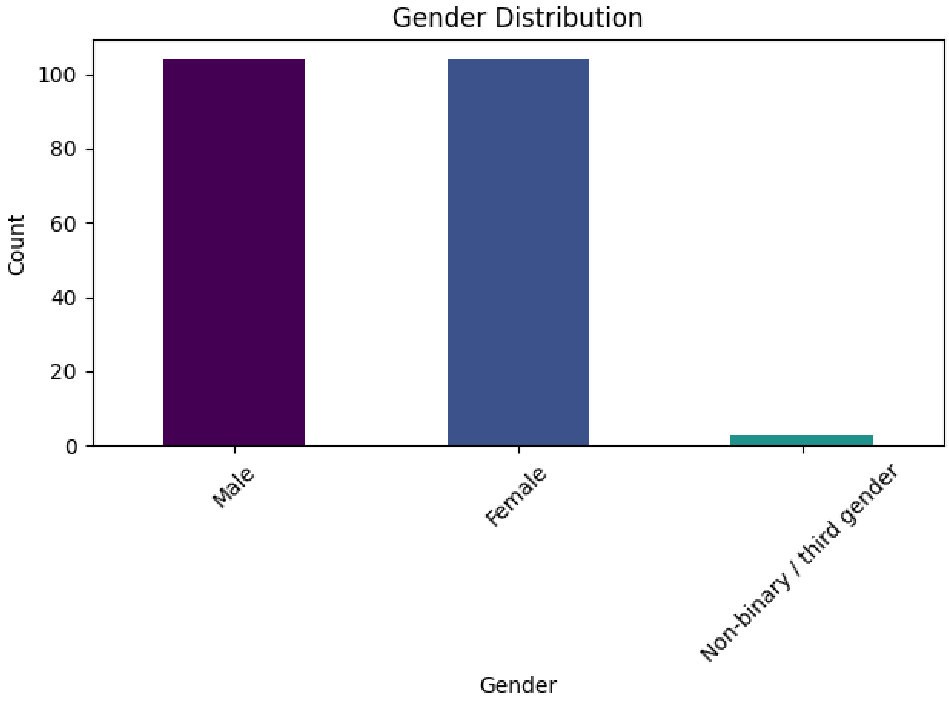


Figure 5.1: Gender distribution in collected data set

Age range	Sample	Population in the world
0-14	0	25.5%
15-24	40.3%	15.5%
25-39	50.2%	22.2%
40-64	9.5%	27.2%
65-74	0	6.0%
75+	0	3.6%

Table 5.1: Age distribution comparison between sample and population in the world

5.1.2. Age

The participant's age distribution is shown in the Figure 5.2. Due to the regulations, we did not collect data from people under 18. As shown the table above, the Sample column represent the percentage of participants within a specific age range and Population in the world column is the percentage of people within a specific range in the world [10]. As what you can see from this table, the most of participants is under 39 and the data from participants above 65 are missing, which means that our sample may represent more about the behaviour of the younger group.

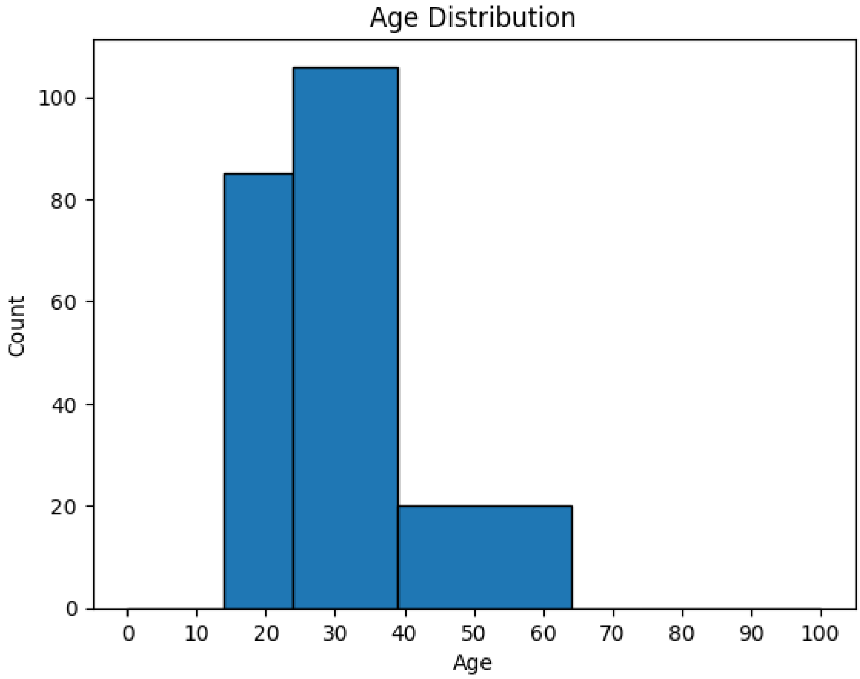


Figure 5.2: Age distribution

5.1.3. Educational level

As shown in Figure 5.3, most participants obtained upper secondary education or tertiary diploma. As shown in table below, the rightmost column represents the distribution of education level in OECD countries and 63.51% participants have a tertiary degree as compared to 40.44% tertiary degree in the OECD countries, which means that our sample may represent more about the users with higher

Highest educational level	Sample	PoPerceived Usability dimensionlation in OECD countries
Below upper secondary	3.32%	19.75%
Upper secondary	33.18%	40.23%
Tertiary	63.51%	40.44%

Table 5.2: Education level comparison between sample and poPerceived Usability dimensionlation in OECD countries

diploma.

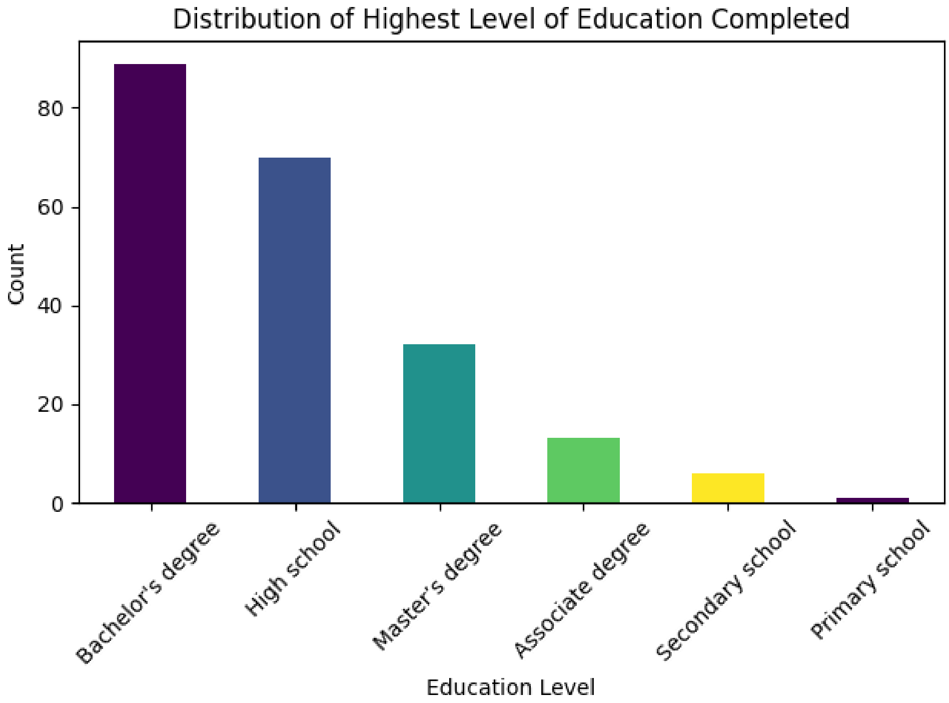


Figure 5.3: Educational level distribution

5.2. Hypothesis Testing

To verify proposed hypotheses (See Chatpter 1) for research questions, the statistical test was conduct for each hypotheis. As discussed in the Chapter 4, the distributions of dependent variables are not in normal distributions so that not all the assumptions for ANOVA are satisfied. Because of this reason, the Kruskal-Wallis test (also called one-way ANOVA on ranks [18]) was utilized for all the hypotheses and the significance threshold was set as 0.05 / 6 = 0.0083 due to the Bonferroni correction.

5.2.1. RQ1 - User behaviour

To verify whether there is significant difference of the user behaviour between users in different levels of ChatGPT expertise in ChatGPT search sessions (**H1a**), the Kruskal-Wallis test was conducted. And we can see from Figure 5.4, the medians of different user behaviours metrics between users with different levels of ChatGPT expertise and similar. Furthermore, since the p-values from all different user behaviour metrics are larger than 0.0083, the results of user behaviour metrics are not significant as shown in Table 5.3, so there is no significant relationship between ChatGPT expertise and user behaviour metrics. Based on these findings, the **H1a** is rejected.

Regarding the hypothesis about the difference of the user behaviour between users in different levels of topical expertise (**H1b**), the results of user behaviours metrics are also not significant. However, as shown in the Figure 5.5, the medians of the number of prompts, number of distinct prompts differ between different levels of topical expertise, which means the users with higher topical expertise tend to use more prompts and distinct prompts as compared to users with relatively low topical expertise. To investigate what could be the reasons why participant with different topical expertise level behaved differently, we looked into the prompts used in the experiment and we found the users with higher topical expertise it is more likely to go deeper in the search session. As compared to the participants with lower topical expertise, the participants with lower topical expertise is more possible to directly use task description in the interaction and ask limited following questions and stopped. That could be the reason why the the participants with relatively high topical expertise use more prompts in average. Furthermore, the user with higher topical expertise tend to use less time for each prompt, which is probably because they are more familiar with the result. Based on these findings, we may assume that the user with higher topical expertise behave differently probably because they have more motivation to explore more and are more familiar with topic. However, since the statistical test results are not significant, the **H1b** is rejected. In terms of the hypothesis about the difference of the user behaviour between users in different levels of topical interest (**H1c**), we noticed that the according to the hypothesis results shown in Table 5.3, there is no correlation between topical interest and user behaviour. So the **H1c** is not supported.

5.2.2. RQ2 - User engagement

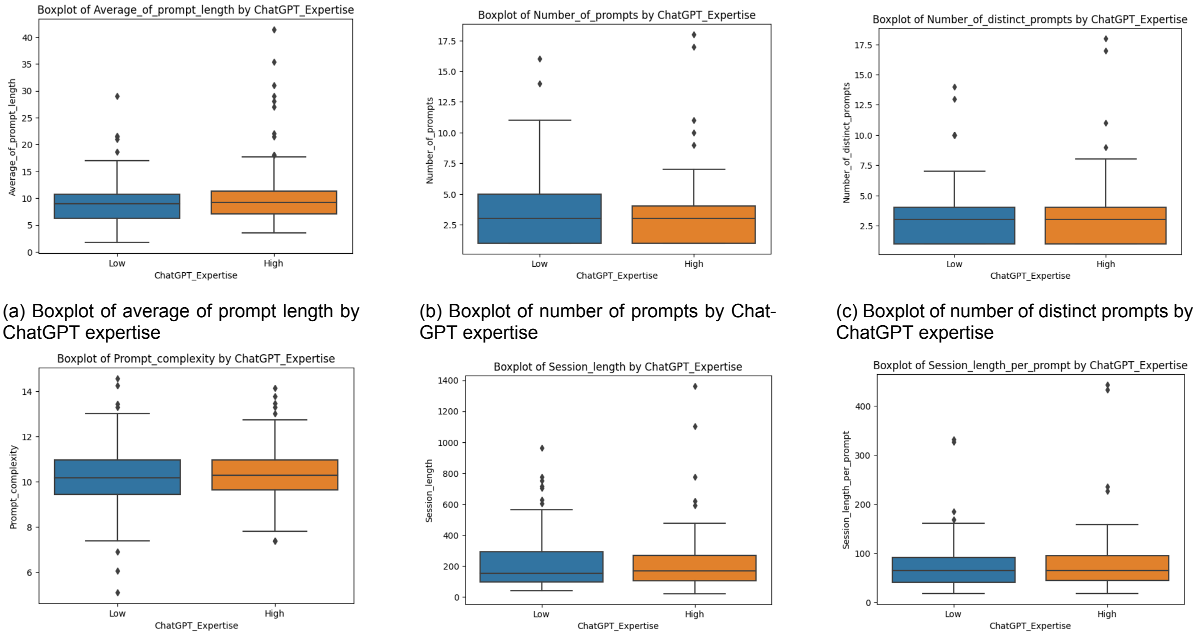
The statistical test results of user engagement are presented in Table 5.3. Since the User Engagement Scale (UES) measures user engagement from 4 different aspects [46], we also investigated whether the ChatGPT expertise, topical expertise, or topical interest influence one of the dimensions of user engagement.

For the hypothesis about the difference in user engagement between users in different levels of ChatGPT expertise (**H2a**), there is no statistically significant relationship between ChatGPT expertise and user engagement according to the result of the Kruskal-Wallis test. However, we found that the effect of ChatGPT expertise on the Reward aspect of user engagement was statistically significant (H Statistic = 2.541, p = 0.001). As shown the Figure 5.7, the median user engagement score Reward of users with higher ChatGPT expertise is significantly higher, which means the users with relatively high ChatGPT expertise may feel more rewarding while conducting search in ChatGPT and we assume that it is because the users with high ChatGPT expertise is able to formulate their prompts and are able to seek required information effectively. Based on the results of the p-values of H2a, the Hypothesis H2a is partially supported.

Regarding the relationship between topical expertise and user engagement (**H2b**) and the relationship between topical interest and user engagement (**H2c**), the significant test results did not show the effect of topical expertise or topical interest on user engagement or one is significant and the results of the corresponding statistical tests can be found in Table 5.3. Hence, the H2b and H2c are rejected since all the p-values of metrics are above 0.0083. But, in the meanwhile, we also noticed that, for the (**H2b**), the users with higher topical expertise have higher user engagement scores, particularly in the Aesthetic Appeal dimension as shown in the boxplots from Figure. 5.8, which means that the users with more topic knowledge may be more likely to be attracted by the design of the interface. We assume that this attraction may originate from their smooth searching experience so that the interface design appears more appealing because of this positive interaction. For the (**H2c**), the medians of the Reward aspect are different between different levels of topical interests, which also means the users with relatively high topical interest may feel more rewarded while conducting the search in ChatGPT and one possible explanation is that the users are interesting on a specific topic could dive deeper into the experiment, leading to a stronger sense of reward.

In general, based on the outcomes of statistical tests, H1a, H1b, H1c, H2b, and H2c are rejected and H2a is partially supported since there is a significant relationship between topical interest and topical and reward aspect of user engagement.

No.	Hypothesis	Metric	p-value	Result
			Continued on Next Page	
H1a	People with relatively higher ChatGPT expertise behave differently compared to people with relatively lower ChatGPT expertise	Average of prompt length	0.087	Not significant
		Number of prompts	0.700	Not significant
		Number of distinct prompts	0.546	Not significant
		Prompt complexity	0.054	Not significant
		Session length	0.933	Not significant
		Average prompt duration	0.629	Not significant
H1b	People with relatively higher topical expertise behave differently compared to people with relatively lower topical expertise	Average of prompt length	0.070	Not significant
		Number of prompts	0.011	Significant
		Number of distinct prompts	0.020	Significant
		Prompt complexity	0.903	Not significant
		Session length	0.235	Not significant
		Average prompt duration	0.019	Significant
H1c	People with relatively higher topical interest behave differently compared to people with relatively lower topical interest	Average of prompt length	0.068	Not significant
		Number of prompts	0.274	Not significant
		Number of distinct prompts	0.286	Not significant
		Prompt complexity	0.721	Not significant
		Session length	0.858	Not significant
		Average prompt duration	0.161	Not significant
H2a	People with relatively higher ChatGPT expertise exhibit different levels of engagement compared to people with relatively lower ChatGPT expertise	User engagement score	0.09	Not significant
		Focused Attention dimension score	0.942	Not significant
		Perceived Usability dimension score	0.794	Not significant
		Aesthetic Appeal dimension score	0.442	Not significant
		Reward dimension score	**0.001**	**Significant**
H2b	People with relatively higher topical expertise exhibit different levels of engagement compared to people with relatively lower topical expertise	User engagement score	0.149	Not significant
		Focused Attention dimension score	0.279	Not significant



(a) Boxplot of average of prompt length by ChatGPT expertise

(b) Boxplot of number of prompts by ChatGPT expertise

(c) Boxplot of number of distinct prompts by ChatGPT expertise

(d) Boxplot of prompt complexity by ChatGPT expertise

(e) Boxplot of session length by ChatGPT expertise

(f) Boxplot of average time of prompt formation by ChatGPT expertise

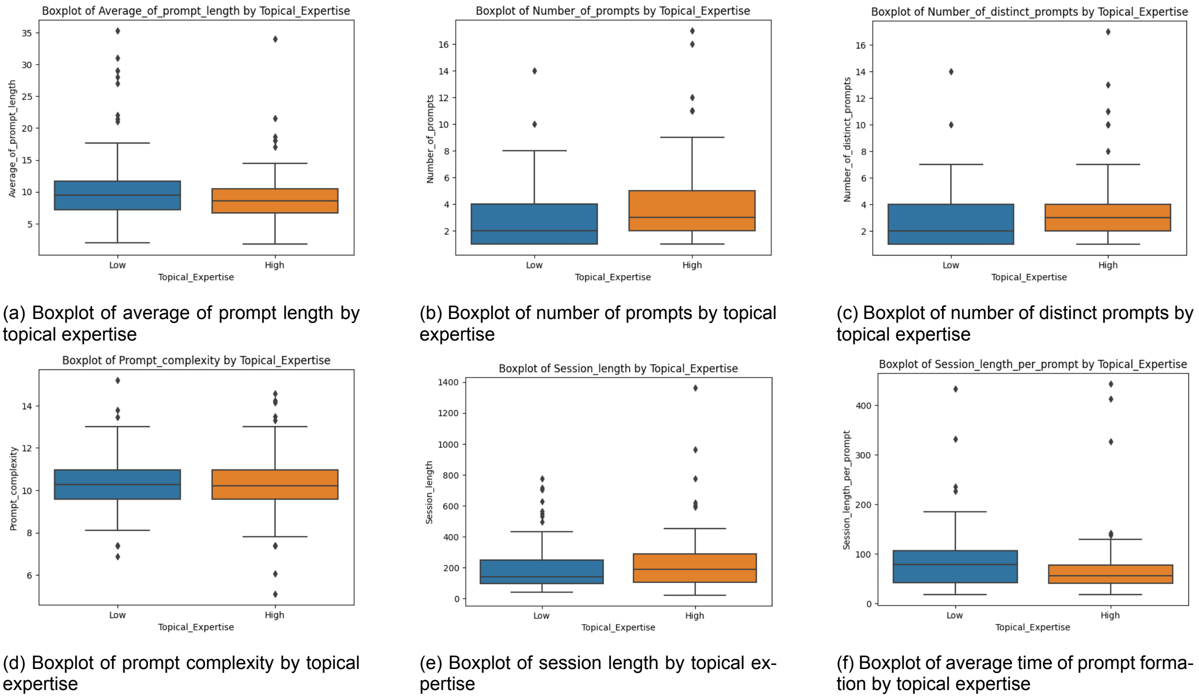
Figure 5.4: Boxplots of user behaviour metrics by ChatGPT expertise

No.	Hypothesis	Metric	p-value	Result
		Perceived Usability dimension score	0.209	Not significant
		Aesthetic Appeal dimension score	0.123	Not significant
		Reward dimension score	0.303	Not significant
H2c	People with relatively higher topical interest exhibit different levels of engagement compared to people with relatively lower topical interest	User engagement score	0.215	Not significant
		Focused Attention dimension score	0.120	Not significant
		Perceived Usability dimension score	0.324	Not significant
		Aesthetic Appeal dimension score	0.648	Not significant
		Reward dimension score	0.075	Not significant

Table 5.3: Statistical results of hypotheses

5.3. Exploratory Findings

As mentioned in Chapter 3, in our research, we also investigated how affinity for technology interaction (ATI), frequency of use of ChatGPT, and previous experience with chatbots or digital assistants affect



(a) Boxplot of average of prompt length by topical expertise

(b) Boxplot of number of prompts by topical expertise

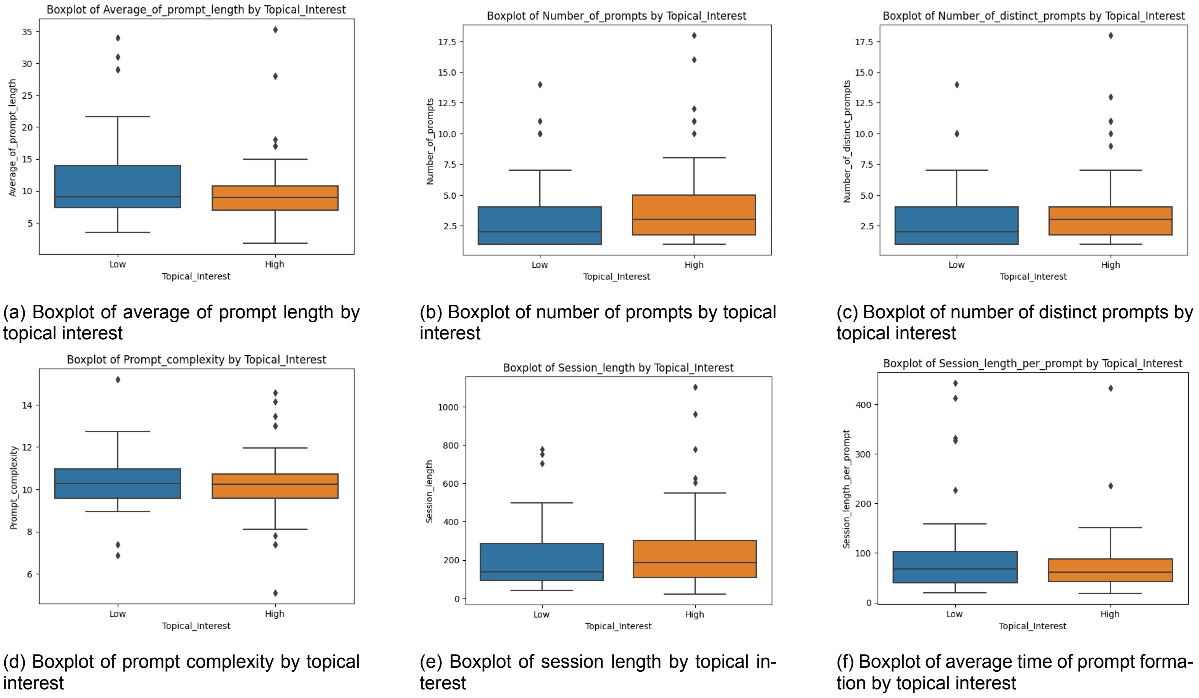
(c) Boxplot of number of distinct prompts by topical expertise

(d) Boxplot of prompt complexity by topical expertise

(e) Boxplot of session length by topical expertise

(f) Boxplot of average time of prompt formation by topical expertise

Figure 5.5: Boxplots of user behaviour metrics by topical expertise



(a) Boxplot of average of prompt length by topical interest

(b) Boxplot of number of prompts by topical interest

(c) Boxplot of number of distinct prompts by topical interest

(d) Boxplot of prompt complexity by topical interest

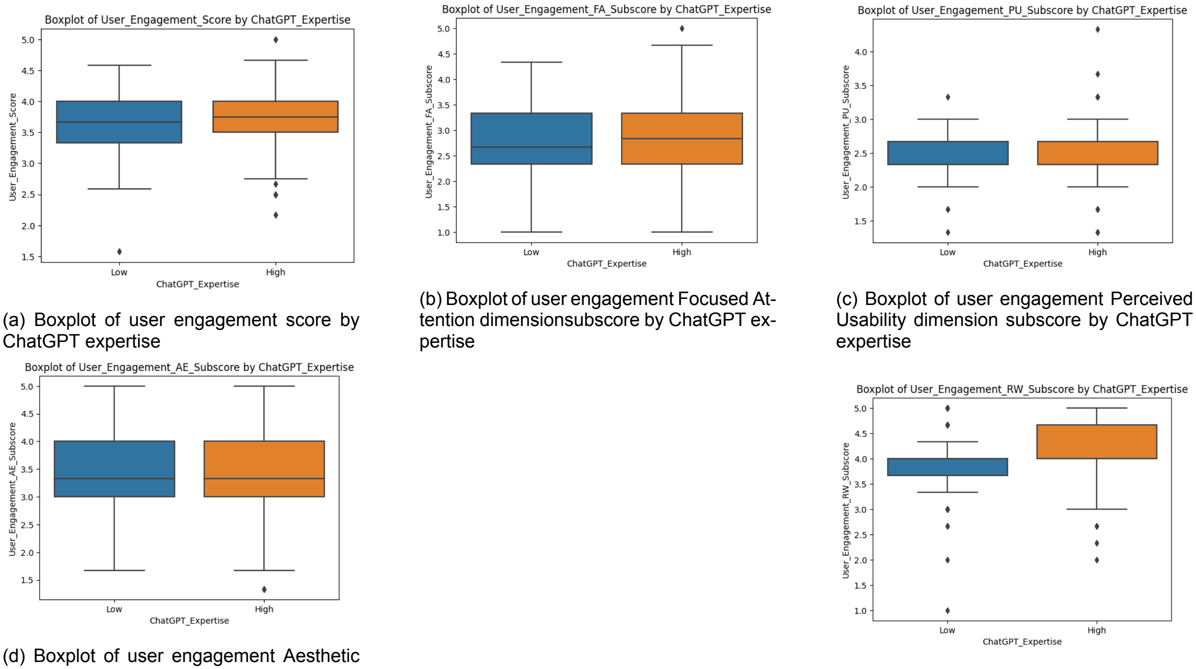
(e) Boxplot of session length by topical interest

(f) Boxplot of average time of prompt formation by topical interest

Figure 5.6: Boxplots of user behaviour metrics by topical interest

	ChatGPT expertise		Topical expertise		Topical interest	
	Low	High	Low	High	Low	High
Average of prompt length	9.24±4.56	10.83±6.36	10.80±6.40	9.09±4.33	11.48±7.00	9.59±4.69
Number of prompts	3.81±3.17	3.41±2.85	2.96±2.29	3.97±3.20	3.20±2.88	3.90±3.26
Number of distinct prompts	3.51±2.88	3.26±2.80	2.79±2.23	3.72±2.91	3.06±2.84	3.63±2.96
Prompt complexity	10.10±1.49	10.44±1.18	10.31±1.28	10.24±1.48	10.34±1.24	10.21±1.30
Session length	234.16±196.92	220.77±194.82	200.90±160.80	233.29±201.44	214.13±177.69	234.86±194.61
Average time of prompt formation	75.47±53.86	79.28±63.83	85.70±63.96	71.45±65.28	94.23±91.93	71.58±53.18

Table 5.4: Mean and standard deviation of user behaviour metrics for participants in different ChatGPT expertise, topical expertise, and topical expertise levels



(a) Boxplot of user engagement score by ChatGPT expertise

(b) Boxplot of user engagement Focused Attention dimensionsubscore by ChatGPT expertise

(c) Boxplot of user engagement Perceived Usability dimension subscore by ChatGPT expertise

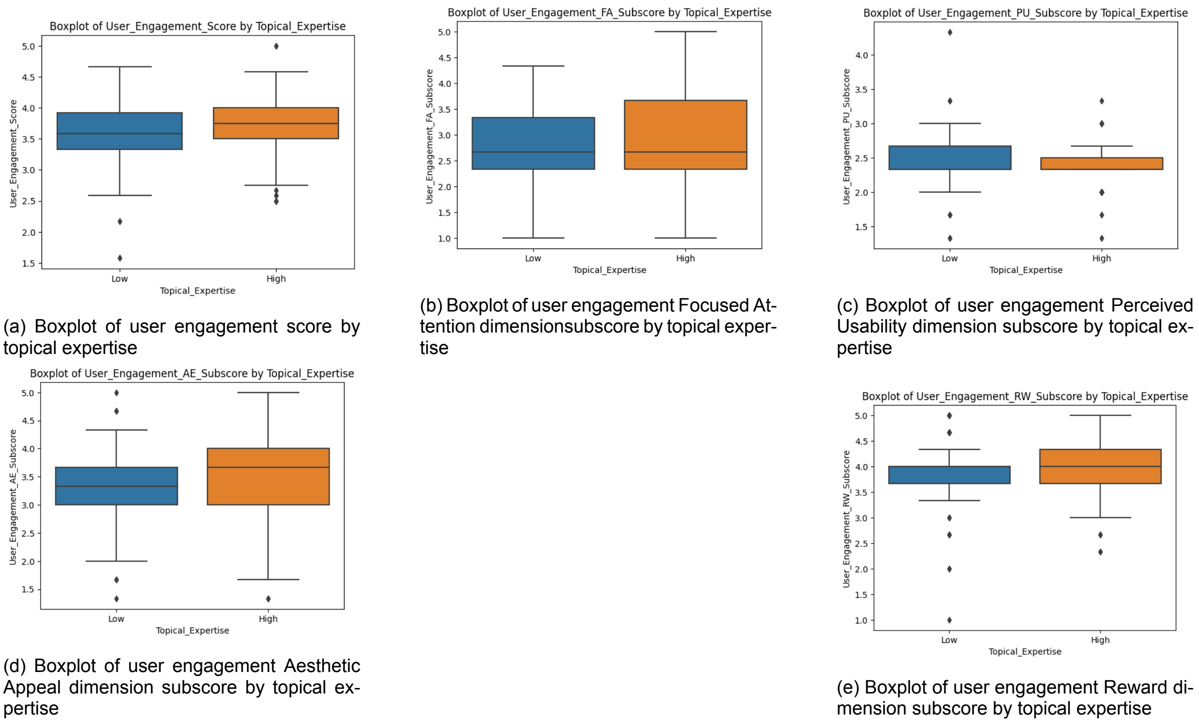
(d) Boxplot of user engagement Aesthetic Appeal dimension subscore by ChatGPT expertise

(e) Boxplot of user engagement Reward dimension subscore by ChatGPT expertise

Figure 5.7: Boxplots of user engagement metrics by ChatGPT expertise

	ChatGPT expertise		Topical expertise		Topical interest	
	Low	High	Low	High	Low	High
User Engagement Score	3.60±0.50	3.71±0.49	3.58±0.49	3.70±0.51	3.58±0.51	3.74±0.51
User_Engagement_FA_Subscore	2.84±0.72	2.87±0.81	2.76±0.71	2.93±0.85	2.81±0.82	3.06±0.78
User_Engagement_Perceived Usability dimension_Subscore	2.39±0.32	2.44±0.40	2.44±0.41	2.37±0.29	2.38±0.30	2.44±0.26
User_Engagement_Aesthetic Appeal dimension_Subscore	3.33±0.69	3.40±0.79	3.28±0.72	3.42±0.78	3.27±0.74	3.40±0.77
User_Engagement_Reward dimension_Subscore	3.85±0.65	4.13±0.59	3.90±0.66	4.01±0.60	3.87±0.74	4.12±0.62
Average time of prompt formation	75.47±53.86	79.28±63.83	85.70±63.96	71.45±65.28	94.23±91.93	71.58±53.18

Table 5.5: Mean and standard deviation of user user engagement metrics for participants in different ChatGPT expertise, topical expertise, and topical expertise levels



(a) Boxplot of user engagement score by topical expertise

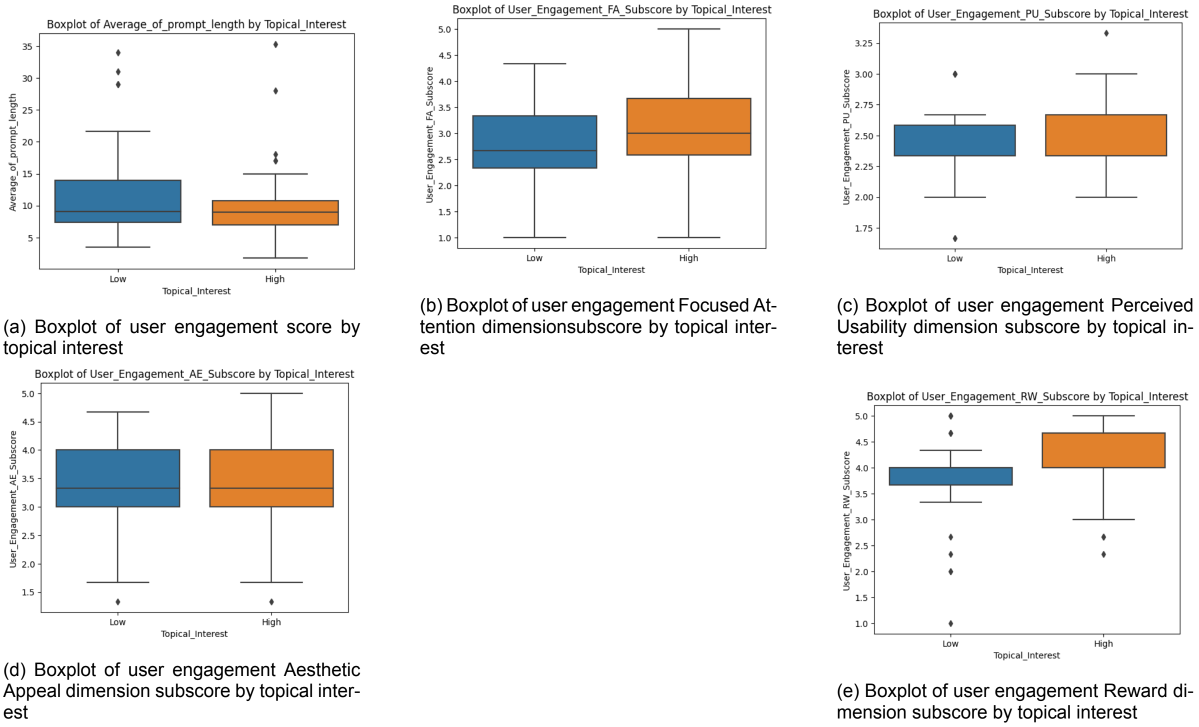
(b) Boxplot of user engagement Focused Attention dimensionsubscore by topical expertise

(c) Boxplot of user engagement Perceived Usability dimension subscore by topical expertise

(d) Boxplot of user engagement Aesthetic Appeal dimension subscore by topical expertise

(e) Boxplot of user engagement Reward dimension subscore by topical expertise

Figure 5.8: Boxplots of user engagement metrics by topical expertise



(a) Boxplot of user engagement score by topical interest

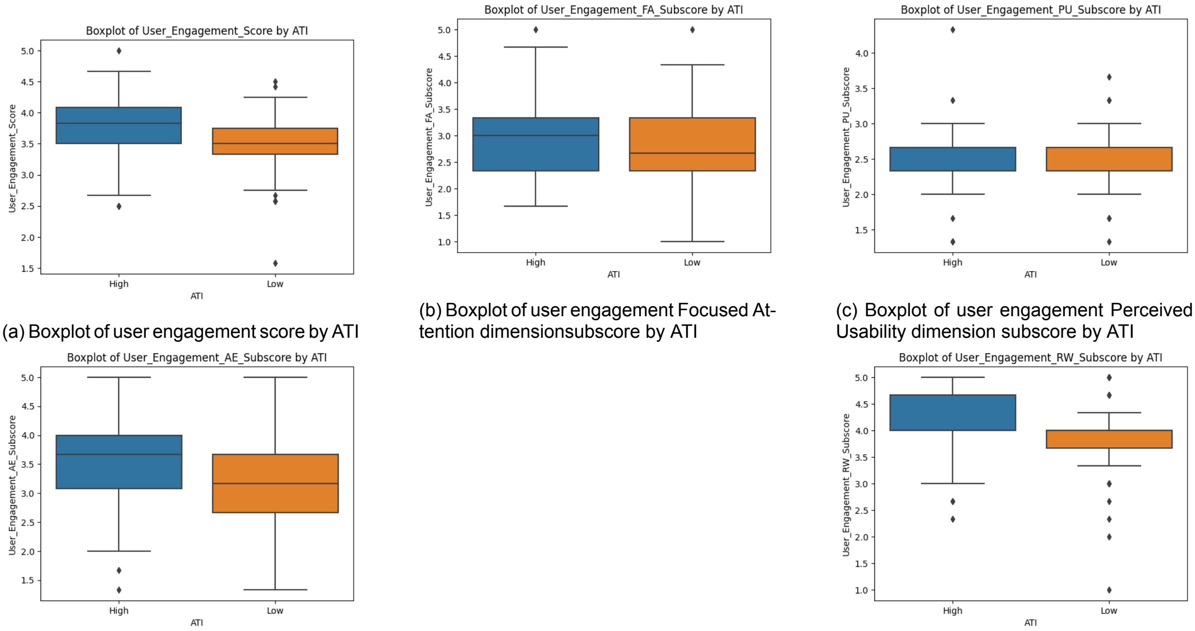
(b) Boxplot of user engagement Focused Attention dimensionsubscore by topical interest

(c) Boxplot of user engagement Perceived Usability dimension subscore by topical interest

(d) Boxplot of user engagement Aesthetic Appeal dimension subscore by topical interest

(e) Boxplot of user engagement Reward dimension subscore by topical interest

Figure 5.9: Boxplots of user engagement by topical interest



(a) Boxplot of user engagement score by ATI

(b) Boxplot of user engagement Focused Attention dimensionsubscore by ATI

(c) Boxplot of user engagement Perceived Usability dimension subscore by ATI

(d) Boxplot of user engagement Aesthetic Appeal dimension subscore by ATI

(e) Boxplot of user engagement Reward dimension subscore by ATI

Figure 5.10: Boxplots of user engagement by ATI

the user behaviour and engagement in search sessions in ChatGPT and here are what we found during the experiment below:

5.3.1. Affinity for technology interaction (ATI)

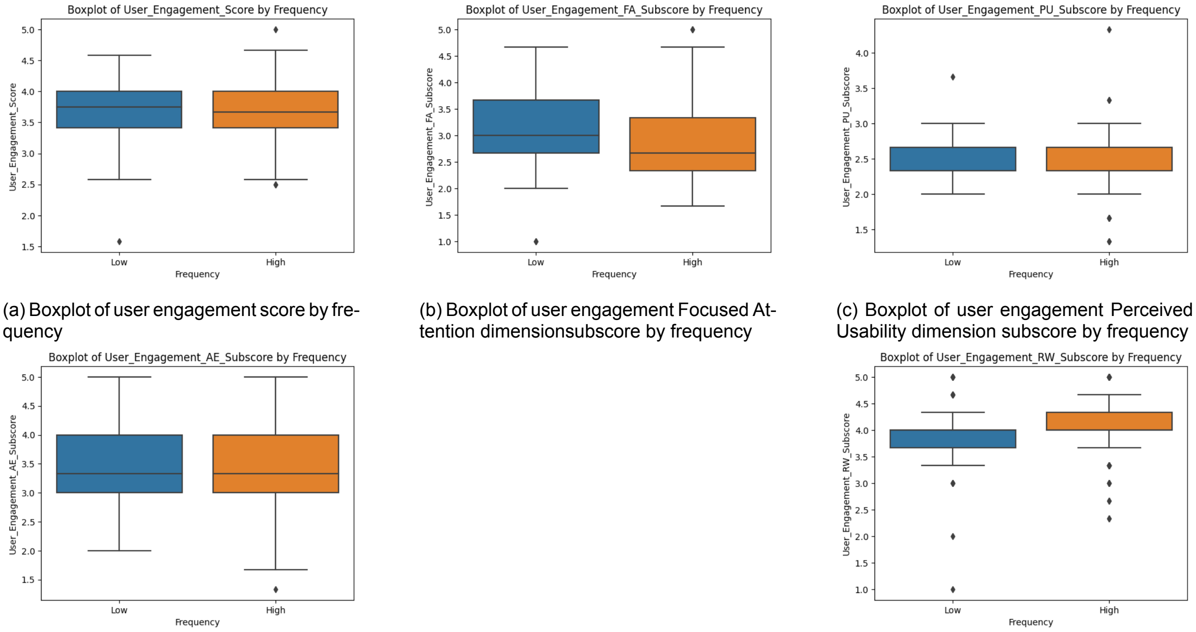
For ATI, we found that there is no statistically significant relationship between ATI and user behaviour since the statistical tests related to the user behaviours failed to reject the null hypotheses. However, we also found that the Kruskal-Wallis test revealed a significant effect of ATI on user engagement (H Statistic = 13.865, p = 0.000). As shown in Figure 5.10, there is a significant median difference in user engagement between the users with high ATI scores and low ATI scores. Furthermore, as shown in the Table 5.8, the effects of ATI on user engagement (Aesthetic Appeal dimension) (H Statistic = 9.562, p = 0.002) and Reward (Reward dimension) (H Statistic = 20.344, p = 0.000) aspects are also significant. According to the boxplots shown, the users with higher ATI tend to have higher Aesthetic Appeal dimension and Reward dimension on average.

However, due to the limited work regarding the relationship between the ATI and user behaviour and user engagement, we are not able to compare our findings with existing work. In this case, we look into the definition of ATI, and we can assume the reason why there are correlations between ATI and user engagement is that users with higher affinity which is probably because users with higher ATI are more willing to approach the interaction with ChatGPT so that they could be more engaged than other users. This assumption is consistent with the definition of the ATI but further exploration is desired to establish a more concrete understanding.

5.3.2. Frequency of use of ChatGPT

As shown in the Jenkins et al. [28]research, frequency sometimes also plays an important role in user behaviour and user engagement.

Regarding the frequency of use of ChatGPT, we did not find significant differences between the frequency of use of ChatGPT and user behaviour or user engagement and the statistical results can be found in Table 5.8. However, we noticed that the users with higher usage frequency tend to have a higher score in terms of the Focused Attention dimension of user engagement as shown in Figure 5.11, which means that the participants with a relatively higher frequency of ChatGPT usage feel time pass shower during the search in session in ChatGPT as compared to the participants with relatively lower ChatGPT usage frequency and it is probably because that the users with higher use frequency of use



(a) Boxplot of user engagement score by frequency

(b) Boxplot of user engagement Focused Attention dimensionsubscore by frequency

(c) Boxplot of user engagement Perceived Usability dimension subscore by frequency

(d) Boxplot of user engagement Aesthetic Appeal dimension subscore by frequency

(e) Boxplot of user engagement Reward dimension subscore by frequency

Figure 5.11: Boxplots of user engagement by frequency

of ChatGPT are more familiar with ChatGPT so that they are less likely to become frustrated during the interaction, which leads to more engaged interaction.

5.3.3. Previous experience with chatbots or digital assistants

Because ChatGPT is a new application, it is not relatively meaningful to measure the experience of users with ChatGPT, in this case, due to the similarity between ChatGPT and chatbots. we assume that the experience with chatbots may also facilitate ChatGPT usage.

Regarding the previous experience with chatbots or digital assistants, we did not find a significant difference in user behaviour between users with different levels of previous experience. But we observe that, for the session length aspect, the users with more experience with chatbots or digital assistants tend to have longer session length on average as shown in Figure 5.12, which is contrary to the finding in the [28] and they found the users with more experience is more likely to spend less time in the search. This difference may be caused by the different experimental settings. Since we encourage users to collect as much information as possible, the users with higher familiarity with ChatGPT may be less likely to give up and stop searching in the middle.

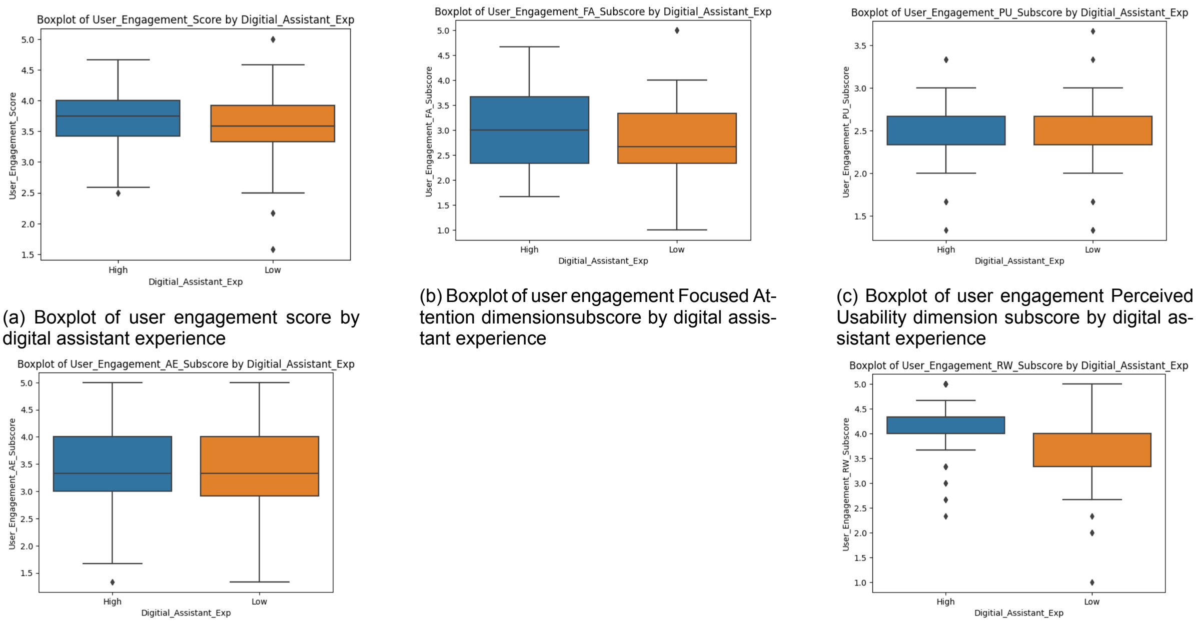
In terms of user engagement, the statistical test uncovered that there is a significant effect of Previous experience on user engagement of Reward. As displayed in Figure 5.12, the users with more experience tend to feel rewarded, which is similar to the finding regarding the influence of frequency.

Explanatory var.	Metric	p-value	Effect size	Result
				Continued on Next Page
ATI	Average of prompt length	0.302		Not significant
	Number of prompts	0.404		Not significant
	Number of distinct prompts	0.354		Not significant

Explanatory var.	Metric	p-value	Effect size	Result
	Prompt complexity	0.721		Not significant
	Session length	0.563		Not significant
	Average prompt duration	0.759		Not significant
Frequency of use of ChatGPT	Average of prompt length	0.191		Not significant
	Number of prompts	0.352		Not significant
	Number of distinct prompts	0.397		Not significant
	Prompt complexity	0.445		Not significant
	Session length	0.381		Not significant
	Average prompt duration	0.841		Not significant
Previous experience with chatbots or digital assistants	Average of prompt length	0.320		Not significant
	Number of prompts	0.177		Not significant
	Number of distinct prompts	0.219		Not significant
	Prompt complexity	0.212		Not significant
	Session length	0.026		Not significant
	Average prompt duration	0.352		Not significant
ATI	**User engagement score**	**0.000**	**0.091**	**Significant**
	Focused Attention dimension score	0.307		Not significant
	Perceived Usability dimension score	0.450		Not significant
	Aesthetic Appeal dimension score	**0.002**	**0.063**	**Significant**
	Reward dimension score	**0.000**	**0.133**	**Significant**
Frequency of use of ChatGPT	User engagement score	0.627		Not significant
	Focused Attention dimension score	0.046		Not significant
	Perceived Usability dimension score	0.076		Not significant
	Aesthetic Appeal dimension score	0.617		Not significant
	Reward dimension score	0.120		Not significant

Explanatory var.	Metric	p-value	Effect size	Result
Previous experience with chatbots or digital assistants	User engagement score	0.070		Not significant
	Focused Attention dimension score	0.235		Not significant
	Perceived Usability dimension score	0.655		Not significant
	Aesthetic Appeal dimension score	0.800		Not significant
	Reward dimension score	**0.000**	**0.066**	**Significant**

Table 5.8: Statistical results of exploratory variables



(a) Boxplot of user engagement score by digital assistant experience

(b) Boxplot of user engagement Focused Attention dimensionsubscore by digital assistant experience

(c) Boxplot of user engagement Perceived Usability dimension subscore by digital assistant experience

(d) Boxplot of user engagement Aesthetic Appeal dimension subscore by digital assistant experience

(e) Boxplot of user engagement Reward dimension subscore by digital assistant experience

Figure 5.12: Boxplots of user engagement by digital assistant experience

	ATI		Frequency		Digital Assistant Experience	
	Low	High	Low	High	Low	High
Average of prompt length	9.67±5.80	10.54±6.22	9.63±5.60	10.76±6.11	10.21±5.69	9.98±5.79
Number of prompts	3.99±3.50	3.18±2.38	3.70±2.81	3.45±3.04	4.13±3.46	3.22±2.52
Number of distinct prompts	3.79±3.38	3.01±2.36	3.51±2.69	3.30±2.92	3.90±3.31	3.06±2.34
Prompt complexity	10.25±1.36	10.25±1.39	10.22±1.63	10.35±1.14	10.41±1.51	10.26±1.23
Session length	240.84±207.54	212.26±186.91	239.96±191.48	215.88±194.73	259.48±211.07	198.69±173.69

Table 5.6: Mean and standard deviation of user user engagement metrics for participants in different ChatGPT expertise, topical expertise, and topical expertise levels

	ATI		Frequency		Digital Assistant Experience	
	Low	High	Low	High	Low	High
User Engagement Score	3.50±0.46	3.78±0.49	3.67±0.52	3.68±0.47	3.57±0.55	3.70±0.44
User_Engagement_FA_Subscore	2.79±0.71	2.94±0.77	3.02±0.76	2.82±0.79	2.79±0.81	2.94±0.74
User_Engagement_Perceived Usability dimension_Subscore	2.44±0.39	2.41±0.38	2.48±0.29	2.39±0.38	2.42±0.36	2.40±0.31
User_Engagement_Aesthetic Appeal dimension_Subscore	3.16±0.74	3.53±0.74	3.38±0.66	3.42±0.75	3.32±0.78	3.34±0.72
User_Engagement_Reward dimension_Subscore	3.77±0.60	4.16±0.59	3.92±0.69	4.07±0.58	3.82±0.72	4.12±0.53

Table 5.7: Mean and standard deviation of user user engagement metrics for participants in different ATI, frequency, and digital assistant experience levels

6 Discussion

In this study, to investigate how ChatGPT expertise, topical expertise, topical interest influence user behaviour and user engagement in search sessions in ChatGPT, we devised and carried find out the answers. In this Chapter, we discuss the implication and limitation based on the experiment result. Furthermore, we also go into the potential influence of exploratory variables to user behaviours and engagement in ChatGPT search.

6.1. Results Implications

6.1.1. RQ1 - User behaviour

Regarding the influence of ChatGPT expertise, topical expertise, and topical interest on user behaviour, , for the influence of ChatGPT expertise on user behaviour (**H1a**), the result shows that people with relatively higher ChatGPT expertise do not behave differently compared to people with relatively lower ChatGPT expertise. Regarding the result of boxplots, the difference between medians is not significant. Since the research relating to the ChatGPT expertise is limited, we looked in the works about the influence of search expertise on user behaviour and Meley [39] found that users with higher search expertise are more likely to use longer queries and more queries. However, the phenomenon is not obvious in our results. One potential reason is that the interaction pattern in ChatGPT differs from one in the Web search. Users may more rely on the conversations. Furthermore, the responses of ChatGPT typically are long summaries. Due to the wealth of information in the responses, the users need to ask fewer times. This result indicates that, as the technology gets measured, the dependency on application expertise may moderate.

For the hypothesis about the relationship between topical expertise and user (**H1b**), in terms of the number of distinct prompts, number of prompts, and session length per prompt, people with relatively higher topical expertise behave differently compared to people with relatively lower topical expertise. Our finding is similar to the result of White et al. [60] and they found that the users with relatively high topical expertise may use more queries in a web search session because the... However, there are different experimental settings in our research. In our experimental setting, we give concrete tasks to participants but [60] analysed the log information of qualified users without specific tasks. Besides, our finding is contrary to the result of Mao et al. [38] and they found that users with higher topical expertise tend to use fewer queries and shorter session length in the web search session because experts are able to finalize the task more efficiently. One of the reasons why the result is different is probably because of the task setting. We both have concrete tasks in the experiment but, in ours, we encourage users to seek as much as information for the task, which could be why the users tend to have more prompts according to the findings. Our findings contribute to uncovering the properties that may impact the manifest of topical expertise in the search task, which emphasizes the importance of task contexts in searching especially in environments like ChatGPT.

As mentioned in Chapter 5, we did not find the user behaviour difference between users with different topical interest levels (**H1c**). However, the boxplots show that users with higher topical interest may use more prompts in the ChatGPT search sessions. But, according to the research [15], they did not find this situation happened and the number of queries in different levels are similar. One possible reason is the nature of ChatGPT as an interactive tool, which may encourage users to conduct more conversations and satisfy their information needs and it is less likely to happen in the traditional search. This finding may be valuable to designing a more responsive search tool to cater for users with different topical interest levels.

6.1.2. RQ2 - User engagement

For research question 2, regarding how ChatGPT expertise influences user engagement (**H2a**), we found people with higher ChatGPT expertise show higher Focus Attention and Reward scores in the boxplots but the differences are not significant in the statistical tests. Smith et al. [55] found that the users with training for search show higher engagement, which shows the same trend as our result. However, because of the lack of statistical significance, the relationship could be less strong. If further research can confirm this relationship in ChatGPT, it is reasonable to customize the applications for users with different ChatGPT expertise levels.

For the relationship between topical expertise and engagement(**H2b**), we found that users with higher topical expertise show higher engagement scores particularly in Aesthetic Appeal on average according to the result of the box plot but not significant. However, due to the limited existing research about the relationship between topical expertise and user engagement, we are currently not able to compare this result with other works further. Regarding the relationship between topical interest and user engagement (**H2c**), as mentioned in Chapter 5, we did not find that people with different levels of topical interest exhibit different levels of engagement while searching in ChatGPT but the medians of Reward aspect for different topical expertise levels of users differ in boxplots. O'Brien et al. [45] found that there is a correlation between topical interest and Reward of user engagement.Edward et al. [15] found users with higher topical interest levels tend to have high Focus Attention levels in web searches. According to our result shown in the Chapter 5, we found a similar phenomenon from the boxplot but the statistical result is not significant, which is because some of the properties of ChatGPT may weaken the user engagement. For instance, the users have to wait for responses in ChatGPT, which may cause the user to lose focus during the interaction. As compared to the web search, the latency is not significant in web search most of the time. This result shows similar trends between web search and ChatGPT but it may also highlight the latency feature of ChatGPT as a search tool and it could moderate the influence of topical interest on user engagement.

6.1.3. Exploratory variables

In terms of exploratory variables in the research, for Affinity for Technology Interaction (ATI), we found that ChatGPT users present different levels of user engagement. There is also a significant difference between Aesthetic Appeal and Reward aspects. Based on the statistical test and box plots, this outcome reveals that the people with relatively higher ATI may be more engaged than people with relatively lower ATI and they will also feel more rewarded and time pass is slow as compared to one with relatively low ChatGPT expertise. Besides, we have not found existing research investigating the relationship between ATI and user engagement so far. But Liu et al. [37] found there is a correlation between the concentration and intention to use the technology in e-learning, which means that users tend to be more concentrated if they are more willing to use the technology. In light of this finding, the importance of considering ATI while designing a ChatGPT-like system can be emphasized due to the importance of ATI to user engagement.

Regarding the frequency of use of ChatGPT, as mentioned in Chapter 5, the participants with higher usage frequency showed higher scores in terms of the Focused Attention dimension of user engagement in the boxplot on average, the participants with a relatively higher frequency of ChatGPT usage feel time pass shower during the search in session in ChatGPT as compared to the participants with relatively lower ChatGPT usage frequency. As compared to the result from Jenkins et al. [28], they also found that the users could be easier to get lost in the search while focusing on the search tasks. And our findings indicate that the regularity of use of ChatGPT could be a critical factor in deep engagement.

For the digital experience, we found that the people with lower frequency tend to spend more time, which is probably because they need more time to get familiar with the usage of ChatGPT and, furthermore, they also feel more rewarding than people with high ChatGPT. Holscher et al. [25] found users with higher frequency in search tend to have longer sessions because of familiarity. Even if the difference is not significant the users with less frequency can quickly catch up. The main reason why the result is significant is probably that the influence of familiarity is significant in relatively short search sessions, which means user-friendliness may flatten the learning curve of ChatGPT. Furthermore, the result is also meaningful to the UX designers and they can leverage this insight to offer a more user-friendly interface to ensure all the users have a smooth experience.

6.2. Limitations

Even if the research and experiment design were discussed and analysed before starting the experiment, there are still some limitations not being realized:

6.2.1. Threshold selection

Threshold selection could be one of the limitations. We concluded that the ChatGPT expertise does not affect the user behaviour in search sessions in ChatGPT. However, this conclusion is based on the division by 37.5th and 62.5th percentiles and it is no guarantee that there is no significant difference after moving the thresholds further from the centre like the 25th and 75th due to less noise. However, we were not able to conduct this test because more data points could be excluded and the requirement of power analysis could not be satisfied. Furthermore, it is possible that the samples only represent a part of the population. Then it the conclusion can only represent this interval if the low/high is classified in this way instead of using an absolute threshold. Furthermore, there is also some limitation in the classification of topical expertise and topical interest. As shown in Chapter 4, there are different distributions for different topics. Because we expect ideally only 37.5th and 62.5th on the scale will be used. However, if the distribution is skewed, then more noise could involved, which could reduce the accuracy of the result.

6.2.2. Tool and measurement limitations

ChatGPT expertise could be also one of the limitations. Since there is no existing questionnaire, we adapted the technical expertise questionnaire from [67]. However, the efficacy of the adapted questionnaire was not verified because of limited time and budget. We hypothesized that ChatGPT expertise may have an impact on user behaviour and user engagement. However, the statistical results show that the ChatGPT expertise level may have affected the Reward in user engagement. If the efficacy of the ChatGPT expertise was verified before the experiment then it is possible to find more correlation between user behaviour and user engagement.

Furthermore, the measurement of topical interest could be also one of the limitations. In our research, we use a self-reported questionnaire with one question to measure the interest in each topic. But it could be better to also involve the implicit features to measure the topical interest.

6.2.3. Interface design limitation

The similarity of interface design is also important in our experiment. For our research, if the similarity between the mock product and the real one is not high enough, it could cause some negative effects on the measurement of dependent variables. For instance, the session length could get longer for users with high ChatGPT expertise, which could lead to a smaller difference between users with different levels of ChatGPT expertise. However, because the real ChatGPT application is a relatively mature business product, it is not easy to make an application which has high similarity with ChatGPT in a limited time.

7 Conclusion and future work

7.1. Conclusion

ChatGPT emerged as a significant application with great potential in the search domain. For this reason, this study investigates the impact of ChatGPT expertise, topical expertise, and topical interest on user behaviour and engagement in the search in ChatGPT during search sessions within ChatGPT. To address these questions, we designed the experiment and recruited participants from the crowdsourcing platform. Contrary to what was expected, our finding indicates that ChatGPT expertise, topical expertise, and topical interest do not influence user behaviour in the ChatGPT search sessions **(RQ1)**. Regarding user engagement, we found that the ChatGPT expertise has a partial influence on user engagement but the topical expertise and topical interest do not affect the user engagement in the search sessions in ChatGPT **(RQ2)**, which may emphasize the influence of the AI expertise on the user interactions. Moreover, our research also indicates that the Affinity for Technology Interaction (ATI) affects user engagement, which highlights the importance of understanding the psychological aspects in the era of AI. Additionally, our research indicates that experience with chatbots has a marginal but noticeable impact on user engagement. This finding may open a new avenue for exploring how the previous experience of technology shapes user interaction. However, there are still some limitations involved in the research as the selection of thresholds for high/low levels or the lack of validation of the ChatGPT expertise questionnaire, which may skew the accuracy of the results.

Even if there are some limitations in our research such as the selections of Low/High threshold, and ChatGPT questionnaire to be verified, our research provides valuable insights. For instance, it enhances our understanding of user behaviour and engagement in ChatGPT search sessions. Specifically, our research reveals how different levels of ChatGPT expertise, topical expertise, and topical interest affect the way that users interact with ChatGPT in search, which is important for researchers and developers aiming at enhancing the ChatGPT functionality. By integrating these insights, there is the possibility to enable ChatGPT functionalities to be more effective to enrich the search experience for diverse users.

7.2. Future work

Based on the insights into how the ChatGPT expertise, topical expertise, and topical interest impact user behaviour and engagement in ChatGPT search, future research can focus on these potential directions:

- **ChatGPT expertise questionnaire.** As mentioned in the Chatpter 6, a verified ChatGPT expertise questionnaire is still an open question. Due to the limitation of time and budget, we were not able to verify the effectiveness of the ChatGPT expertise questionnaire. However, due to the importance of ChatGPT expertise, coming out with a ChatGPT expertise questionnaire may be impactful.

- **User behaviour and engagement in multilingual languages.** In our experiment, the submissions in non-English were excluded because different languages may have their own habit so that the user behaviour. In this case, it is interesting to look into how the users behave in different languages, which may help us to understand the users' behaviours in the different languages better and based on the results, we may be able to customize ChatGPT in different languages.

- **ChatGPT Interface.** The interface design may also play an important role in a similar experiment. Due to the limitation of time, it is challenging to implement an interface that highly closely resembles the actual ChatGPT interface. There are open-source ChatGPT-like interfaces available on the Internet. However, the design goal of these available interfaces is not for HCI research, which means that there is a possibility of causing some data loss in the experiment due to the complex architecture.

- **ATI for personalized ChatGPT search.** Given the significant influence of ATI on user engagement, particularly in Aesthetic Appeal and Reward aspects, it becomes important to consider personalizing responses based on users' ATI levels. This approach is driven by the correlation between higher ATI and increased user engagement, which means tailoring interactions to match the technological affinity of users could enhance their experience in the ChatGPT search.

A
Informed Consent

Here is the informed consent leveraged at the start of the experiment below, if the participant agreed with the contents mentioned, then they are allowed to proceed. Otherwise, the experiment will be directed to the end.

Informed Consent

We are researchers for the Web-Information-Systems group at the Delft University of Technology and we aim to better understand how people interact with ChatGPT.

You will be first asked to provide demographic information (e.g., your age, gender, and highest level of education completed). Then you will be asked to answer some questions regarding the levels of your ChatGPT expertise, topical expertise, topical interest, and your experience. Afterwards, you will be directed to a ChatGPT API-based interactive interface, where you can chat with ChatGPT to satisfy your information needs. Finally, you will be asked to answer a questionnaire about what you searched in ChatGPT and how you feel about this experience.

In our study, we are especially interested in specific levels of ChatGPT expertise and will exclude participants after the first part if they do not have such a ChatGPT expertise. If that is the case, your submission will be automatically returned, and you will receive a partial payment of £0.10.

All data will be stored securely in a password-protected electronic format. We will not store any information that can be used to identify who you are (e.g., your IP address). Be aware that the data we gather with this task might be published in an anonymized form later. Such an anonymized data set would include the answers you provide in this study but no personal information (e.g., your Prolific ID), so the answers will not be traceable. Furthermore, you can choose to withdraw your data within three months of the experiment date.

Completing the survey will take about 10 minutes. Participation in this task is entirely voluntary, and you can withdraw anytime. To complete the survey you will have to answer all questions. This study has been approved by TU Delft's ethics committee. If you have any questions about this study, please contact xxx@student.tudelft.nl.

By selecting "yes" below, you confirm that you have read, understood, and consent to the above information.

B

ChatGPT Expertise Questionnaire

In this section, we exhibit the ChatGPT expertise questionnaire adapted from the technical expertise questionnaire [67]. This questionnaire contains 5 questions based on the 7 Likert scale was leveraged at the start of the experiment to measure the ChatGPT expertise.

ChatGPT Expertise Questionnaire

1. My technical interest with ChatGPT is ... (e.g. interest in the functionality, mechanism, strengths, and limitations of the ChatGPT)
2. My enthusiasm for technology of ChatGPT is ... (e.g. enthusiasm about the functionality, mechanism, and strengths of the ChatGPT)
3. My technical literacy with ChatGPT is .... (e.g. knowledge about formulating prompts effectively, knowledge about limitations of ChatGPT)
4. My ability in dealing with ChatGPT is ... (e.g. capability or confidence in using the technology)
5. My distrust in technology of ChatGPT is... (e.g. distrust against the functionality, mechanism, strengths, and limitations of the ChatGPT)

Bibliography

[1] Eytan Adar et al. “Why we search: visualizing and predicting user behavior”. In: *Proceedings of the 16th international conference on World Wide Web*. 2007, pp. 161–170.

[2] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. “Intent-aware query obfuscation for privacy protection in personalized web search”. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 2018, pp. 285–294.

[3] Wasi Uddin Ahmad, Md Masudur Rahman, and Hongning Wang. “Topic model based privacy protection in personalized web search”. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 2016, pp. 1025–1028.

[4] Doaa Alrefaei et al. “Using Eye Tracking to Measure User Engagement with a Decision Aid”. In: *International Conference on Human-Computer Interaction*. Springer. 2023, pp. 57–70.

[5] Myoung-a An and Sang-Lin Han. “Effects of experiential motivation and customer engagement on customer value creation: Analysis of psychological process in the experience-based retail environment”. In: *Journal of Business Research* 120 (2020), pp. 389–397.

[6] Ioannis Arapakis et al. “User engagement in online N ews: Under the scope of sentiment, interest, affect, and gaze”. In: *Journal of the Association for Information Science and Technology* 65.10 (2014), pp. 1988–2005.

[7] Arian Askari et al. “A Test Collection of Synthetic Documents for Training Rankers: ChatGPT vs. Human Experts”. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2023, pp. 5311–5315.

[8] Anne Aula and Klaus Nordhausen. “Modeling successful performance in web searching”. In: *Journal of the american society for information science and technology* 57.12 (2006), pp. 1678–1693.

[9] Shenghua Bao et al. “Optimizing Web Search Using Social Annotations”. In: *Proceedings of the 16th International Conference on World Wide Web*. WWW ’07. Banff, Alberta, Canada: Association for Computing Machinery, 2007, pp. 501–510. ISBN: 9781595936547. DOI: 10.1145/1242572.1242640. URL: https://doi.org/10.1145/1242572.1242640.

[10] Sonia Blachier. *Age structure – UNCTAD Handbook of Statistics 2022*. URL: https://hbs.unctad.org/age-structure/.

[11] Lingjiao Chen, Matei Zaharia, and James Zou. “How is ChatGPT’s behavior changing over time?” In: *arXiv preprint arXiv:2307.09009* (2023).

[12] Geoffrey M Currie. “Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy?” In: *Seminars in Nuclear Medicine*. Elsevier. 2023.

[13] Tim Draws et al. “This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 295–305.

[14] Geoffrey B Duggan and Stephen J Payne. “Knowledge in the head and on the web: Using topic expertise to aid search”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2008, pp. 39–48.

[15] Ashlee Edwards and Diane Kelly. “Engaged or frustrated? Disambiguating emotional state in search”. In: *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 2017, pp. 125–134.

[16] Carsten Eickhoff et al. “Lessons from the journey: a query log analysis of within-session learning”. In: *Proceedings of the 7th ACM international conference on Web search and data mining*. 2014, pp. 223–232.

[17] Franz Faul et al. “Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses”. In: *Behavior research methods* 41.4 (2009), pp. 1149–1160.

[18] Andy Field, Jeremy Miles, and Zoe Field. *Discovering statistics using R*. W. Ross MacDonald School Resource Services Library, 2017.

[19] Steve Fox et al. “Evaluating implicit measures to improve web search”. In: *ACM Transactions on Information Systems (TOIS)* 23.2 (2005), pp. 147–168.

[20] Thomas Franke, Christiane Attig, and Daniel Wessel. “A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale”. In: *International Journal of Human–Computer Interaction* 35.6 (2019), pp. 456–467.

[21] Luanne Freund and Elaine G Toms. “Enterprise search behaviour of software engineers”. In: *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*. 2006, pp. 645–646.

[22] Ujwal Gadiraju et al. “Analyzing knowledge gain of users in informational search sessions on the web”. In: *Proceedings of the 2018 conference on human information interaction & retrieval*. 2018, pp. 2–11.

[23] Helene A Hembrooke et al. “The effects of expertise and feedback on search term selection and subsequent learning”. In: *Journal of the American Society for Information Science and Technology* 56.8 (2005), pp. 861–871.

[24] Suzanne E Hidi and John A McLaren. “Motivational factors and writing: The role of topic interestingness”. In: *European journal of psychology of education* 6 (1991), pp. 187–197.

[25] Christoph Hölscher and Gerhard Strube. “Web search behavior of Internet experts and newbies”. In: *Computer networks* 33.1-6 (2000), pp. 337–346.

[26] Mark I Hwang and Ron G Thorn. “The effect of user engagement on system success: a meta-analytical integration of research findings”. In: *Information & management* 35.4 (1999), pp. 229–236.

[27] Sampath Jayarathna, Atish Patra, and Frank Shipman. “Mining user interest from search tasks and annotations”. In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2013, pp. 1849–1852.

[28] Christine Jenkins, Cynthia L Corritore, and Susan Wiedenbeck. “Patterns of information seeking on the Web: A qualitative study of domain expertise and Web expertise”. In: *IT & society* 1.3 (2003), pp. 64–89.

[29] Thorsten Joachims. “Optimizing search engines using clickthrough data”. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, pp. 133–142.

[30] Veronika Karnowski et al. “From incidental news exposure to news engagement. How perceptions of the news post and news usage patterns influence engagement with news articles encountered on Facebook”. In: *Computers in Human Behavior* 76 (2017), pp. 42–50.

[31] V Karunakaran and Amita Sharma. “User Engagement Analysis of E-Commerce Websites from the Perspective of Eye Tracking”. In: *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*. IEEE. 2022, pp. 1–6.

[32] Abhishek Kaushik and Gareth JF Jones. “Exploring current user web search behaviours in analysis tasks to be supported in conversational search”. In: *arXiv preprint arXiv:2104.04501* (2021).

[33] Diane Kelly and Colleen Cool. “The effects of topic familiarity on information search behavior”. In: *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*. 2002, pp. 74–75.

[34] Mounia Lalmas and Liangjie Hong. “Tutorial on metrics of user engagement: Applications to news, search and E-commerce”. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 2018, pp. 781–782.

[35] Mounia Lalmas, Heather O’Brien, and Elad Yom-Tov. *Measuring user engagement*. Springer Nature, 2022.

[36] Bulou Liu et al. “Conversational vs traditional: Comparing search behavior and outcome in legal case retrieval”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 1622–1626.

[37] Su-Houn Liu, Hsiu-Li Liao, and Cheng-Jun Peng. “Applying the technology acceptance model and flow theory to online e-learning users’ acceptance behavior”. In: *E-learning* 4.H6 (2005), H8.

[38] Jiaxin Mao et al. “How does domain expertise affect users’ search interaction and outcome in exploratory search?” In: *ACM Transactions on Information Systems (TOIS)* 36.4 (2018), pp. 1–30.

[39] Karen Markey. “Twenty-five years of end-user searching, Part 1: Research findings”. In: *Journal of the American Society for Information Science and Technology* 58.8 (2007), pp. 1071–1081.

[40] Mohamad Noorman Masrek et al. “User engagement and satisfaction: The case of web digital library”. In: *International Journal of Engineering and Technology (UAE)* 7.4 (2018), pp. 19–24.

[41] Nicolaas Matthijs and Filip Radlinski. “Personalizing web search using long term browsing history”. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011, pp. 25–34.

[42] David Maxwell and Claudia Hauff. “LogUI: Contemporary Logging Infrastructure for Web-Based Experiments”. In: *Advances in Information Retrieval (Proc. ECIR)*. 2021, pp. 525–530.

[43] Rishabh Mehrotra and Emine Yilmaz. “Terms, topics & tasks: Enhanced user modelling for better personalization”. In: *Proceedings of the 2015 international conference on the theory of information retrieval*. 2015, pp. 131–140.

[44] Masahiro Morita and Yoichi Shinoda. “Information filtering based on user behavior analysis and best match text retrieval”. In: *SIGIR’94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*. Springer. 1994, pp. 272–281.

[45] Heather L O’Brien, Jaime Arguello, and Rob Capra. “An empirical study of interest, task complexity, and search behaviour on user engagement”. In: *Information Processing & Management* 57.3 (2020), p. 102226.

[46] Heather L O’Brien, Paul Cairns, and Mark Hall. “A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form”. In: *International Journal of Human-Computer Studies* 112 (2018), pp. 28–39.

[47] Denis Parra and Peter Brusilovsky. “User-controllable personalization: A case study with SetFusion”. In: *International Journal of Human-Computer Studies* 78 (2015), pp. 43–67.

[48] Feng Qiu and Junghoo Cho. “Automatic identification of user interest for personalized search”. In: *Proceedings of the 15th international conference on World Wide Web*. 2006, pp. 727–736.

[49] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. “Improving worker engagement through conversational microtask crowdsourcing”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–12.

[50] Partha Pratim Ray. “ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope”. In: *Internet of Things and Cyber-Physical Systems* (2023).

[51] *reCAPTCHA v3*. URL: `https://developers.google.com/recaptcha/docs/v3#:~:text=By%20default%2C%20you%20can%20use%20a%20threshold%20of%200.5.`.

[52] Alexandre Salle et al. “Studying the effectiveness of conversational search refinement through user simulation”. In: *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43*. Springer. 2021, pp. 587–602.

[53] Paulus Insap Santosa, Kwok Kee Wei, and Hock Chuan Chan. “User involvement and user satisfaction with information-seeking activity”. In: *European Journal of Information Systems* 14.4 (2005), pp. 361–370.

[54] Phillip Schneider et al. “Investigating Conversational Search Behavior for Domain Exploration”. In: *European Conference on Information Retrieval*. Springer. 2023, pp. 608–616.

[55] Catherine L Smith. “Domain-independent search expertise: A description of procedural knowledge gained during guided instruction”. In: *Journal of the Association for Information Science and Technology* 66.7 (2015), pp. 1388–1405.

[56] Sofia Eleni Spatharioti et al. “Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment”. In: *arXiv preprint arXiv:2307.03744* (2023).

[57] Weiwei Sun et al. *Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent*. 2023. arXiv: `2304.09542 [cs.CL]`.

[58] Pertti Vakkari, Mikko Pennanen, and Sami Serola. “Changes of search terms and tactics while writing a research proposal: A longitudinal case study”. In: *Information processing & management* 39.3 (2003), pp. 445–463.

[59] Jules White et al. “A prompt pattern catalog to enhance prompt engineering with chatgpt”. In: *arXiv preprint arXiv:2302.11382* (2023).

[60] Ryen W White, Susan T Dumais, and Jaime Teevan. “Characterizing the influence of domain expertise on web search behavior”. In: *Proceedings of the second ACM international conference on web search and data mining*. 2009, pp. 132–141.

[61] Ryen W White, Susan T Dumais, and Jaime Teevan. “Characterizing the influence of domain expertise on web search behavior”. In: *Proceedings of the second ACM international conference on web search and data mining*. 2009, pp. 132–141.

[62] Ryen W White et al. “Enhancing personalized search by mining and modeling task behavior”. In: *Proceedings of the 22nd international conference on World Wide Web*. 2013, pp. 1411–1420.

[63] *World - The World Factbook*. URL: `https://www.cia.gov/the-world-factbook/countries/world/#people-and-society`.

[64] Tianyu Wu et al. “A brief overview of ChatGPT: The history, status quo and potential future development”. In: *IEEE/CAA Journal of Automatica Sinica* 10.5 (2023), pp. 1122–1136.

[65] Ruiyun Xu, Yue Feng, and Hailiang Chen. “Chatgpt vs. google: A comparative study of search performance and user experience”. In: *arXiv preprint arXiv:2307.01135* (2023).

[66] Mengdie Zhuang, Gianluca Demartini, and Elaine G Toms. “Understanding engagement through search behaviour”. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2017, pp. 1957–1966.

[67] Martina Ziefle and Anne Kathrin Schaar. “Technical expertise and its influence on the acceptance of future medical technologies: what is influencing what to which extent?” In: *HCI in Work and Learning, Life and Leisure: 6th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering, USAB 2010, Klagenfurt, Austria, November 4-5, 2010. Proceedings 6*. Springer. 2010, pp. 513–529.

[68] Guido Zuccon and Bevan Koopman. “Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness”. In: *arXiv preprint arXiv:2302.13793* (2023).