


5-2024

AN EMPIRICAL STUDY ON THE EFFICACY OF LLM-POWERED CHATBOTS IN BASIC INFORMATION RETRIEVAL TASKS

Naja Faysal

Follow this and additional works at: <https://scholarworks.lib.csusb.edu/etd>

 Part of the [Artificial Intelligence and Robotics Commons](#), [Databases and Information Systems Commons](#), [Graphic Design Commons](#), [Graphics and Human Computer Interfaces Commons](#), [Industrial and Product Design Commons](#), and the [Interactive Arts Commons](#)

Recommended Citation

Faysal, Naja, "AN EMPIRICAL STUDY ON THE EFFICACY OF LLM-POWERED CHATBOTS IN BASIC INFORMATION RETRIEVAL TASKS" (2024). *Electronic Theses, Projects, and Dissertations*. 1938. <https://scholarworks.lib.csusb.edu/etd/1938>

This Project is brought to you for free and open access by the Office of Graduate Studies at CSUSB ScholarWorks. It has been accepted for inclusion in Electronic Theses, Projects, and Dissertations by an authorized administrator of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.

AN EMPIRICAL STUDY ON THE EFFICACY OF LLM-POWERED CHATBOTS
IN BASIC INFORMATION RETRIEVAL TASKS

A Project
Presented to the
Faculty of
California State University,
San Bernardino

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
in
Information Systems and Technology:
Business Intelligence and Analytics

by
Naja M. Faysal
May 2024

AN EMPIRICAL STUDY ON THE EFFICACY OF LLM-POWERED CHATBOTS
IN BASIC INFORMATION RETRIEVAL TASKS

A Project
Presented to the
Faculty of
California State University,
San Bernardino

by
Naja M. Faysal
May 2024

Approved by:
Conrad Shayo, Committee Chair
Oluwatosin Ogundare, Committee Member

© 2024 Naja Faysal

ABSTRACT

The rise of conversational user interfaces (CUIs) powered by large language models (LLMs) is transforming human-computer interaction. This study evaluates the efficacy of LLM-powered chatbots, trained on website data, compared to browsing websites for finding information about organizations across diverse sectors. A within-subjects experiment with 165 participants was conducted, involving similar information retrieval (IR) tasks using both websites (GUIs) and chatbots (CUIs). The research questions are: (Q1) Which interface helps users find information faster: LLM chatbots or websites? (Q2) Which interface helps users find more accurate information: LLM chatbots or websites? The findings are: (Q1) Participants found information significantly faster using LLM-chatbots, Q2. Participants found more accurate information using LLM chatbots. The conclusions are: (Q1) LLM-chatbots are highly efficient, and (Q2). LLM chatbots are highly reliable for information lookup tasks. These findings highlight the potential of LLM-powered CUIs to revolutionize user experience and advocate for integrating advanced AI capabilities in future interface design. Future research should investigate the following: 1. LLM-chatbot interaction speed over time to measure efficiency, especially with more complex questions, 2. The precision of these models over larger knowledge bases and complex questions, 3. Improvements in chatbot's usability and its impact on user experience and human-computer interaction (HCI), and 4. Gauge user preference over prolonged interactions over more complex questions.

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF TABLES.....	vi
LIST OF FIGURES.....	vii
CHAPTER ONE: INTRODUCTION.....	1
Research Questions.....	4
Research Hypotheses	4
Limitations	5
CHAPTER TWO: LITERATURE REVIEW	6
Background	6
Rule-Based to LLM-Powered Chatbots.....	7
Evolution of NLP Through LLMs	8
Addressing Research Questions.....	9
CHAPTER THREE: RESEARCH METHODS.....	17
Introduction.....	17
Q1: Which Interface Helps Users Find Information Faster: CUI or GUI? .	18
Q2: Which Interface Helps Users Find More Accurate Information: CUI or GUI?	20
CHAPTER FOUR: DATA COLLECTION, ANALYSIS AND FINDINGS.....	24
Descriptive Analysis of Survey Participants	24
Hypothesis Testing.....	29
CHAPTER FIVE: DISCUSSION, CONCLUSION, AND AREAS FOR FURTHER STUDY	38

Q1: Which Interface Helps Users Find Information Faster: CUI or GUI? .	38
Q2: Which Interface Helps Users Find More Accurate Information: CUI or GUI?	40
APPENDIX A: IRB APPROVALS.....	42
APPENDIX B: ONLINE SURVEY	45
APPENDIX C: DATA SAMPLE	49
REFERENCES.....	52

LIST OF TABLES

Table 1 Completions	24
Table 2 Participants Age	26
Table 3 Tech Savviness	27
Table 4 Devices	28
Table 5 Group Statistics for Hypothesis #1	30
Table 6 Paired Samples Correlation for Hypothesis #1	31
Table 7 Paired Samples T-test, $\alpha=0.05$, $df = 164$ for Hypothesis #1	31
Table 8 One Tailed T-test, $\alpha=0.05$, $df = 328$ for Hypothesis #1	32
Table 9 Group Statistics for Hypothesis #2	34
Table 10 χ^2 Contingency Table for Hypothesis #2, $\alpha=0.05$, $df = 1$	35
Table 11 Expected Frequencies, $\alpha=0.05$, $df = 1$ for Hypothesis #2	35
Table 12 Chi-Square Points, $\alpha=0.05$, $df = 1$ for Hypothesis #2	36
Table 13 Chi-Square Tests for Hypothesis #2	36

LIST OF FIGURES

Figure 1 Completions Pie Chart.....	25
Figure 2 Participants Age Pie Chart.....	26
Figure 3 Tech Savviness Pie Chart	27
Figure 4 Devices Pie Chart	28
Figure 5 Relationship Between CUI Time and GUI Time	30
Figure 6 Relationship Between CUI Accuracy and GUI Accuracy	34

CHAPTER ONE: INTRODUCTION

Advancements in natural language processing (NLP) and machine learning (ML) models, mainly due to innovations like the ‘transformer architecture’ proposed by Vaswani et al., (2017) have enabled a new generation of sophisticated conversational agents powered by large language models (LLMs) (Bhayana, 2024). The capabilities and sophistication of these chatbots in handling human-like conversations allow them to tackle diverse applications (Hadi et al., 2023), (Wu et al., 2023), (Ai et al., 2023), (Chen et al., 2023a), (Vaswani et al., 2017), (Teubner et al., 2023). Information lookup or retrieval is one of these applications. Currently, people browse and interact with websites via the point-and-click Graphical User Interface (GUI). While it is widely accepted that GUIs are an upgrade to the Command Line Interface (CLI) of the early computers, GUIs have their limitations, especially when dealing with information-dense websites. Websites belonging to sectors like education, hospitality, healthcare, e-commerce, and financial services, among others, can contain large amounts of information, mainly text data, organized in menus and sub-menus, making the task of finding the information needed troublesome. Users who are not accustomed to the menu layout may struggle to navigate websites and locate information, particularly if they are unfamiliar with how the information is organized (Nguyen, et al, 2022). These sectors could benefit greatly from innovative technological interventions to improve their user experience (UX).

Although LLMs have demonstrated exceptional capabilities in text understanding, generation, and knowledge inference (Ai et al., 2023), it is unclear whether they can be used as an alternative to the GUIs.

It is important to differentiate between LLM-powered chatbots and traditional rule-based chatbots. Rule-based chatbots are popular among websites, especially those dealing with frequent customer service inquiries. These chatbots act as complementary to GUI websites, serve narrow tasks, aren't capable of handling complex conversations (Arz Von Straussenburg, 2023), and often end up connecting users to a human representative. LLM chatbots like ChatGPT 3.5 and 4, on the other hand, are major advancements in Artificial Intelligence (AI), notably for their capability in handling human-like conversations (Teubner et al., 2023). One of these applications is to fine-tune these models on specific organizational data. A fine-tuned model could then be used to power and transform an organization's traditional rule-based chatbot. Such chatbots could outperform traditional rule-based chatbots in existing tasks and extend their use to a variety of new tasks they cannot perform. This study envisions LLM-powered chatbots as an alternative user interface (UI) to the point-and-click GUIs - which is currently the primary interface of human-computer interaction (HCI). This new Conversational User Interface (CUI) can replace GUIs on many devices and software applications including websites. Users may find themselves conversing with hardware and software using natural language without the need for a keyboard, mouse, or screen. In this study, we assess the

efficacy of these chatbots in one of the most basic tasks of finding specific information about an organization.

A comparative study between an organization's CUI (chatbot) and its GUIs (website) in the task of information retrieval has significant implications. This knowledge can transform the fields of HCI, UI, and business operations, and can even have social implications, especially in education and healthcare. Although Nguyen et al, (2022) studied NLP-driven chatbots and menu-based interfaces, their findings are predominantly rooted in the era of rule-based systems. This culminating experience project acknowledges their foundational work but also emphasizes the transformative impact of LLMs, which became popular in late 2022 (after Nguyen, Sidorova, and Torres published their study). This project also serves as a natural extension, building on Nguyen et al, (2022), particularly focusing on the advanced capabilities of LLMs and their implications for user interaction and satisfaction, which is an evolution from the technologies discussed in their research.

Research Questions

Q1: Which interface helps users find information faster: LLM chatbots or websites?

Q2: Which interface helps users find more accurate information: LLM chatbots or websites?

Research Hypotheses

Time Efficiency

H_0 : There is no difference between LLM-chatbot and GUI in terms of speed.

$$H_0: \mu_{GUI\ time} = \mu_{CUI\ time}$$

H_1 : LLM-chatbot users find information faster than GUI users. $H_1: \mu_{GUI\ time} >$

$$\mu_{CUI\ time}$$

(Where $\mu_{GUI\ time}$ the mean time to complete tasks for $\mu_{CUI\ time}$ is the mean time to complete tasks for CUIs.)

Accuracy

H_0 : There is no difference between LLM-chatbot and GUI in terms of Accuracy.

$$H_0: p_{GUI\ accuracy} = p_{CUI\ accuracy}$$

H_1 : LLM-chatbot users find more precise information than GUI users. $H_1:$

$$p_{GUI\ accuracy} \neq p_{CUI\ accuracy}$$

(Where $p_{GUI\ accuracy}$ the proportion of correct responses for GUIs, and

$p_{CUI\ accuracy}$ is the proportion of correct responses for CUIs.

Limitations

Like all technologies, LLMs have their limitations. Issues like bias, toxicity, fairness, controllability (Naveed et al., 2023), lack of reasoning, limited domain knowledge, high resource needs, ethical concerns, robustness, and interpretability (Hadi et al., 2023) remain persistent challenges that need to be tackled. In addition, according to Caldarini et al., (2022), LLMs lack contextual understanding and emotional intelligence, and face difficulties in handling long conversations. Furthermore, LLMs have massive environmental impacts due to high computational needs (Hadi et al., 2023). Wide-spread adoption of LLMs comes with implications and open issues including the importance of prompt engineering, legal concerns regarding copyrighted training data, and potential impacts on education and research (Teubner et al., 2023). Other concerns were raised by Wu et al., (2023) regarding academic integrity, intellectual property, safety challenges, factual errors (or what is known as the hallucination problem), lack of explicit knowledge modeling, and high research costs are additional limitations (Wu et al., 2023). According to Xu et al., (2023), LLMs like ChatGPT may lead to over-reliance and generate or replicate misinformation, yielding inconsistent results. Finally, when it comes to response selection, Tao et al., (2021)'s survey found that there are still challenges in effectively modeling multi-turn conversations, ensuring logical consistency, and adapting models to shifting conversation domains that need further research.

CHAPTER TWO: LITERATURE REVIEW

Background

Chatbots, or Conversational Agents (CA), aren't new technology; they date back to the 1960s. However, advances in AI and NLP, along with the rise of messaging platforms, have made them more prominent recently (TONTTS, 2019). These bots are usually present on a website via a tiny pop icon located at the bottom right corner of the page, which either launches automatically or by a click, followed by asking for a prompt from the user (Khan and Walcott, 2023). Chatbots have been deployed in a diverse range of sectors like healthcare, education, entertainment, and e-commerce, with narrow tasks using rule-based approaches (Caldarini et al., 2022). Due to their versatility, chatbots came to handle a variety of tasks, including simulating, streamlining, and automating customer service tasks (Khan and Walcott, 2023). In addition, CUIs are being used for starting and controlling software applications, like in a voice-based video repository, in-car systems, and accessibility applications (Jaber and McMillan, 2020). Conversational User Interfaces (CUIs) are usually powered by a single or multi-modal chatbot that can receive input and return outputs in a variety of formats (text, audio, or imagery) interchangeably, Apple's Siri is a popular early example of a CUI (Jaber and McMillan, 2020).

Rule-Based to LLM-Powered Chatbots

Chatbots have evolved from early rule-based systems like ELIZA to more advanced systems using a suite of AI technologies (Caldarini et al., 2022). While the user interface of rule-based and LLM-based chatbots may appear similar, the technology behind the two is far apart. For instance, Arz Von Straussenburg, (2023) found that rule-based chatbots face challenges in understanding user intent and generating human-like responses, they are limited to specific tasks and follow decision tree-like paths. The study also distinguished that input processing in chatbots can be either rule-based or ML-based, shedding light on the fundamental differences in the technology's architecture. LLMs on the other hand, represent a major advancement in AI capabilities for generating human-like text (Teubner et al., 2023). They can be used to build chatbots that accept natural language prompts (Wei et al., 2023). OpenAI's ChatGPT, for instance, is one of these chatbots that utilize LLMs to power its ability to handle challenging language understanding, generation, and knowledge inference tasks in conversational format (Wu et al., 2023) and (Ai et al., 2023). It has shown impressive capabilities in NLP tasks including translation, summarization, and question answering (Hadi et al., 2023) and (Ogundare and Araya, 2023). In addition, LLMs have remarkable capabilities in instruction comprehension, commonsense reasoning, and human interaction (Huang et al., 2023). OpenAI's GPT-4, for example, performs close to human-level across a general array of

tasks and is considered an early form of artificial general intelligence (AGI) (Bubeck et al., 2023).

Evolution of NLP Through LLMs

Classical theories in computational linguistics like generative grammar and POS tagging laid the foundation for modern NLP accomplishments (Ogundare and Araya, 2023). Recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) were commonly used in NLP tasks before the rise of the transformer architecture, however, these models struggled with challenges like vanishing gradients and modeling long-range dependencies (Hochreiter et al., 2001). LLMs evolved through four major stages: 1. statistical models, 2. neural models, 3. contextual word embeddings, and 4. large-scale pre-training (Hadi et al., 2023). The transformer architecture (Vaswani et al., 2017) was a breakthrough in the field of NLP, where it got rid of recurrence and convolution in favor of attention mechanisms. The encoder-decoder architecture with multi-headed self-attention allows transformers to model long-range dependencies in text effectively (Devlin et al., 2018), and in doing so, made the models not only better at tasks like translation but also faster and more efficient to train. Modern LLMs are trained on a massive amount of text data using unsupervised learning to predict words based on context. They are then fine-tuned using reinforcement learning (RL) with human feedback on downstream tasks (Teubner et al., 2023) and (Hadi et al., 2023). Unsupervised learning objectives like masked language modeling (MLM) (Devlin et al., 2018) and causal language modeling (Radford et

al., 2018) allow the model to learn powerful representations of language. The pre-trained model was then fine-tuned using supervised learning and much smaller task-specific data sets. This transfer-learning approach allows the model to adapt to various end tasks (Howard and Ruder, 2018). Training these models involves data collection, data cleaning, architecture design, unsupervised pre-training, and supervised fine-tuning (Hadi et al., 2023). Advanced models like GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and Switch (Fedus et al., 2022), are driven by an increase in model size (e.g. GPT-3's 175B parameters) and utilize a mix of architectural innovations (e.g. sparse attention), training strategies (e.g. mixed precision), and various objectives (e.g. masked language modeling) to scale performance, while techniques like instruction tuning (e.g. Flan-T5) and human preference learning refine task-specific outcomes (Naveed et al., 2023).

Addressing Research Questions

Q1: Which Interface Helps Users Find Information Faster: CUI or GUI?

Few studies have touched on the speed of finding relevant information using LLMs such as ChatGPT compared to regular point-and-click browsing. Arz Von Straussenburg, (2023) found that LLMs are designed to process and understand natural language queries efficiently, which can significantly reduce the time it takes for users to find information. Unlike traditional search methods that may require navigating through multiple web pages or links, the same study found that LLM-powered chatbots can directly provide concise and relevant

responses to user queries. Furthermore, the same study found that LLMs offer advanced chatbot capabilities beyond existing NLP and ML techniques, excelling in generating human-like text and understanding unstructured data sources. The study suggests that LLMs could process and understand user queries more effectively than traditional search methods on websites and could generalize and adapt to specific tasks using a relatively small amount of task-specific data and transfer downstream tasks easily. This flexibility and efficiency in handling various tasks could lead to faster information retrieval compared to navigating through website interfaces (Arz Von Straussenburg, 2023). While LLM-powered chatbots may offer faster information retrieval in theory, it's important to consider factors such as the chatbot's training data, the complexity of the queries, and the specific implementation of the chatbot. These factors can influence the efficiency and effectiveness of the chatbot in providing quick responses (Arz Von Straussenburg, 2023).

Another study, (Xu et al., 2023), explored user behavior differences between using ChatGPT-like tools and Google Search-like tools for information-seeking tasks. It found that the group using a ChatGPT-like tool consistently spent less time on all tasks compared to the group using a Google Search-like tool, with no significant difference in overall task performance between the two groups. This suggests that ChatGPT can level user search performance across different education levels and is particularly effective for straightforward questions, providing information quickly and efficiently (Xu et al., 2023). Teubner

et al., (2023), on the other hand, emphasizes the human-like conversation capability of LLMs, which could enable them to understand user queries more naturally and provide relevant information more swiftly than websites, where users might need to navigate through multiple pages or menus. In addition, Ai et al., (2023) and Ogundare & Araya, (2023) found that LLMs facilitate generative retrieval and offer improved solutions for user understanding, model evaluation, and user-system interactions. This advantage indicates that LLM chatbots can provide more relevant and personalized responses quickly, enhancing the speed of information retrieval compared to the traditional web browsing experience.

EI-Ansari et al., (2023), found that the use of advanced LLMs in chatbots, such as BERT for generating responses, has been shown to be effective in understanding the context of conversations and providing relevant and personalized answers, showcasing the potential of LLM-powered chatbots in offering a more efficient information retrieval process compared to traditional search methods (EI-Ansari et al., 2023). Another paper (Hadi et al., 2023) concluded that AI-powered chatbots offer significant speed and efficiency advantages. The study found they can operate 24/7, handle large volumes of inquiries simultaneously, and provide quick and consistent responses. This efficiency is particularly beneficial in scenarios where scalability and rapid information retrieval are crucial. Additionally, as the study continues, these chatbots can perform routine tasks, answer common questions, and provide instant access to information. Importantly, AI-driven chatbots continuously learn

and improve from user interactions, becoming more accurate and efficient over time (Hadi et al., 2023).

Despite the promising theoretical advantages of LLM-powered chatbots, empirical evidence directly comparing the speed of information retrieval between these chatbots and traditional website interfaces remains limited. (Xu et al., 2023) suggested future research could explore other types of search tasks, understand user interactions with AI-powered conversational systems versus traditional search engines, investigate long-term effects on search behaviors and the search engine market, and examine the integration of chat and search functionalities to find an optimal balance between conversational and keyword-based approaches. This culminating experience project addresses this gap by providing quantitative data on the time efficiency of LLM chatbots versus websites in real-world tasks. This contributes to the understanding of their practical effectiveness in information retrieval and could inform the design of future user interfaces for enhanced speed and efficiency.

Q2: Which Interface Helps Users Find More Accurate Information: CUI or GUI?

The pursuit of accuracy in information retrieval is vital in determining the effectiveness of a user interface. This becomes very critical when users rely on these systems for sensitive information. Several studies looked at the accuracy of LLM-generated responses. Arz Von Straussenburg, (2023) suggested that combining the accuracy guarantees of rule-based chatbots with the

conversational capabilities of LLMs can significantly improve response precision and user satisfaction. This hybrid approach leverages LLMs to generate human-like responses while ensuring the information is accurate and reliable by relying on structured information from traditional chatbots (Arz von Straussenburg, 2023). In addition, Bubeck et al., (2023) discussed LLMs' tendency to generate errors or "hallucinations," which appear reasonable but are inaccurate. The study highlighted the challenge of identifying these errors without close inspection and the importance of review and quality assurance to ensure accuracy, especially in high-stakes domains (Bubeck et al., 2023). The same study evaluated the truthfulness of responses generated by LLMs using standard similarity metrics like ROUGE, BLEU, and BLEURT. The authors concluded that while GPT-4 responses were closer to "gold" answers than GPT-3, manual inspections are necessary for quality assurance (Bubeck et al., 2023). Khan (2023) studied the development of chatbots using conversational systems that aimed to provide more accurate responses to user queries. A deep learning model, specifically a retrieval-based chatbot model, was implemented to accept a set of queries and provide the most accurate responses possible. Khan's (2023) study on chatbots utilizing the Sequential model from Keras emphasizes their ability to enhance the accuracy and relevance of information, directly contributing to our examination of accurate information retrieval. By employing a response prediction mechanism that carefully selects the most appropriate answer based on probabilities, this approach ensures that user inquiries are matched with highly relevant responses. Xu et al., (2023) observed that ChatGPT 3.5 often aligned with the input query,

replicating inaccuracies in subsequent responses. For instance, when participants were asked to fact-check a statement about the 2009 United Nations Climate Change Conference, ChatGPT 3.5 sometimes replicated the inaccuracies of the prompt in its responses. However, it was also capable of providing the correct answer when the prompt was rephrased, demonstrating variability in its accuracy (Xu et al., 2023). The same research noted a significant overreliance on ChatGPT 3.5 by participants, with 70.8% of them in the ChatGPT 3.5 group taking the chatbot's responses at face value without further verification. This tendency contributed to lower accuracy levels in tasks that required critical evaluation of information. The study also highlighted a significant difference in performance on fact-checking tasks, with the ChatGPT 3.5 group performing worse than the Google Search group (average scores: 5.83 vs. 8.37, respectively). This suggests that Google Search may be more reliable for tasks requiring accuracy in information verification. Furthermore, the same researchers have observed that ChatGPT 3.5 provides inconsistent answers for the same prompt during multiple trials. Although it occasionally recognized statements as incorrect, it sometimes failed to provide accurate or complete information, affecting the overall accuracy of the information provided (Xu et al., 2023).

Accuracy in various contexts is paramount to the success of LLMs and chatbots, serving as a basis for comparing their performance against GUIs in delivering precise information. For example, Caldarini et al., (2022) emphasized the importance of improving chatbots' contextual and emotional understanding,

addressing gender biases, and enhancing human-chatbot interactions. It suggested that advancements in these areas are crucial for developing more engaging and user-friendly chatbots, marking them as key topics for future research. Hadi et al., (2023) emphasized exploring ethical considerations, improving data efficiency, enhancing interpretability, and addressing training data contamination from AI-generated content as key areas for future research. It also highlighted the importance of addressing bias and fairness in LLM applications. The emphasis on accuracy in different contexts for LLMs and chatbots presents an important area for research, suggesting that this empirical study could contribute to understanding how CUIs compare to GUIs in accuracy.

Research Gaps

Very few studies have compared LLMs to GUIs. Xu et al., (2023) compared ChatGPT with Google in search performance and user experience and found that participants using ChatGPT report better user experience in terms of usefulness, enjoyment, and satisfaction. The ChatGPT group, the study demonstrated, consistently spends less time on tasks compared to the Google Search group, with no significant difference in overall task performance. The study also found that users perceive ChatGPT's responses to have higher information quality compared to Google Search. In their comparative work on the evolution of LLMs, Ogundare and Araya (2023) found that ChatGPT demonstrated strong language generation abilities across major NLP tasks like machine translation, summarizing, question answering, and language generation compared to mainstream algorithms. While these studies are encouraging, they

don't specifically assess whether LLM chatbots outperform, or can potentially replace, the point-and-click GUIs in information lookup from websites of diverse sectors. Wang et al. (2022) called for future research to explore additional variables and outcomes related to chatbot use. Caldarini et al. (2022), recommended future work includes improved evaluation frameworks. While metrics for evaluating chatbots such as BLEU, METEOR, F-score, and perplexity are common, there is no agreed-upon standard, hence human evaluation is still necessary (Caldarini et al., 2022).

CHAPTER THREE: RESEARCH METHODS

Introduction

LLM's influence on the field of human-computer interaction (HCI) is yet to be felt due to the recent breakthroughs in LLM. The mass adoption of these models indicates that we are at the cusp of a new era in HCI. However, we still don't have empirical evidence of the efficacy of chatbots powered by these models and trained on domain-specific data compared to the traditional point-and-click GUIs in information lookup tasks. In this culminating experience project, we employ a within-subjects study design chosen for its effectiveness in controlling for individual differences among participants, thereby providing a more accurate comparison of the efficacy of LLM-powered chatbots versus traditional website browsing in information retrieval tasks. The project employed the following research methods to tackle each of the research questions: For speed of information retrieval (Q1), we measured the time participants took to find information using both interfaces, while for accuracy (Q2), we measured through the correctness of information participants retrieve, comparing results from chatbots and websites.

Q1: Which Interface Helps Users Find Information Faster: CUI or GUI?

The efficacy of information retrieval in terms of speed is a critical component of user experience in digital interfaces. In this project, we assessed and compared the time efficiency between LLM-powered chatbots (CUI) and traditional graphical user interfaces (GUI) in various sectors. The core objective was to determine which interface allows users to find information more rapidly (Hypotheses #1: LLM-chatbot users find information faster than GUI users).

Research Methods for Measuring Speed

Utilizing a "Within-Subjects" experimental design, we enlisted participants to engage with both interfaces, performing predefined tasks aimed at information lookup. This design is chosen for its robustness in controlling individual differences, as each participant acts as their own control. 165 human subjects participated.

The tasks were designed to mirror real-world information lookup scenarios that users might encounter. These tasks were balanced in complexity and depth across both interfaces to ensure fairness in the comparison. For instance, finding specific information on a university's website versus querying a chatbot trained on the same website's data. The tasks were crafted to require a few clicks but were not overly simplistic, mimicking a realistic search effort.

The primary metric for this question was the time taken by participants to complete each task. Participants were instructed to perform the tasks as quickly

and accurately as possible. A digital timer integrated into the online survey platform (Qualtrics) recorded the duration from the initiation to the completion of each task. This method allowed for a direct, quantitative comparison of speed between the GUI and CUI interfaces.

Participants began the experiment by consenting to participate through an informed consent page, ensuring ethical standards are met. Following consent, they were briefed on the study and the tasks ahead. The tasks were then presented in a randomized order to minimize order effects and bias. Upon completing a task, the digital timer automatically records the completion time.

Data on task completion times were collected seamlessly via the online survey platform and stored securely. This quantitative data was then extracted for statistical analysis, where we compared the average time taken to complete tasks via GUIs versus CUIs. The analysis aimed to identify significant differences in speed, providing empirical evidence to support one interface over the other in terms of efficiency.

Conclusion and Implications

This detailed methodological approach to measuring the speed of information retrieval yielded insights into the comparative efficiency of LLM-powered chatbots and traditional websites. By maintaining the integrity of the experimental design and ensuring a balanced and fair comparison, we aim to

contribute valuable empirical evidence to the discourse on HCI and the potential of LLMs in enhancing user experience through faster information retrieval.

Q2: Which Interface Helps Users Find More Accurate Information: CUI or GUI?

Accuracy in information retrieval is pivotal for assessing the effectiveness of user interfaces in delivering correct and relevant information upon request.

This segment of our methodology focuses on comparing the accuracy of information retrieved by participants when using LLM-powered chatbots versus traditional graphical user interfaces across diverse sectors (Hypotheses #2: LLM-chatbot users find more precise information than GUI users).

Research Methods for Measuring Accuracy

The study employed a "Within-Subjects" design, enlisting participants to interact with both LLM-powered chatbots and GUIs. This approach ensured each participant's experience with both interfaces can be directly compared, enhancing the reliability of our findings on accuracy. The sample size (165 participants) allowed room for calculating statistical significance while ensuring a broad demographic representation.

Accuracy was evaluated through tasks designed to simulate typical user interactions with both interfaces. Each task was crafted to require retrieval of specific information that can be quantitatively assessed for correctness.

The core metric for this question was the correctness of the information retrieved by participants. Tasks were designed with clear, objective answers, allowing responses to be evaluated against a predetermined set of correct answers. This quantitative measure provided a direct comparison of the interfaces' ability to aid users in finding accurate information.

Following consent and an initial briefing, participants were presented with a series of tasks, the order of which is randomized to control for sequence effects. They were instructed to seek specific pieces of information using both interfaces. Upon task completion, participants' responses were collected for accuracy assessment.

Responses were collected through the Qualtrics platform and analyzed for correctness. The analysis involved comparing the proportion of correct answers obtained through each interface, aiming to identify any statistically significant differences in accuracy. This comparison shed light on whether GUIs or CUIs were more effective in guiding users to accurate information.

Conclusion and Implications

This approach to assessing accuracy in information retrieval was designed to provide empirical evidence on the comparative effectiveness of LLM-powered chatbots and traditional websites. By systematically evaluating the correctness of

information retrieved through each interface, the study uncovered valuable insights into their respective strengths and weaknesses in facilitating accurate user interactions.

Limitations

It is important to acknowledge the potential limitations of the methodology. The first limitation is the lack of control over participants' environment and while the study attempts to minimize the impact of these limitations by providing detailed instructions, future studies should conduct similar studies with controlled environments. The order of the tasks also poses a challenge, if participants consistently start with one interface, their experience with the second might be influenced by their experience with the first. That is why the survey used a randomizer to make sure different participants start with a different interface test. Fatigue might affect the performance of participants. Some may take a long time to find the information and decide to quit or just randomly select from the options in the rest of the study. The study addressed that by simplifying and minimizing the number of tasks to two simple ones. In addition, the pre-test and post-test survey questions are also very short and simple, making them easy to complete. While the number of tasks is minimized to reduce fatigue, it is also a limitation because it puts a lot of pressure on the questions to be representative of the experience. For instance, there might be easier or harder information on a website to find. To mitigate this limitation, the study carefully selects two questions that are similar in nature and strike a balance between not being too

obvious and not being too hidden. The tasks also make reasonable sense, which actual users of the website are likely to be looking for. To eliminate guessing, the questions in the tasks are open-ended, so participants can't simply select an answer randomly. However, the answer is simple to find, mainly a number (a price or the number of guests), and it wouldn't impact fatigue. Finally, the websites used are too few to be considered a fair representation of GUIs. To minimize this limitation, the study employs x5 websites from x5 different sectors, but it is too small of a sample size of all websites on the internet. In addition, chatbots were built on the Botpress platform (botpress.com) and using the ChatGPT 3.5 model. The chatbots were only exposed to the website data without any further configuration, which may not be the ideal representation of the user experience of LLM chatbots. Further studies are needed to test a diverse number of websites and a diverse number of chatbots and platforms.

CHAPTER FOUR:
DATA COLLECTION, ANALYSIS AND FINDINGS

Descriptive Analysis of Survey Participants

Before delving into the hypothesis testing, this section provides a descriptive overview of the survey participants, setting the stage for a deeper understanding of the context in which the subsequent findings are situated. The survey and a sample of the data collected are available in Appendix (B).

Completions

This segment gives details on the completion rates of our study survey, looking at participants' engagement.

	Frequency	Percent
Not Finished	42	20.2 %
Finished	166	79.8 %
Total	208	100.0 %

Table 1 Completions

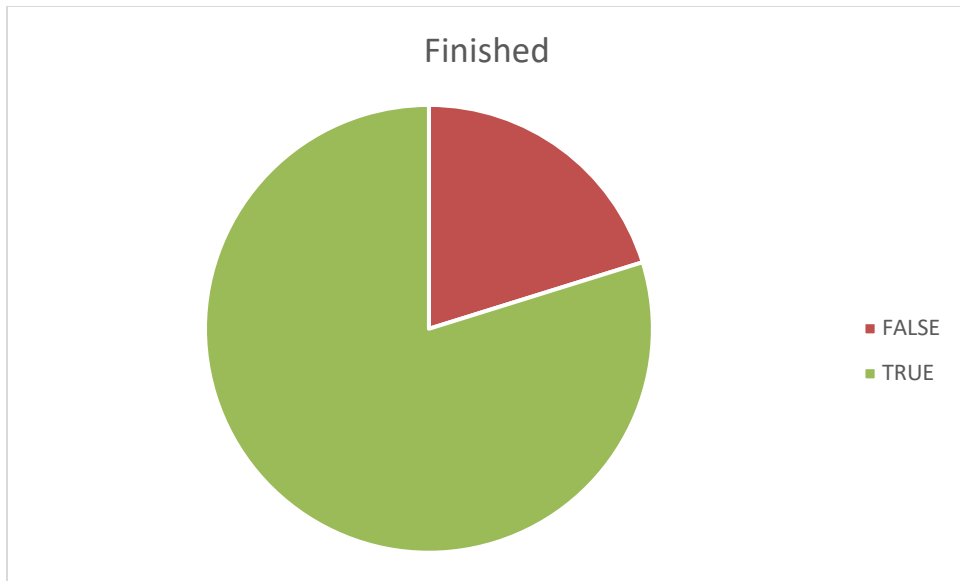


Figure 1 Completions Pie Chart

Table (1) and Figure (1) demonstrate the number of participants who completed the survey vs those who didn't. Uncompleted surveys were removed from future analysis. In addition, one blank response was also removed. Therefore, the total number of participants remaining is 165, and 43 responses were eliminated from future charts and tables.

Age

The following is the age distribution of our participants:

Age (mid-point)	Frequency	Percent
22	141	85.5 %
35	19	11.5 %
51	3	1.8 %
68	2	1.2 %

Table 2 Participants Age

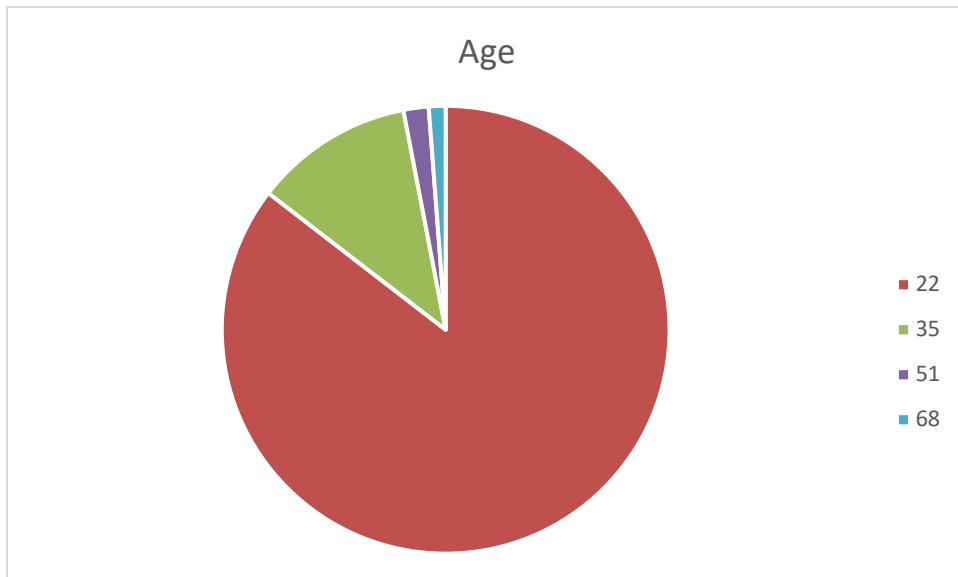


Figure 2 Participants Age Pie Chart

Table (2) and Figure (2) provide the participants' age groups, where the mid-point of an age range is used. Most of our participants (85%) belong to the age group of 18-26 years old followed by the age group 27-42 (11.5%).

Tech-Savviness

The following showcases participants' tech-savviness:

Valid	Frequency	Percent
Struggling	17	10.3 %
Manage	108	65.5 %
Savvy	40	24.2 %

Table 3 Tech Savviness

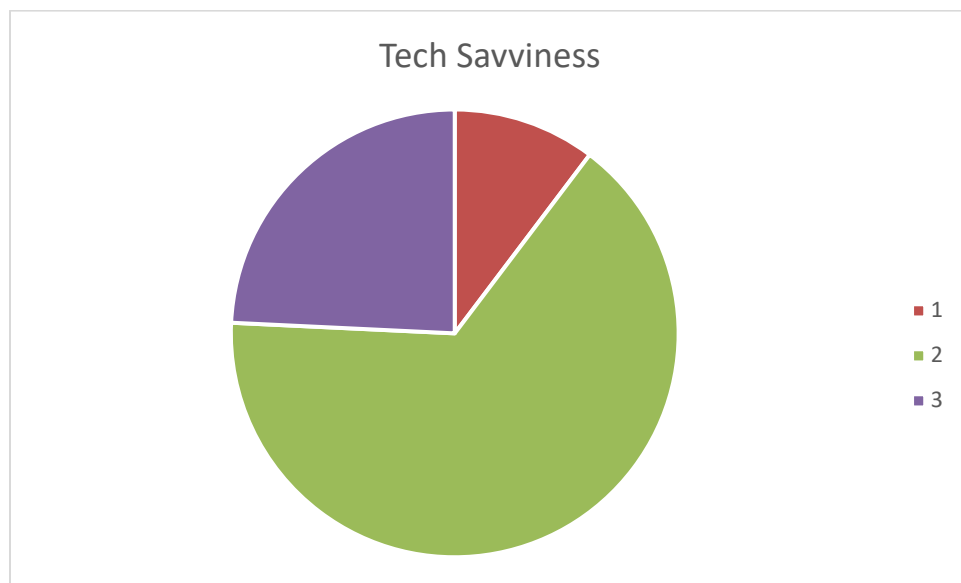


Figure 3 Tech Savviness Pie Chart

Table (3) and figure (3) demonstrate that most participants (65.5%) stated that they manage when it comes to technology; they aren't tech-savvy nor struggling with technology. 24% stated that they are indeed tech savvy, and 10.3% stated they struggle with technology.

Devices

Devices	Frequency	Percent
Desktop/Laptop	114	69.1 %
Mobile/Tablet	51	30.9 %

Table 4 Devices

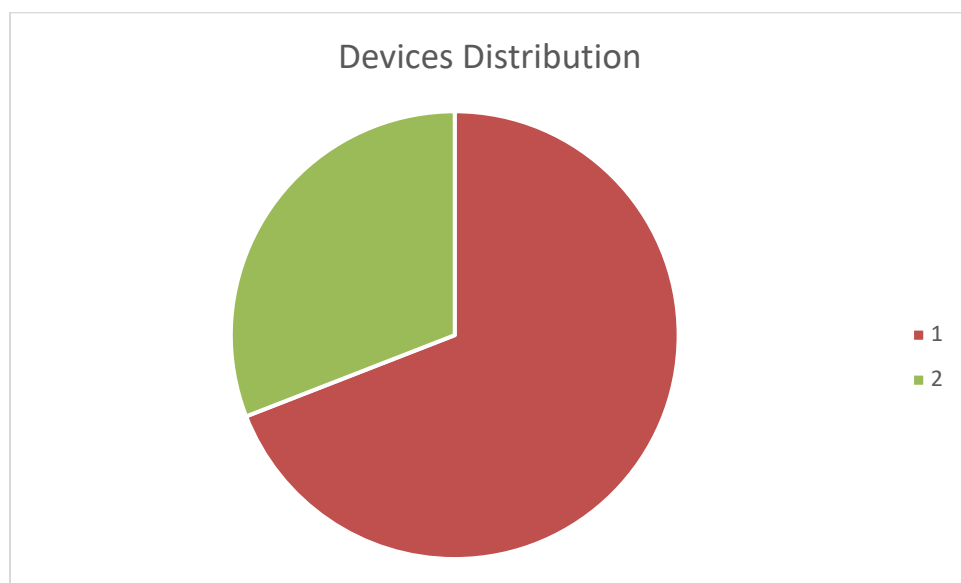


Figure 4 Devices Pie Chart

About sixty-nine percent of the participants (69.1%) took the survey using a desktop or laptop, whereas 30.9% took the survey using their mobile phone or tablet.

Hypothesis Testing

Hypothesis #1: Time Efficiency

- H_0 : There is no difference between LLM-chatbot and GUI in terms of speed.
- $H_0: \mu_{GUI\ Time} = \mu_{CUI\ Time}$
- H_1 : LLM-chatbot users find information faster than GUI users.
- $H_1: \mu_{GUI\ Time} > \mu_{CUI\ Time}$

(Where $\mu_{GUI\ Time}$ is the mean time to complete tasks for GUIs and $\mu_{CUI\ Time}$ is the mean time to complete tasks for CUIs.)

In Figure 5 below, we show that the relationship between CUI Time and GUI Time exhibits very little linearity. This project does not investigate the reason for this non-linear behavior, but further research should.

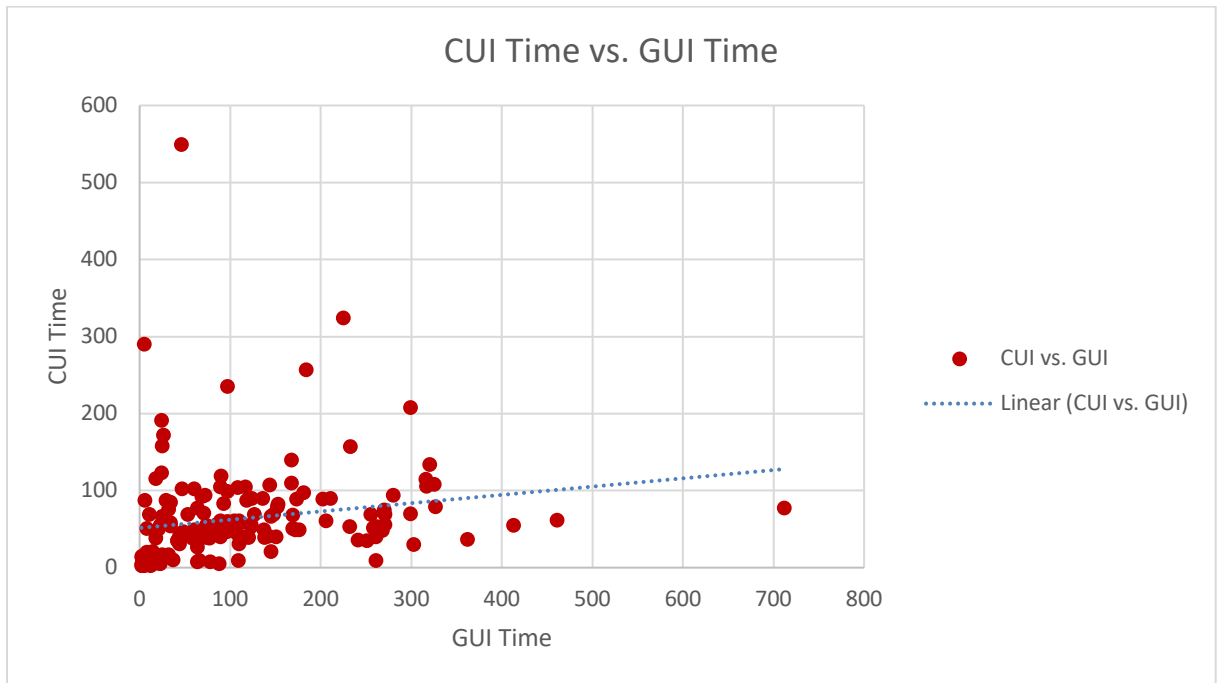


Figure 5 Relationship Between CUI Time and GUI Time

GROUP	N	Mean	Standard Deviation	Std. Error Mean
Experimental Group (CUI)	165	63.52	64.26	5.00
Control Group (GUI)	165	111.68	108.65	8.46

Table 5 Group Statistics for Hypothesis #1

		N	Correlation
Pair 1	CUI & GUI	165	0.0329

Table 6 Paired Samples Correlation for Hypothesis #1

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
GUI - CUI	48.16	115.76	9.012	30.36	65.95	5.3438	164	<0.0001

Table 7 Paired Samples T-test, $\alpha=0.05$, $df = 164$ for Hypothesis #1

The group statistics are presented in Table 5. The correlation between the paired samples is shown in Table 6. The calculated two-tailed p -value < 0.0001 is significant at the chosen level of significance, $\alpha=0.05$, as shown in Table 7.

Therefore, the difference between the distribution of CUI Time and GUI Time is statistically significant. Subsequently, we perform a one-tailed T-test to test the hypothesis in the suspected direction of the effect, and the result is shown in

Table 8 below:

GUI - CUI	T-test for Equality of Means						
	t	df	Sig. (1-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
Equal Variances Not Assumed	4.90	328	<0.0001	48.16	9.827	28.82	67.49

Table 8 One Tailed T-test, $\alpha=0.05$, $df = 328$ for Hypothesis #1

The results from the independent samples t-test, as shown in Table 8, statistically confirm the difference in time efficiency between the GUI and CUI interfaces. With a t-value of 4.90 and degrees of freedom (df) of 328, the significance level (Sig. 1-tailed) is less than 0.0001, firmly indicating a substantial difference. The mean difference in completion time is 48.16 seconds, with a standard error difference of 9.827 seconds. This statistical analysis further solidifies the evidence that the time taken to complete tasks using CUI is significantly less compared to GUI, within a 95% confidence interval ranging from 28.82 to 67.49 seconds.

Decision Rule: Reject null hypothesis if the p-value < 0.05 i.e., the absolute value of the calculated t-statistic is greater than the critical value. Otherwise, we fail to reject the null hypothesis. Since the p-value < 0.0001 which is significant at the chosen level of significance, $\alpha=0.05$, we REJECT the NULL hypothesis.

Hypothesis #2: Accuracy

- H_0 : There is no difference between LLM-chatbot and GUI in terms of Accuracy.
- $H_0 : p_{GUI\ accuracy} = p_{CUI\ accuracy}$
- H_1 : LLM-chatbot users' accuracy is different from that of GUI users.
- $H_1 : p_{GUI\ accuracy} \neq p_{CUI\ accuracy}$

(Where $p_{GUI\ accuracy}$ is the proportion of correct responses for GUIs and $p_{CUI\ accuracy}$ is the proportion of correct responses for CUIs.)

In Figure 6 below, we show how accurately users performed on the assigned IR tasks when using CUI compared to GUI. It appears that when users use CUI, they are more likely to satisfy the IR task objective. The reasons for the disparity of the distribution of accuracy over the two paradigms are not investigated in this research.

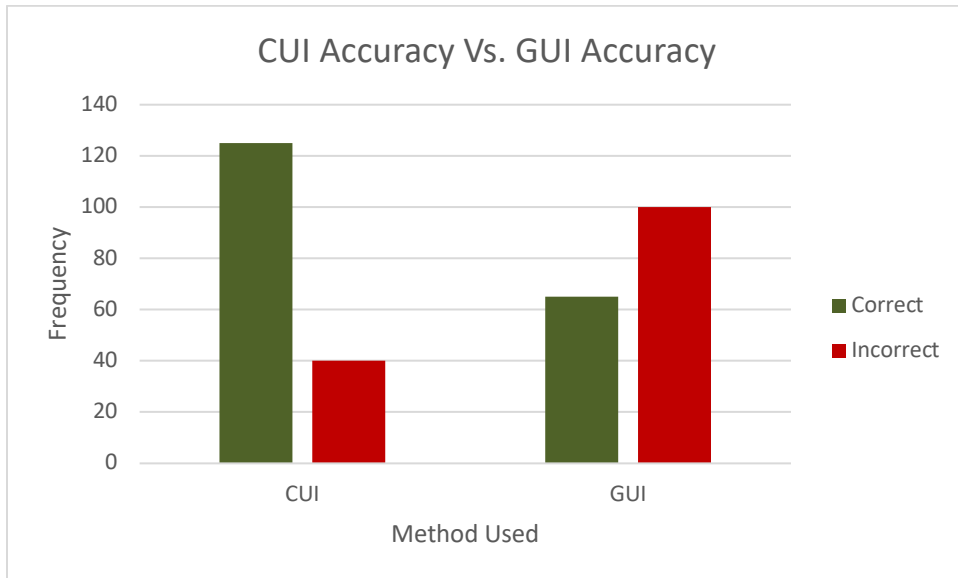


Figure 6 Relationship Between CUI Accuracy and GUI Accuracy

GROUP	N	Mean	Standard Deviation	Std. Error Mean
Experimental Group (CUI)	165	0.7576	0.43	0.033
Control Group (GUI)	165	0.3963	0.49	0.038

Table 9 Group Statistics for Hypothesis #2

Table 9 reveals group statistics, highlighting the mean accuracy rates for both CUI and GUI, with CUI showing a higher mean accuracy.

	CORRECT	INCORRECT	TOTAL
CUI	125	40	165
GUI	65	100	165
Total	190	140	

Table 10 χ^2 Contingency Table for Hypothesis #2, $\alpha=0.05$, $df = 1$

Table 10, a χ^2 contingency table, contrasts correct and incorrect responses between interfaces, indicating a higher correctness rate for CUI.

	CORRECT	INCORRECT
CUI	95	70
GUI	95	70

Table 11 Expected Frequencies, $\alpha=0.05$, $df = 1$ for Hypothesis #2

Expected frequencies in Table 11 support the differences in accuracy between CUI and GUI presented in Table 10.

	CORRECT	INCORRECT
CUI	9.474	12.857
GUI	9.474	12.857

Table 12 Chi-Square Points, $\alpha=0.05$, $df = 1$ for Hypothesis #2

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	44.6617	1	0.000
Continuity Correction ^b	43.1853	1	0.000
N of Valid Cases	165		
(b. Computed for a 2x2 table)			

Table 13 Chi-Square Tests for Hypothesis #2

Table 12 and Table 13 present χ^2 points and tests, respectively, confirming the significant difference in accuracy between CUI and GUI, as evidenced by a significant Pearson Chi-Square value.

Decision Rule: Reject the null hypothesis if the p-value < 0.05. Otherwise, we fail to reject the null hypothesis. Since the p-value < 0.0001, which is significant at the chosen level of significance, $\alpha=0.05$, we REJECT the NULL hypothesis, i.e., the difference in the distribution of the accuracy of CUI users and GUI users for identical Information Retrieval (IR) tasks is significant at the chosen level of significance, $\alpha=0.05$.

CHAPTER FIVE:

DISCUSSION, CONCLUSION, AND AREAS FOR FURTHER STUDY

This chapter discusses the results presented in Chapter 4, followed by a conclusion and areas for further study.

Q1: Which Interface Helps Users Find Information Faster: CUI or GUI?

Discussion

This section discusses how did CUI (chatbots) do compared to GUI (websites) in the IR tasks given to participants. Table 5 demonstrates the speed efficiency of information retrieval tasks performed using a Conversational User Interface (CUI) as compared to a Graphical User Interface (GUI). The mean time to complete tasks via the CUI was significantly lower ($M = 63.52$, $SD = 64.264$) than the GUI ($M = 111.68$, $SD = 108.654$), with a mean difference of 48.158 seconds (95% CI (30.363, 65.952)). The results of the t-test were highly significant ($p < .0001$), indicating that participants completed the tasks more quickly using the CUI than the GUI. This finding supports the research hypothesis that LLM-chatbot users find information faster than GUI users. These results suggest that the implementation of a Conversational User Interface powered by Large Language Models may contribute to enhanced time efficiency in information retrieval tasks across the tested sectors, which aligns with current literature advocating for the potential of LLM-powered interfaces to improve user experience in terms of speed and efficiency.

Conclusion

The results demonstrated a significant difference in speed, with CUIs enabling faster information retrieval than GUIs. This suggests that CUIs, powered by LLMs, are more efficient in processing user queries and providing rapid responses, a critical factor in enhancing user experience in digital interfaces.

Areas for Further Study

Future research directions could include:

- **Analysis of prolonged interactions:** Investigating the interaction speed of LLM-chatbots over time, especially with more complex questions.
- **Usability improvements:** Evaluating how improvements in chatbot usability affect the user experience. This encompasses examining the effects of user interface (UI) design improvements, multimodality of user's input and model's output, and interface intuitiveness on overall user satisfaction and engagement.
- **Long-term user preferences:** Gauging user preferences for LLM-chatbots over prolonged interactions and with more complex questions. This investigation would offer insights into how sustained use and the complexity of tasks affect user satisfaction and preference trends.

Q2: Which Interface Helps Users Find More Accurate Information: CUI or GUI?

Discussion

In evaluating the accuracy of information retrieval, our study finds a clear advantage for Conversational User Interfaces (CUIs) over Graphical User Interfaces (GUIs). Analysis revealed that participants using CUIs demonstrated a higher mean accuracy ($M = 0.7576$, $SD = 0.43$) compared to those using GUIs ($M = 0.3963$, $SD = 0.49$), with a statistically significant difference ($p < .0001$) supporting the superiority of LLM-chatbots in guiding users to correct information. This significant finding suggests that the advanced processing capabilities of LLM-powered chatbots enhance the precision of responses provided to users. By providing more accurate information, CUIs not only enhance user experience but also present a compelling case for their integration across various digital platforms, particularly in sectors where the precision of information is crucial.

Conclusion

The study revealed that CUIs were more accurate in providing correct information compared to GUIs. This superior performance of CUIs can be attributed to the advanced capabilities of LLMs in understanding and processing natural language, which is pivotal in sectors where precision of information is paramount. These findings underline the transformative potential of LLM-powered chatbots in revolutionizing the way information is accessed and interacted with across various sectors.

Areas for Further Study

Future research directions could include:

- **Precision at scale:** Exploring how LLM models maintain or improve their accuracy when dealing with larger knowledge bases and more complex questions.
- Investigating accuracy disparity: Future research could explore why accuracy varies between chatbots and websites, aiming to improve chatbot information processing.

APPENDIX A:
IRB APPROVALS

Dr. Oluwatosin Ogundare CITI Certificate



Completion Date 25-Oct-2023
Expiration Date 25-Oct-2026
Record ID 59262188

This is to certify that:

Oluwatosin Ogundare

Has completed the following CITI Program course:

Faculty/Staff/Outside Collaborators 2
(Curriculum Group)
Faculty/Staff/Outside Collaborators/Students
(Course Learner Group)
1 - Basic Course
(Stage)

Under requirements set by:

California State University, San Bernardino

Not valid for renewal of certification through CME.

CITI
Collaborative Institutional Training Initiative
101 NE 3rd Avenue, Suite 320
Fort Lauderdale, FL 33301 US
www.citiprogram.org

Verify at www.citiprogram.org/verify/?w209f8d27-791f-4a5f-b73e-cd030ca1e614-59262188

Naja Faysal CITI Certificate



Completion Date 05-Oct-2023
Expiration Date 05-Oct-2028
Record ID 58908563

This is to certify that:

Naja Faysal

Has completed the following CITI Program course:

Not valid for renewal of
certification through CME.

Human Research
(Curriculum Group)
Social Behavioral Research Investigators and Key Personnel
(Course Learner Group)
1 - Basic Course
(Stage)

Under requirements set by:

California State University, San Bernardino



101 NE 3rd Avenue, Suite 320
Fort Lauderdale, FL 33301 US
www.citiprogram.org

Verify at www.citiprogram.org/verify/?w170726c2-e13b-41bc-8505-0370e04100aa-58908563

IRB Approval Email

IRB-FY2024-70 - Initial: IRB Admin./Exempt Review Determination Letter

do-not-reply@cayuse.com <do-not-reply@cayuse.com>
To: 007607945@csusb.edu, Oluwatosin.Ogundare@csusb.edu

Tue, Oct 31, 2023 at 10:37 AM



October 31, 2023

CSUSB INSTITUTIONAL REVIEW BOARD
Administrative/Exempt Review Determination
Status: Determined Exempt
IRB-FY2024-70

Prof. Oluwatosin Ogundare and Mr. Naja Faysal
JHBC - Info & Decision Sci, ITS-Information Security
California State University, San Bernardino
5500 University Parkway
San Bernardino, California 92407

Dear Prof. Oluwatosin Ogundare and Mr. Naja Faysal:

Your application to use human subjects, titled "Comparative Study on the Efficacy of LLM-Chatbots in Information Retrieval Across Diverse Sectors" has been reviewed and determined exempt by the Chair of the Institutional Review Board (IRB) of CSU, San Bernardino. An exempt determination means your study had met the federal requirements for exempt status under 45 CFR 46.104. The CSUSB IRB has weighed the risks and benefits of the study to ensure the protection of human participants.

This approval notice does not replace any departmental or additional campus approvals which may be required including access to CSUSB campus facilities and affiliate campuses. Investigators should consider the changing COVID-19 circumstances based on current CDC, California Department of Public Health, and campus guidance and submit appropriate protocol modifications to the IRB as needed. CSUSB campus and affiliate health screenings should be completed for all campus human research related activities. Human research activities conducted at off-campus sites should follow CDC, California Department of Public Health, and local guidance. See CSUSB's [COVID-19 Prevention Plan](#) for more information regarding campus requirements.

You are required to notify the IRB of the following as mandated by the Office of Human Research Protections (OHRP) federal regulations 45 CFR 46 and CSUSB IRB policy. You can find the modification, renewal, unanticipated/adverse event, study closure forms in the Cayuse IRB System. Some instructions are provided on the [IRB Online Submission webpage](#) toward the bottom of the page. Failure to notify the IRB of the following requirements may result in disciplinary action. The Cayuse IRB system will notify you when your protocol is due for renewal. Ensure you file your protocol renewal and continuing review form through the Cayuse IRB system to keep your protocol current and active unless you have completed your study.

- Ensure your CITI Human Subjects Training is kept up-to-date and current throughout the study.
- Submit a protocol modification (change) if any changes (no matter how minor) are proposed in your study for review and approval by the IRB before being implemented in your study.
- Notify the IRB within 5 days of any unanticipated or adverse events are experienced by subjects during your research.
- Submit a study closure through the Cayuse IRB submission system once your study has ended.

If you have any questions regarding the IRB decision, please contact Michael Gillespie, the Research Compliance Officer. Mr. Michael Gillespie can be reached by phone at (909) 537-7588, by fax at (909) 537-7028, or by email at mgillesp@csusb.edu. Please include your application approval number IRB-FY2024-70 in all correspondence. Any complaints you receive from participants and/or others related to your research may be directed to Mr. Gillespie.

Best of luck with your research.

Sincerely,

King-To Yeung

King-To Yeung, Ph.D., IRB Chair
CSUSB Institutional Review Board

KY/MG

APPENDIX B:
ONLINE SURVEY

This research is conducted entirely through the following online survey:

Page #1: Informed Consent

University: California State University, San Bernardino

College: CSUSB's Jack Brown College of Business

Department: Information and Decision Sciences

Study: Comparative Study between Graphical User Interface vs. Conversational

User Interface

Researcher: Naja Faysal

Faculty Advisors: Dr. Oluwatosin Ogundare and Dr. Conrad Shayo

CSUSB IRB Approved Study: IRB-FY2024-70

Purpose: The objective of this research is to assess and compare the user experience and efficacy of Graphical User Interfaces (GUI) versus Conversational User Interfaces (CUI) powered by Large Language Models (LLMs) in information retrieval tasks within various sectors. Participants will engage with both websites (GUI) and chatbots (CUI) to complete two specific tasks and provide feedback.

Duration: Participation in this study will take approximately 15 minutes.

Confidentiality & Anonymity: All responses to this study are entirely anonymous. Optional email data for a prize draw is stored confidentially and will solely be used for the raffle draw.

Voluntary Participation: Your involvement in this study is entirely voluntary. You have the right to withdraw your participation at any time without any repercussions.

Risks & Benefits: There are no foreseeable risks associated with this study, as it involves tasks that you would typically perform on your computer or smartphone.

Participants will also have a chance to enter a draw to win a \$100 Amazon gift card.

Contact Information: For questions or additional information, please contact Naja Faysal (007607945@coyote.csusb.edu).

Consent: To participate in this study, please confirm the following statement:

"I'm over 18 years old, have read the description of the study, and voluntarily agree to participate. I understand that I can withdraw at any time without any repercussions." By clicking "I Agree," you provide your consent to participate in this research.

Page #2: Pre-Task Questions

How comfortable are you with technology?

- "I'm a tech-savvy."
- "I manage, but I'm not an expert."
- "I often struggle with technology."

What device are you using right now?

- Desktop/Laptop
- Smart Phone/Tablet

What's your age group?

- 18-26
- 27-42

- 43-58

- 59-77

Page #3: Instructions

- Ensure a stable internet connection for the tasks.
- Perform tasks in a quiet spot without distractions.
- For the website task, navigate to the site's home page and search for answers.

Do not use any other tool.

- For the chatbot task, ask the chatbot for an answer. Do not use any other tool.
- Complete all tasks in one go, ensure your responses are accurate, and try to finish as fast as possible.

Page #4: Tasks

GUI Task (Sample):

Please visit the Yaamava Resort Website and find out which restaurant(s) offer a breakfast menu at 8 a.m. on weekdays (Mon-Thurs). Copy & paste or type the name(s) of the restaurants below:

CUI Task (Sample):

Visit the Yaamava Resort Chatbot and find out the price of pool cabana packages on weekdays (Mon-Thurs). Copy & paste or type the price below:

Page #5: Post-Task Questions

Which interface did you find easier to complete the tasks?

- Website
- Chatbot

- Both were similar

Which interface did you enjoy and/or prefer using to complete such tasks?

- Website

- Chatbot

- Both are similar

Would you like to enter a draw for the chance to win a USD 100 gift card? If so,
please enter your email address below:














Page #6: Thank you note








Your response has been recorded, thank you for your participation.

APPENDIX C:
DATA SAMPLE

🔧 Id	📊 tech_sav	🎮 device	🔧 age
1	3	1	22
2	3	1	22
3	2	2	35
4	2	1	22
5	3	2	22
6	2	1	22
7	2	1	22
8	2	1	22
9	1	1	22
10	2	1	22
11	2	1	22
12	2	2	22
13	2	1	22
14	2	2	22
15	2	1	22
16	2	2	22
17	1	2	22
18	3	2	22
19	2	1	22
20	1	2	22

🔧 gui_time	🔧 hospGUI_time	🔧 ecomGUI_time	🔧 hlthGUI_time	🔧 eduGUI_time	🔧 finGUI_time
413	.	413	.	.	.
151	.	.	151	.	.
109	109
202	202
2	.	2	.	.	.
712	.	712	.	.	.
461	461
68	.	.	.	68	.
327	327
87	87
264	.	.	264	.	.
76	76
261	.	.	.	261	.
77	77
18	.	.	18	.	.
74	74
5	.	.	.	5	.
325	.	.	325	.	.
90	.	90	.	.	.
96	96

 cui_time	 hospCUI_time	 ecomCUI_time	 finCUI_time	 hlthCUI_time	 eduCUI_time	
55	.	55	.	.	.	
40	.	.	.	40	.	
9	.	.	9	.	.	
89	.	.	89	.	.	
4	.	4	.	.	.	
77	.	77	.	.	.	
62	62	
41	41	
79	.	.	79	.	.	
43	43	
52	.	.	.	52	.	
49	.	.	49	.	.	
40	40	
38	.	.	38	.	.	
38	.	.	.	38	.	
39	.	.	39	.	.	
290	290	
108	.	.	.	108	.	
119	.	119	.	.	.	
47	.	.	47	.	.	
 gui_accuracy	 hospGUI	 hlthGUI	 eduGUI	 finGUI	 ecomGUI	 cui_accuracy
1	1	1
0	.	0	.	.	.	0
0	.	.	.	0	.	1
0	.	.	.	0	.	1
0	0	0
0	0	1
1	1	1
1	.	.	1	.	.	1
0	.	.	.	0	.	1
1	1	1
1	.	1	.	.	.	1
0	.	.	.	0	.	1
0	.	.	0	.	.	1
0	.	.	.	0	.	1
0	.	0	.	.	.	0
0	.	.	.	0	.	1
0	.	.	0	.	.	0
0	.	0	.	.	.	1
0	0	0
1	.	.	.	1	.	1

 hospCUI	 ecomCUI	 finCUI	 hlthCUI	 eduCUI	 cog_load	 cus_sat
.	1	.	.	.	2	2
.	.	.	0	.	3	3
.	.	1	.	.	2	2
.	.	1	.	.	2	2
.	0	.	.	.	3	2
.	1	.	.	.	1	1
1	2	2
.	.	.	.	1	2	2
.	.	1	.	.	3	3
1	2	2
.	.	.	1	.	1	1
.	.	1	.	.	2	2
.	.	.	.	1	2	3
.	.	1	.	.	2	2
.	.	.	0	.	1	1
.	.	1	.	.	2	2
.	.	.	.	0	1	1
.	.	.	1	.	2	2
.	0	.	.	.	2	2
.	.	1	.	.	2	2

REFERENCES

- (Bhayana, 2024) Bhayana, R. (2024). Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. *Radiology*, 310(1), e232756.
- (Nguyen et al., 2022) Nguyen, Q. N., Sidorova, A., & Torres, R. (2022). User interactions with chatbot interfaces vs. Menu-based interfaces: An empirical study. *Computers in Human Behavior*, 128, 107093.
- (Ai et al., 2023) Ai, Q., Bai, T., Cao, Z., Chang, Y., Chen, J., Chen, Z., Cheng, Z., Dong, S., Dou, Z., Feng, F., et al. (2023). Information retrieval meets large language models: A strategic report from Chinese IR community. *AI Open*, 4:80–90.
- (Arz von Straussenburg, 2023) Arz von Straussenburg, A. F. (2023). Towards hybrid architectures: Integrating large language models in informative chatbots.
- (Au Yeung et al., 2023) Au Yeung, J., Kraljevic, Z., Luintel, A., Balston, A., Idowu, E., Dobson, R. J., and Teo, J. T. (2023). AI chatbots not yet ready for clinical use. *Frontiers in Digital Health*, 5:60.
- (Brown et al., 2020) Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

- (Bubeck et al., 2023) Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4.
- (Caldarini et al., 2022) Caldarini, G., Jaf, S., and McGarry, K. (2022). A literature survey of recent advances in chatbots. *Information*, 13(1):41.
- (Chen et al., 2023a) Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., Pu, Y., Lei, Y., Chen, X., Wang, X., Lian, D., and Chen, E. (2023a). When large language models meet personalization: Perspectives of challenges and opportunities.
- (Chen et al., 2023b) Chen, T., Gascó-Hernandez, M., and Esteve, M. (2023b). The adoption and implementation of artificial intelligence chatbots in public organizations: Evidence from US state governments. *The American Review of Public Administration*, page 02750740231200522.
- (Chowdhery et al., 2022) Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). PALM: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.
- (Devlin et al., 2018) Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

- (El-Ansari and Beni-Hssane, 2023) El-Ansari, A. and Beni-Hssane, A. (2023). Sentiment analysis for personalized chatbots in e-commerce applications. *Wireless Personal Communications*, 129(3):1623–1644.
- (Fedus et al., 2022) Fedus, W., Zoph, B., and Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270.
- (Følstad and Skjuve, 2019) Følstad, A. and Skjuve, M. (2019). Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st international conference on conversational user interfaces*, pages 1–9.
- (Hadi et al., 2023) Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M., Akhtar, N., Wu, J., and Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*.
- (Hochreiter et al., 2001) Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., et al. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.
- (Howard and Ruder, 2018) Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- (Huang et al., 2023) Huang, X., Lian, J., Lei, Y., Yao, J., Lian, D., and Xie, X. (2023). Recommender AI agent: Integrating large language models for interactive recommendations.

- (Jaber and McMillan, 2020) Jaber, R. and McMillan, D. (2020). Conversational user interfaces on mobile devices: Survey. In Proceedings of the 2nd Conference on Conversational User Interfaces, pages 1–11.
- (Khan and Walcott, 2023) Khan, H. A. and Walcott, T. H. (2023). The role of chatbots in Industry 4.0. In 2023 International Conference on Computing, Electronics & Communications Engineering (iCCECE), pages 85–88. IEEE.
- (Ma et al., 2021) Ma, Z., Dou, Z., Zhu, Y., Zhong, H., and Wen, J.-R. (2021). One chatbot per person: Creating personalized chatbots based on implicit user profiles. In Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, pages 555–564.
- (Naveed et al., 2023) Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2023). A comprehensive overview of large language models.
- (Nguyen et al., 2021) Nguyen, T.-P., Razniewski, S., and Weikum, G. (2021). Advanced semantics for commonsense knowledge extraction. In Proceedings of the Web Conference 2021, pages 2636–2647.
- (Nicolescu and Tudorache, 2022) Nicolescu, L. and Tudorache, M. T. (2022). Human-computer interaction in customer service: the experience with AI chatbots—a systematic literature review. *Electronics*, 11(10):1579.

(Ogundare and Araya, 2023) Ogundare, O. and Araya, G. Q. (2023).

Comparative analysis of ChatGPT and the evolution of language models.

arXiv preprint arXiv:2304.02468.

(OpenAI, 2023) OpenAI (2023). GPT-4 technical report.

(Radford et al., 2018) Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et

al. (2018). Improving language understanding by generative pre-training.

(Tao et al., 2021) Tao, C., Feng, J., Yan, R., Wu, W., and Jiang, D. (2021). A

survey on response selection for retrieval-based dialogues. In IJCAI,

pages 4619–4626.

(Teubner et al., 2023) Teubner, T., Flath, C. M., Weinhardt, C., van der Aalst, W.,

and Hinz, O. (2023). Welcome to the era of ChatGPT et al., the prospects

of large language models. Business & Information Systems Engineering,

65(2):95–101.

(TONTS, 2019) TONTS, S. (2019). Chatbots, will they ever be ready? Pragmatic

shortcomings in communication with chatbots.

(Vaswani et al., 2017) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J.,

Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is

all you need. Advances in neural information processing systems, 30.

(Wang et al., 2022) Wang, X., Lin, X., and Shao, B. (2022). How does artificial

intelligence create business agility? Evidence from chatbots. International

Journal of Information Management, 66:102535.

(Wei et al., 2023) Wei, J., Kim, S., Jung, H., and Kim, Y.-H. (2023). Leveraging large language models to power chatbots for collecting user self-reported data.

(Wu et al., 2023) Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., and Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo, and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

(Xiao, 2022) Xiao, Y. (2022). A transformer-based attention flow model for intelligent question and answering chatbot. In 2022 14th International Conference on Computer Research and Development (ICCRD), pages 167–170. IEEE.

(Xu et al., 2023) Xu, R., Feng, Y., and Chen, H. (2023). ChatGPT vs. Google: A comparative study of search performance and user experience.