MIJST
MIST International Journal of Science and Technology

System Usability and Design Evaluation of AI Chatbots: A Comparative Analysis of ChatGPT, Google Bard, and Bing Chat

Sumaiya Nuha Mustafina[1], Nusrat Kaniz Khan[2], Muhammad Nazrul Islam[*3], Fatema Siddiqua Nusrat[4], and M. Akhtaruzzaman[5]

[1]Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh
[2-5]Department of Computer Science and Engineering, Military Institute of Science and Technology, Dhaka, Bangladesh

Corresponding Email: *nazrul@cse.mist.ac.bd*

ARTICLE INFO

Article History:

Received: 26th March 2025
Revised: 05th June 2025
Accepted: 16th June 2025
Published: 30th June 2025

Keywords:

SUS
System Usability Score
HE
Heuristic Evaluation
HCI
Human Computer Interaction

ABSTRACT

Artificial intelligence (AI) has brought significant advancements in technology while the chatbots like ChatGPT, Google Bard, and Bing Chat are some of its remarkable innovations. These chatbots are helping users with diverse backgrounds by generating ideas, providing resources, and overall knowledge management. We acknowledge that these chatbots are still in their experimental stages of use. Evaluating the usability and user experience of chatbots becomes crucial to make them more usable, accessible, and intuitive to end users around the globe. Thus, the objectives of this research are to make a comparative usability analysis of AI-generated chatbots: Google Bard, ChatGPT, and Bing Chat. To achieve these goals, firstly, the System Usability Score (SUS) through questionnaire surveys and secondly, Heuristic Evaluation (HE) through expert observation were used. Through HE, we investigated characteristics of design, user engagement, and some other specific usability lacking along with a severity score that suggests both urgent and gradual usability improvement action. As an outcome, this study found that the SUS evaluation provided a comprehensive view of user satisfaction. Google Bard and Bing Chat received lower SUS scores, while ChatGPT demonstrated comparatively better usability, with a SUS score above 70. Again, a comparative usability analysis of AI-generated chatbots (ChatGPT, Google Bard and Bing Chat) reveals that, while all these applications suffer from a notable number of usability problems, ChatGPT demonstrates better usability performance compared to Google Bard and Bing Chat.

1. INTRODUCTION

People have used conventional agents in personal smart devices such as Siri, Google Assistant, and Alexa while Artificial Intelligence (AI) has reached its grand success in present days by introducing chatbots as a free and paid website version, mobile application, and integrated with other applications (Kundu, Kabir, & Islam, 2020). AI has introduced natural conversational systems, commonly known as chatbots, as a model for human-computer interaction, enabling conversations with humans (Adamopoulou & Moussiades, 2020). These chatbots utilize Natural Language Processing (NLP) and sentiment analysis to comprehend and respond to human language effectively. They are also called artificial conversational entities, interactive agents, smart bots, and digital assistants (Adamopoulou & Moussiades, 2020).

The pervasive acceptance of mobile internet and messaging platforms has resulted in the heightened utilization of chatbots. These chatbots are now capable of engaging with a significant segment of the global online community through popular mobile messaging applications, including Facebook Messenger, WeChat, Skype, Telegram, Slack, Viber, and Kik, which collectively boast approximately 3 billion users worldwide (Brandtzaeg & Folstad, 2017). Again, AI-generated chatbots can be used by a wide range of users, including customers, patients, students, employees, researchers, etc. for example, a customer might use a chatbot to get help with setting up a new Wi-Fi router; a patient might use a chatbot to schedule an appointment with their doctor or to get treatment-related information; a student might use a chatbot to get help with a math problem or to learn about a new topic in history; an employee might use a

chatbot to file a vacation request or to get help with a technical issue, etc.

Given the increasing user base of chatbots, spanning various age groups from youngsters to senior citizens, chatbots must offer a highly usable user interface. Chat- bots should also be able to handle a wide range of requests and questions. The central objective of using chatbots revolves around productivity, characterized by delivering rapid and accurate responses. Research indicates that interactions with chatbots tend to be lengthier compared to interactions with human strangers (Brandtzaeg & Folstad, 2017; Jain et al., 2018).

Users favored chatbots that either had a human-like capacity to converse in regular language or offered a fun experience that made use of the advantages of the well- known turn-based messaging interface. Xu (Xu et al., 2017), revealed that user interactions with customer service chatbots are mostly driven by a desire for emotional engagement rather than purely seeking information. Failing to establish this emotional connection can impede the chatbot's effectiveness in fulfilling user needs. Development of an effective chatbot is required to address specific design considerations to facilitate easy comprehension of their functional features, particularly for new, unfamiliar, and disabled users. Moreover, they should incorporate appealing features aligned with user requirements, employing an intelligent approach.

Chatbots represent a novel technological advancement, primarily embraced by innovators and early adopters, who may have distinct needs and preferences compared to the broader population (Brandtzaeg & Folstad, 2017). Examples of recent transformative NLP models are ChatGPT by OpenAI and Google's counterpart-Bard, and Microsoft Bing Chat. These generative language models exhibit the capacity to produce responses that resemble human communication when presented with open-ended prompts, encompassing questions, statements, or prompts about academic subject matter (Fuchs, 2023). Critique should be directed towards these chatbots with regards to their user interfaces, input and out- put functionalities, user satisfaction, usability, as well as their responses to identical situations.

Usability testing is crucial in Chatbots as it ensures that these AI-powered conversational agents are intuitive, understandable, and easy to use for a diverse range of end-users. Usability is defined as the ability of a software application to help users achieve their goals efficiently, effectively, and with satisfaction (International Organization for Standardization, 2018). And, in the context of Chatbots, usability is particularly important because they interact with users through natural language, making it essential to ensure that they can understand and respond appropriately to user queries. The importance of usability testing in Chatbots is also highlighted by the fact that usability is a key quality assessment for any software application, regardless of its background. However, explicitly focusing on Chatbots usability assessment, considering design issues, has not been conducted at a mass level before. This underscores the need for usability testing in Chatbots to ensure that they meet the diverse needs of end-users with different demographic profiles. As usability testing can help to identify the usability issues in the chatbots, increase user satisfaction, improve user experience and enhance accuracy in response.

Therefore, the objective of this study is to evaluate the usability of various AI-generated chatbots, compare different chatbots based on their usability, and suggest possible solutions to enhance the usability and user experience of chatbot applications. Acknowledging that the utilization of these chatbots remains in an experimental phase, the usability evaluation of ChatGPT, Google Bard, and Bing Chat may contribute to improving the effectiveness, efficiency, and universality of such applications. To attain these objectives, Heuristic Evaluation (HE) (Nielsen, 1995) through expert observation and the System Usability Score (SUS) (McLellan, Muddimer, & Peres, 2012) through questionnaire were employed. As such the contributions of the research are highlighted as:

- A comprehensive usability evaluation of three widely used AI chatbots including ChatGPT, Google Bard, and Bing Chat using the System Usability Scale (SUS) and expert-based evaluation through Nielsen's Heuristic Evaluation (HE).
- A dual-method approach that strengthens the reliability and depth of usability analysis in conversational agent research.
- Identification of key design flaws, interaction challenges, and user experience issues specific to each chatbot interface.
- Actionable recommendations and insights to guide the improvement of AI chatbot user interfaces and enhance overall user satisfaction.

1.1 Related Work

The ChatGPT system is an AI-based chatbot capable of producing text in various styles, including formal, informal, and creative writing (Shidiq, 2023). In (Liu et al., 2023), a study was conducted to assess the reliability of ChatGPT in simulating standardized patients for clinical training and education. Ten patient histories were compiled and reviewed by senior physicians to verify the accuracy of ChatGPT's generated information. In another study, Hill Yardin (Hill-Yardin et al., 2023) highlighted the role of ChatGPT, focusing on the future of scientific publishing. Additional studies have explored the uses of ChatGPT in public health, global warming, and problem-solving domains (Rane et al., 2023; Rane, 2023).

A recent evaluation of Google Bard on Vietnamese high school biology examinations was performed (Nguyen, 2023). Comparisons of accuracy between AI-based chatbots like ChatGPT and Google Bard have also been conducted (Ram & Verma, 2023). Bing Chat assists users with creative endeavors, such as writing poems, essays, or songs, and creating graphics from text (Zdnet, 2023). In summary, research has flourished on chatbots' performance in academic writing, medical diagnosis, creative writing, and accuracy comparison. Therefore, chatbots are becoming increasingly relevant to people from diverse backgrounds.

In recent years, usability evaluation of diverse applications has gained significant recognition due to the global necessity of human-computer interaction (HCI). For instance, a

usability evaluation of a pregnancy tracker was performed in (Kundu, Kabir, & Islam, 2020). Usability tests on ride-sharing applications have become common; for example, heuristic and semiotic evaluations of truck hiring mobile applications were addressed in (Muaz, Islam, & Islam, 2021). In (Tasfia et al., 2023), HE and SUS evaluations were conducted for children's AR-based learning applications. Comparative web and mobile user studies were conducted in (Munim et al., 2020). The adoption of HCI and usability concepts for fourth industrial revolution applications was studied in (Munim et al., 2020). In (Hossainet al., 2020), the user experience for the Multichain (Blockchain) platform was introduced. This body of research underscores the importance of incorporating HCI in end-user applications.

Usability evaluation, while significant in the field of chatbot user interfaces, has not been extensively explored on a large scale. As advanced systems evolve, the importance of capabilities and user requirements in system development processes becomes increasingly evident. The literature suggests that combining usability evaluation with chatbot user interfaces can open a new, specific research domain. There is a significant need for studies that focus on the ease of use and challenges of chatbots, especially at these experimental stages. Therefore, this research concentrates on evaluating the usability of the three most widely used chatbots for end users. Again, the chatbots selected for this study are still in their beta stage, and there is a noticeable lack of research on usability measures for these chatbots. This research gap underscores the need for thorough usability evaluations to ensure the reliability of chatbots for a broad range of end-users over time. To address this gap, this study aims to evaluate the usability of chatbots through heuristic evaluation (HE) and the System Usability Scale (SUS). By employing these evaluation methods, the study seeks to enhance chatbot performance in terms of user satisfaction.

1.2 Theoretical Background

This section covers the background studies of the usability methods and a brief introduction to the three chatbots. Contemporary investigations into usability evaluation aim to identify cost-effective methods that yield favorable outcomes for both users and developers within an increasingly competitive industry (Hvannberg, Law, & Lérusdóttir, 2007).

Usability problems should be handled by combining problem sets from multiple evaluators to identify unique problems, and probable severity as well as to propose design solutions. The prevalent methods employed for usability evaluation include surveys/questionnaires, HE, SUS, usability metrics, cognitive walkthrough, automated software-based evaluation, focus groups, eye tracking, cognitive task analysis, semiotic inspection methodology, and simplified pluralistic walk-through, and others (McLellan, Muddimer, & Peres, 2012),(Munim et al., 2020),(Khairat, Priyadi, & Adrian, 2022). Among these, HE and SUS methods can be employed to evaluate the usability of the selected chatbots which are discussed in the following subsections.

1.2.1 Heuristic Evaluation

Heuristic evaluation (HE) is a usability inspection technique for software that aids in detecting usability issues with the creation of the user interface. It entails evaluators particularly looking at the interface and determining whether or not it complies with accepted usability rules. There are a few heuristics that focus on user cognition rather than guide rules (Hvannberg, Law, & Lérusdóttir, 2007). Nielsen's 10 Heuristics, Shneiderman's 8 Golden Rule, Tognazzini's 16 Principles, and Gerhardt Powals Rules are some of the invented heuristic methods (Hvannberg, Law, & Lérusdóttir, 2007). Heuristic evaluation involves utilizing different sets of heuristics, some of which have overlapping criteria, including consistency, task relevance, visual presentation, user control, reducing cognitive load, effective error management, and offering guidance and support (Folmer & Bosch, 2004). Nielsen's heuristics have been developed through empirical studies across diverse contexts that can quickly identify the problems of a user interface (Mack & Nielsen, 1993) (Nielsen, 1995). In (Langevin, 2021), Nielsen conducted an assessment of 249 usability issues and distilled them into a concise set of ten guidelines (see table 1). These guidelines, known as "Rules of Thumb," hold significance in the field of Human-Computer Interaction (HCI) (Nielsen, 1995). Heuristic evaluation is conducted generally by 3 to 5 heuristics and usability experts assign a severity rating of 0 to 4; where 0 indicates no usability problem and 4 denotes a catastrophic usability issue (Kundu, Kabir, & Islam, 2020).

1.2.2 System Usability Scale (SUS)

The System Usability Scale (SUS) is a survey scale intended for rapid and straightforward evaluation of the usability of a particular product or service. The SUS method consists of 10 survey questions (see Table 2) on a 5-point scale that can provide a good usability reflection of the selected system (HubSpot, 2018). The SUS generates final scores on a scale of 0 to 100, with higher scores reflecting enhanced usability.

Table 1: Nielsen's ten heuristics

Index	Heuristics
H1	Visibility of system status
H2	Match between the system and the real world
H3	User control and freedom
H4	Consistency and standards
H5	Error prevention
H6	Recognition rather than recall
H7	Flexibility and efficiency of use
H8	Aesthetic and minimalist design
H9	Help users recognize, diagnose, and recover from errors
H10	Help and documentation

Table 2: System Usability Scale (SUS) questionnaire

Index	SUS Questionnaires
1	I think that I would like to use this system frequently.
2	I found the system unnecessarily complex.
3	I thought the system was easy to use.
4	I think that I would need the support of a technical person to be able to use this system.
5	I found the various functions in this system were well integrated.
6	I thought there was too much inconsistency in this system.
7	I would imagine that most people would learn to use this system very quickly.
8	I found the system very cumbersome to use.
9	I felt very confident using the system.
10	I needed to learn a lot of things before I could get going with this system.

It's important to exercise caution when scoring the survey because the statements within it alternate between positive and negative aspects. Brooke (Brooke, 1996), emphasized that individual item scores are not meaningful by themselves. However, practitioners often analyze individual statements to gain insights into specific interface aspects. Hence, practitioners should examine these scores closely to provide informed insights in such cases (Brooke, 1996). SUS scores can be calculated using the following equation (Public Affairs, 2023).

$$\text{SUS Score} = (X + Y) \times 2.5 \quad (1)$$

where, X = Sum of the points for all odd-numbered questions – 5 and Y = 25 – Sum of the points for all even-numbered questions.

A SUS score of 68 serves as the minimal accepted score. Increment from this average can provide a good assessment of the overall usability of the design solution and decrement results in poor usability assessment. Typically, a score below 50 signifies an unacceptable level of usability. Exceptional usability is considered as a score of 80 or above. Further, one-way ANOVA can be performed using group-level summary statistics such as means, variances, and sample sizes. This method allows estimation of between-group and within-group variability to assess statistical significance (McLellan, Muddimer, & Peres, 2012).

1.2.3 Chatbots

Computer programs called chatbots are made to mimic human users in communication. They frequently work in customer service settings where they can give details and respond to inquiries. Additionally, chatbots can be employed for fun activities like storytelling or gameplay. Advancements in the chatbot field have been progressing rapidly at an extraordinary pace and are applicable in a wider range of situations. Open AI's ChatGPT, Google's Bard, Microsoft's Bing Chat, and Elon Musk's TruthCPT (Kasinathan, 2023) are rival chatbots in today's era (Rudolph, Tan, & Tan, 2023). In this research, the following three most used chatbots are selected for usability evaluation.

ChatGPT: OpenAI debuted ChatGPT, a chatbot. It is constructed on top of the GPT-3 family of big language models from OpenAI and is modified using supervised and reinforcement learning methods (Hill-Yardin et al., 2023). There has been considerable excitement surrounding ChatGPT following its launching (Rudolph, Tan, & Tan, 2023). The integration of natural language processing (NLP) models in healthcare holds great promise for improving access to medical information. Large language models (LLMs), such as ChatGPT based on GPT-3.5, have shown their ability to understand and generate human-like text efficiently through a two-stage training process. ChatGPT, in particular, has gained popularity and potential in medical education and clinical decision support (Johnson et al., 2023). The ChatGPT system is an AI-based chatbot that can produce text in a variety of styles, including formal, informal, and creative writing. Composing creatively including composing poetry, short tales, novels, and other genres of writing is made very simple by the Chat-GPT system's capacity to comprehend human language (Shidiq, 2023).

Google Bard: Google Bard is a large language model (LLM) chatbot created by Google AI. It can produce text of human quality, translate languages, write many types of creative content, and provide users with helpful answers because it was trained on a sizable dataset of text and code (Rahaman et al., 2023). A conversational AI chatbot that can produce any type of text, Google Bard is similar to ChatGPT. As long as it doesn't contravene its content restrictions, users are free to ask Bard any question, and it will respond. Bard is a much more capable AI assistant than Google Assistant, even though it hasn't yet taken its position (Martindale, 2023). Bard's full capabilities are yet to be disclosed, and its range of tasks remains uncertain. However, it is anticipated that this chatbot will present a significant competition to OpenAI's ChatGPT, which has rapidly gained popularity with over 100 million users during just two months of public testing (Ram & Verma, 2023).

Bing Chat: Microsoft is actively integrating Bing Chat powered by GPT- 4 and a GPT-based Copilot embedded within Microsoft 365. Microsoft promotes its Copilot in Word feature as a tool that provides users with a first draft to edit and iterate on, leading to substantial time savings in writing, sourcing, and editing tasks (Rudolph, Tan, & Tan, 2023). With the Bing Chat, a user can ask the AI chatbot questions and receive thorough, believable responses with footnotes linking to the sources and current data. The chatbot may assist with the user's creative endeavors as well, like writing a poem, essay, or song as well as creating graphics from text utilizing the same platform's Bing Image Creator (Zdnet, 2023).

2. MATERIALS AND METHODS

The methodological outline of the research is presented in Figure 1. The study incorporates both user-centered

evaluation (System usability scale) and expert evaluation (Heuristic evaluation) to obtain a holistic view of the chatbots' usability. A comparative analysis is performed to obtain the research objectives. In heuristic evaluation (HE), three usability experts evaluated each of the selected chatbots. They identified the usability problems of the chatbots and assigned severity ratings for each of the problems. The usability problems and severity ratings were then aggregated to get a comprehensive understanding of the usability issues found in the chatbots. In the SUS evaluation, a total of 26 users (frequent and moderate-level users) participated and responded to SUS questionnaires. The SUS scores were then calculated to find out the user satisfaction with the chatbots.

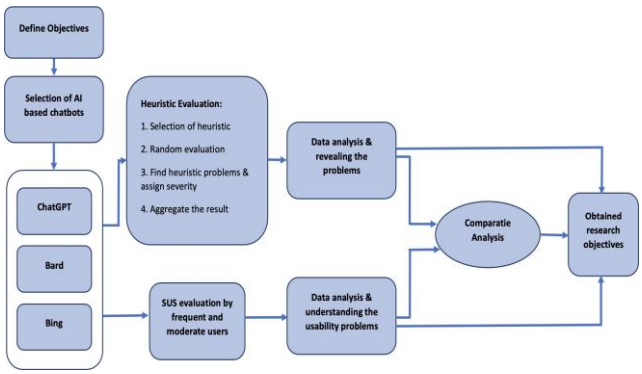

Figure 1: A methodological overview of usability analysis of chatbots.

2.1 Heuristic Evaluation

2.1.1 Evaluators' Profile

Three usability experts conducted a thorough assessment of each chatbot system based on Nielsen's 10 heuristics. All three experts are graduates with a major in computer science. All the experts have theoretical and practical knowledge of human-computer interaction, HE, and SUS evaluation. They have an average age of 25.33±1.58 years and have experience in UI design and evaluation of 2-3 years. They have evaluated the usability of 5-8 web and mobile applications. According to Nielsen's guidelines, employing 3 to 5 evaluators are typically sufficient for heuristic evaluation, as this number balances thoroughness with efficiency. However, the use of three expert evaluators in our study, each with a strong background in HCI and heuristic evaluation provided a solid foundation for reliable and meaningful results.

2.1.2 Procedure

In this evaluation process, the experts identified instances where the design and functionality of the chatbots didn't align with the established usability heuristics outlined by Nielsen (Nielsen, 1995). They looked into the details of the chatbots' appearance, how users moved through them, and the overall experience of using them. This allowed them to detect various issues that might create problems for users, highlighting areas that needed improvement.

All the experts have individually evaluated the three chatbots to identify the usability issues and determine which specific usability rules were being violated by each problem. Problems revealed by each evaluator are then aggregated. Additionally, they assigned severity ratings (as suggested in (Nielsen, 1995)) to each of the problems to outline the significance of the problems. The severity ratings given by the three experts were then aggregated by taking the average value.

Then the severity ratings of the usability problems were categorized in the following manner (Islam, Bouwman, & Islam, 2020): (a) Usability problems with an average rating below 1.5 were labeled as Cosmetic. (b) Usability problems with an average rating falling between 1.5 and 2.5 were categorized as Minor. (c) Usability problems with an average rating ranging from 2.5 to 3.5 were considered as Major. (d) Usability problems with an average rating equal to or exceeding 3.5 were designated as Catastrophic.

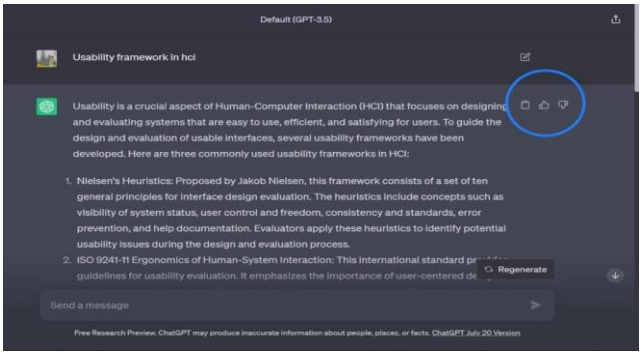

Figure 2: An example problem of ChatGPT.

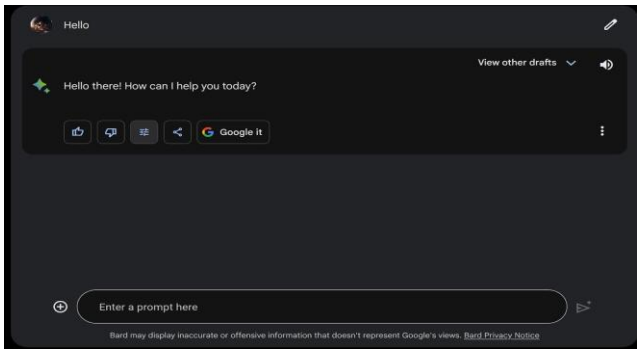

Figure 3: An example problem of Google Bard.

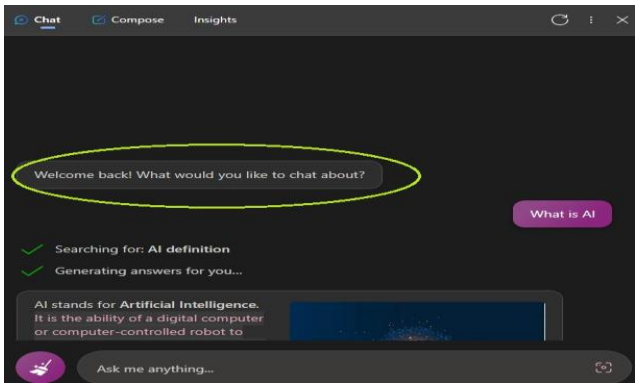

Figure 4: An example problem of Bing Chat

2.1.3 Evaluators' Findings

Several usability issues were found in the chatbots. One such problem centers around the utilization of a copy icon within ChatGPT, which lacks any form of visual indication or hint

when a user hovers their mouse cursor over it (see Figure 2). Novice users may not recognize this icon. According to the three usability experts, this absence of hint can lead to a potential loss of user control as system status is not clearly visible within the system. Furthermore, it can render the chatbot inefficient to use, as users may find themselves at a loss regarding how to effectively employ this particular icon. This issue can be seen as a breach of three critical usability heuristics: heuristic 1 which is Visibility of System Status, heuristic 3, which pertains to User Control and Freedom, and heuristic 7, which addresses Flexibility and Efficiency of Use. Notably, two of the experts classified this issue as a major usability concern, assigning it a severity rating of 3 on a scale of 0 to 4. On the other hand, the third expert, rated the problem as catastrophic, assigning it a higher severity rating of 4. The average severity rating of this problem was found to be 3.33, that makes it a major usability problem (Islam, Bouwman, & Islam, 2020).

One of the most prominent usability issues with the Google Bard platform is that users cannot stop the answer from being generated after it has started, as seen in Figure 3. The lack of a button or mechanism specifically intended for this purpose results in the user losing control over the system, which is a problematic situation. The experts claim that this problem violates heuristic 3. This issue has been classified by all experts as a catastrophic usability issue, earning an average usability rating of 4. Bing Chat revealed one such catastrophic issue with an average severity rating of 4. As shown in Figure 4, there is no chat history available in Bing Chat that leads the user to not being able to use their previous chat information. The experts have classified it as a violation of heuristic 6, which relates to recognition rather than recall and they have all rated this problem as a catastrophic usability problem.

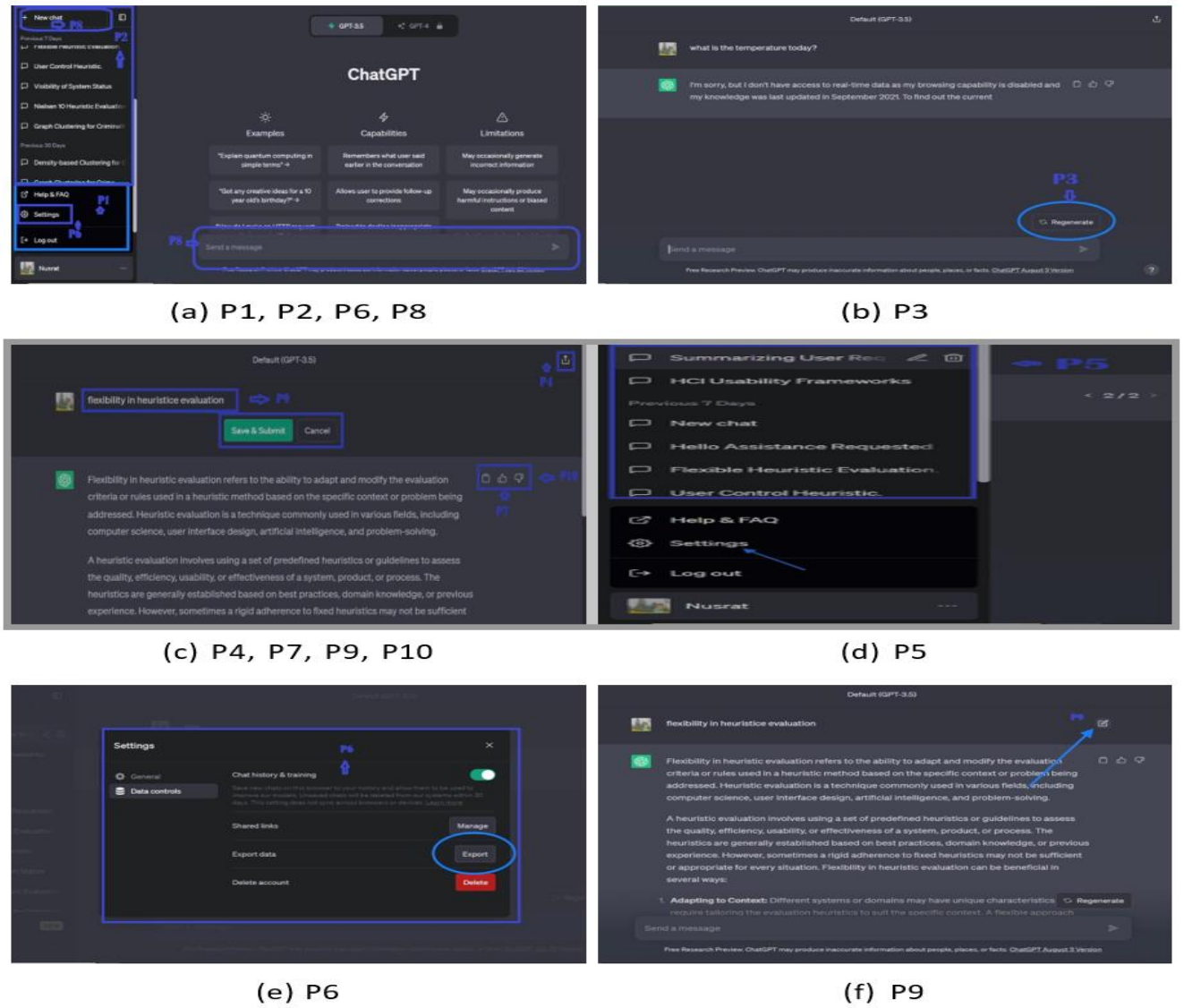

Figure 5: Revealed usability problems of ChatGPT: (a) Represents the problem P1, P2, P6, and P8; (b) Represents the problem P3; (c) Represents the problem P4, P7, P9, and P10; (d), (e), and (f) Represents the problem P5. P6, and P6 correspondingly.

Overall, the heuristic evaluation led us to the discovery of a total of 10, 10, and 13 usability problems for ChatGPT, Google Bard, and Bing Chat respectively with an average severity of 2.6, 2.9, and 2.8 (see Table 7). Each of these revealed problems along with their severity rating and violated heuristics are represented in Table 3, Table 4, and Table 5 respectively. Furthermore, to contribute to the enhancement of these chatbots, necessary recommendations and suggestions have been provided for the identified problems. All the revealed usability problems of ChatGPT, Google Bard and Bing Chat are visually represented in Figure 5, 6 and 7 respectively.

Again, Table 7 shows that Bard and Bing Chat each possesses one and two catastrophic problems. ChatGPT and Bing Chat exhibit an equivalent count of six major usability problems, while Bard presents a total of seven such issues. As for minor usability problems, ChatGPT has four, Bard has two, and Bing Chat has five.

The usability problems found across the chatbots resulted in the violation of nearly all of Nielsen's heuristics. The distribution of problems violating specific heuristics is presented in Table 8. Overall, a total five, seven and six heuristics were violated by ChatGPT, Google Bard and Bing Chat respectively. Examining the study findings, it is observed that ChatGPT had two problems violating heuristic 1, while Bard and Bing Chat had three such problems each, respectively. Heuristic 2 was violated solely by Bard, and heuristic 3 faced violations across all chatbots. Heuristic 4 encountered violations from three problems in Bard and two problems in Bing Chat, whereas Chat- GPT adhered to this heuristic. Bard alone violated heuristic 5, while heuristic 6 was only breached by Bing Chat. Heuristic 7 sustained violations from five problems in ChatGPT and six problems in Bing Chat. Heuristic 9 was not violated by any of the chatbots, whereas ChatGPT and Bing Chat violated H10.

2.2 User Study

In this study, we employed the System Usability Scale (SUS), a widely used and validated questionnaire to assess the usability of our chatbot systems.

2.2.1 Participants' Profile

A total of 40 individuals (22 female and 18 male) participated in this assessment, representing a diverse range of users with various backgrounds and experiences. It's important to note that each of these participants possessed a fundamental level of knowledge in Information and Communication Technology (ICT). Participants were recruited through a combination of personal outreach and digital invitation. Specifically, invitations were distributed via the authors' personal networks and shared on social media platforms to reach a broad audience. Additionally, a targeted invitation was sent via email to a group of final-year undergraduate students from the authors' academic institutions, who were selected based on their relevant academic background and potential familiarity with digital interfaces.

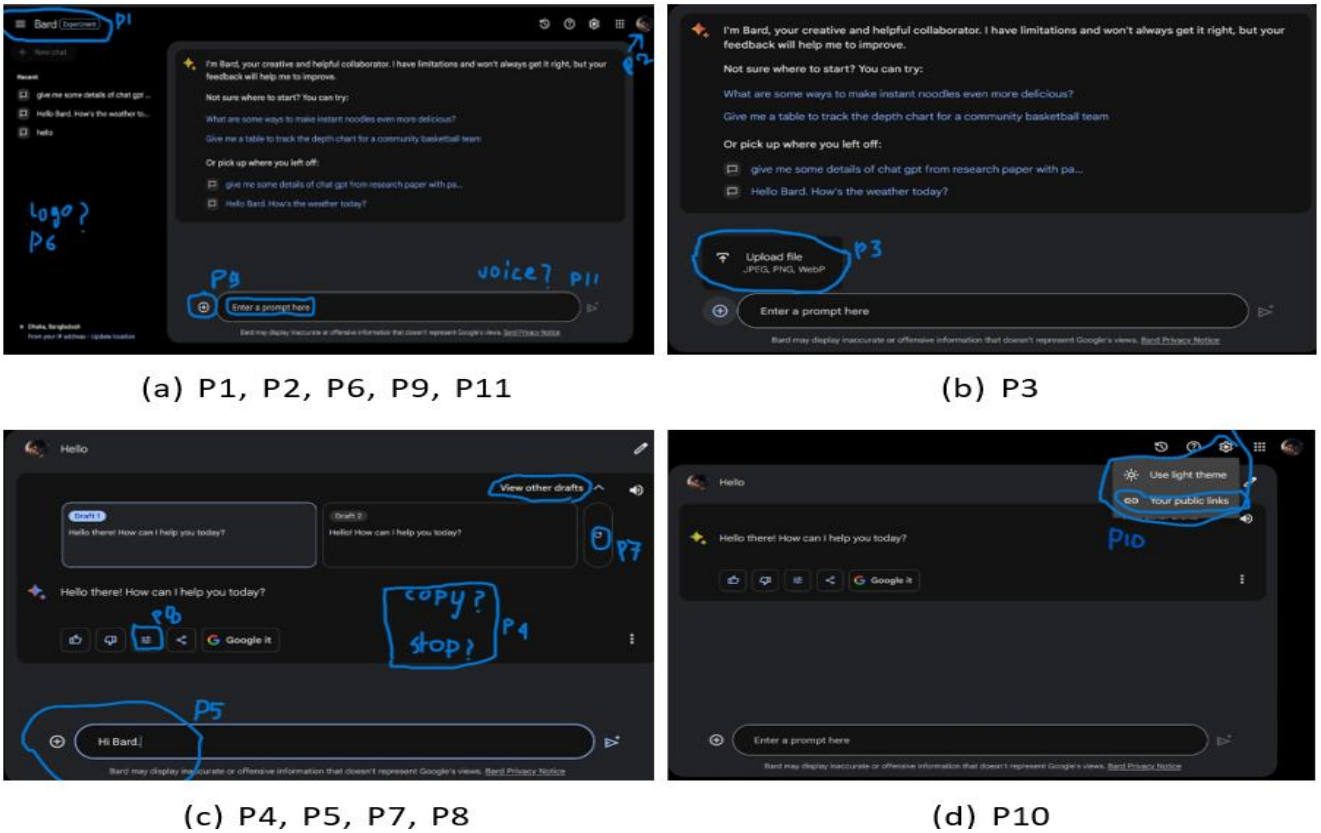

Figure 6: Revealed usability problems of Google Bard: (a) Represents the problem P1, P2, P6, P9 and P11; (b) Represents the problem P3; (c) Represents the problem P4, P5, P7, and P8; (d) Represents the problem P10 correspondingly

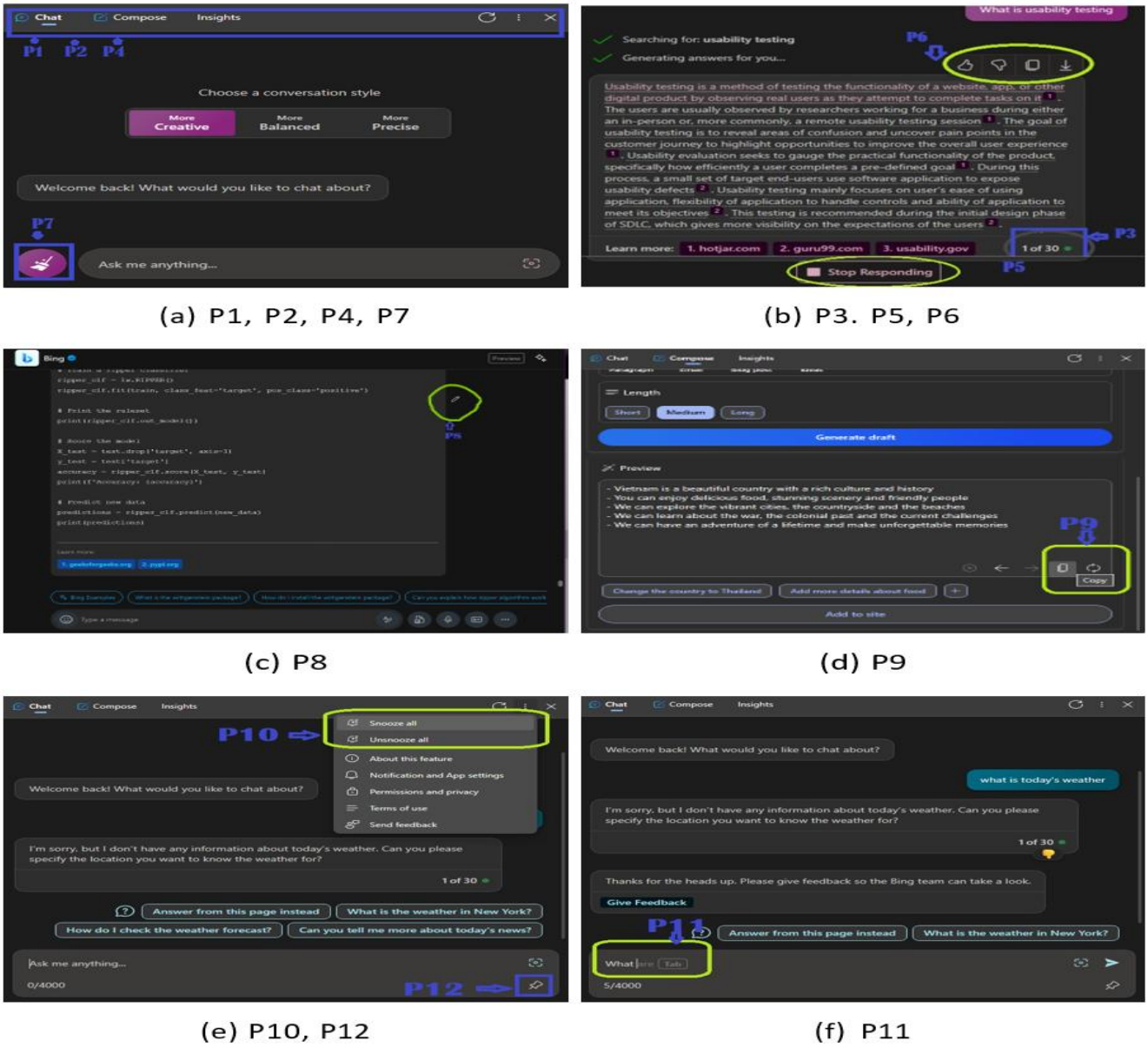

Figure 7: Revealed usability problems of Bing Chat: (a) Represents the problem P1, P2, P4, and P7; (b) Represents the problem P3, P5, and P6; (c) and (d) Represents the problem P8, and P9, (e) Represents the problem P10 and P12 and (e) Represents the problem P11 correspondingly.

Again, 27 of them had a significant degree of familiarity with ChatGPT, indicating frequent interactions with this specific chatbot in their daily lives. Similarly, 25 participants were well acquainted with Google Bard, and 23 participants were well acquainted with Bing Chat. Other participants, a total of 5, 4 and 3 participants fell into the category of moderate users of ChatGPT, Bard and Bing Chat respectively, indicating that they used these chatbots moderately in their daily routines. Among these 40 individuals 31 participants have explored all three chatbots. The demographic profiles of the 40 SUS participants, categorized by gender, age-group, and experience level is shown in Table 6. Among the 22 female and 18 male participants, the majority were between 21 and 30 years old. The participants consisted of frequent and moderate chatbot users, including both male and female individuals aged between 21 and 35, reflecting a diverse range of user backgrounds relevant to chatbot interactions.

Table 6: Demographic Overview of SUS Participants

Gender	User Experience	Age-group
Female (22)	Frequent (15) &Moderate (07)	21 – 25 (9)
		26 – 30 (8)
		31 – 35 (4)
Male (18)	Frequent (13) &Moderate (05)	21 – 25 (7)
		26 – 30 (5)
		31 – 35 (6)

2.2.2 Study Procedure

For the SUS evaluation, participants were assigned to interact with the chatbots. Each participant was allocated a dedicated 30-minute time frame to interact with the chatbots and explore their functionalities and features. During the session, participants were instructed to explore the overall user interface of the all three chatbot web applications. In addition to general navigation, they were asked to engage in typical conversational tasks such as asking general inquiries,

copying text, requesting alternative responses, reacting to replies, and sharing conversations. Furthermore, they were directed to perform more advanced tasks, including summarizing a given text, identifying key terms, and paraphrasing content. Following their interaction with the chatbots, participants were asked to complete the SUS questionnaire to provide their subjective assessment of the chatbot's usability. Before commencing the evaluation process, participants were provided with a clear understanding of the evaluation's purpose and objectives. It was emphasized to them that the evaluation was focused solely on assessing the usability of the chatbots and not on evaluating their personal abilities or skills. After getting the responses from the participants, the SUS score was calculated following the standard guidelines (Public Affairs, 2023). Higher SUS scores indicate better usability and user satisfaction and lower SUS scores indicates lower usability and satisfaction.

2.2.3 Study Findings

Analyzing the responses of the participants found a notable variation in their feedback for each question of SUS related to the three chatbots. These variations are represented using mean ± SD value in Table 9. For example, participants' responses to SUS question 1, which inquired about the participants' likelihood to use the systems frequently, received scores of 3.92 ± 0.93, 3.65 ± 0.85 and 3.35 ± 0.8 for ChatGPT, Bard and Bing Chat respectively. The mean ± SD value of 3.92 ± 0.93, means that most of the participant responses fell within the range of approximately 3.00 to 4.84. In other words, this indicates that the majority of participants expressed agreement or strong agreement regarding their likelihood to use ChatGPT frequently. Their responses clustered around the agree-to-strongly-agree spectrum.

Similarly, most participant responses for Google Bard were within the range of approximately 2.80 to 4.50; which indicates that for Google Bard, participants exhibited a moderate level of agreement regarding their intent to use the system frequently, but there was more variability in responses compared to ChatGPT. Again, for Bing Chat the calculated Mean of 3.35 suggests that, on average, participants had a somewhat lower inclination to use Bing Chat frequently compared to the other two chatbots. Similarly, the analysis of responses to SUS question 2 unveiled ChatGPT as less complex, Google Bard as moderately complex, and Bing Chat as relatively more complex. The average SUS scores of ChatGPT, Bard and Bing Chat were found as 72.88, 66.54, and 61.63 respectively, that indicates ChatGPT has better usability and user satisfaction than the other two chatbots. Considering the SUS score rating, for ChatGPT, its SUS score of 72.88 falls within the range of 68 to 80.3 which indicates that users of ChatGPT experienced a good level of ease, efficiency, and satisfaction while interacting with the system. However, a different scenario unfolds for Google Bard and Bing Chat. Their respective SUS scores of 66.54 and 61.63 position them within the range of 51 to 68 on the SUS score rating table. This range suggests a lower level of user satisfaction in comparison to ChatGPT.

The analysis of SUS scores revealed notable variations between frequent users and moderate users across all three chatbots as shown in Table 10. Frequent users reported engaging with chatbots daily for a range of tasks, from general inquiries to more sophisticated, goal-oriented purposes. In contrast, moderate users typically interacted with the chatbots once or twice a week, primarily to complete two to five general-purpose tasks. Among the 40 participants, 15 female and 13 male users were classified as frequent users, while the remaining 12 participants were categorized as moderate users (Table 10). These SUS scores served as a valuable metric for gauging the perceived usability of each chatbot among these distinct user groups.

Starting with ChatGPT, a notable difference is evident. Frequent users bestowed a SUS score of 71.43, indicating a reasonably positive perception of its usability. In contrast, moderate users assigned a notably higher SUS score of 79. This surprising trend suggests that moderate users found ChatGPT to be even more usable and user-friendly compared to their frequent user counterparts.

Moving on to Google Bard, a similar yet different pattern emerged. Frequent users assigned a SUS score of 67.14 to this chatbot, indicating a favorable perception of its usability. In contrast, moderate users awarded a slightly lower SUS score of 64. Despite the slight disparity, it is evident that both user groups found Google Bard to be reasonably usable, with frequent users leaning towards a higher usability rating.

For Bing Chat, frequent users assigned a SUS score of 61.55, while moderate users provided a slightly higher score of 62. In this case, the distinction between the two user groups was relatively minor. Both frequent and moderate users perceived Bing Chat as reasonably usable, with no significant variation in their assessments.

Table 11 presents the results of a one-way ANOVA conducted on the mean System Usability Scale (SUS) scores collected from participants for ChatGPT, Google Bard, and Bing Chat. ChatGPT had the highest mean SUS score (3.100), followed by Google Bard (3.015), and Bing Chat (2.966). However, the calculated F-statistic ($F = 0.080$) and the corresponding p-value ($p = 0.923$) indicate no statistically significant difference in perceived usability among the three chatbots. This suggests that, based on user responses, the overall usability experience was comparable across all platforms.

Table 7: Number of problems with their severity ratings.

Severity Rating	1 (Cosmetic Problem only)	2 (Minor Usability Problem)	3 (Major Usability Problem)	4 (Usability Catastrophe)	Average Severity Rating
ChatGPT	0	4	6	0	2.6
Google Bard	0	2	7	1	2.9
Bing Chat	0	5	6	2	2.8

Table 8: Number of problems that violated the corresponding heuristics.

Heuristics Violated	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	Total number of Heuristics violated
ChatGPT	2		5				5	2		1	5
Google Bard	3	3	2	3	3		2	1			7
Bing Chat	3		4	2		1	6			1	6

Table 9: Mean ± SD value for each SUS Questions.

SUS Questions	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
ChatGPT	T3.92 ± 0.93	2.08 ± 0.63	4.27 ± 0.53	2.04 ± 0.77	3.92 ± 0.69	2.31 ± 0.74	4.19 ± 0.63	2.12 ± 0.77	3.77 ± 0.59	2.38 ± 0.9
Google Bard	3.65 ± 0.85	2.27 ± 0.6	3.85 ± 0.46	2.15 ± 0.73	3.73 ± 0.67	2.38 ± 0.64	3.58 ± 0.86	2.42 ± 0.76	3.58 ± 0.81	2.54 ± 0.9
Bing Chat	3.35 ± 0.8	2.69 ± 0.88	3.69 ± 0.79	2.31 ± 0.74	3.31 ± 0.79	2.46 ± 0.81	3.35 ± 0.75	2.62 ± 0.9	3.46 ± 0.9	2.42 ± 0.86

Table 10: Comparison of SUS scores for Chat- bots between the frequent and moderate users.

Total Participants (40)	Frequently Used (Female:15, Male:13)	Moderately Used (Female:7, Male:5)
ChatGPT	71.43	79
Google Bard	67.14	64
Bing Chat	61.55	62

Table 11: ANOVA results comparing SUS scores among ChatGPT, Google Bard, and Bing Chat.

Chatbots	Mean SUS Score	Variance	Standard Deviation	F-statistic	p-value
ChatGPT	3.100	0.958	0.979	0.080	0.923
Google Bard	3.015	0.504	0.710		
Bing Chat	2.966	0.262	0.512		

3. RESULTS AND DISCUSSION

A comparative analysis of the SUS evaluation results and the findings from the heuristic evaluation show a clear and intriguing pattern. There appears to be a somewhat linear relationship between the SUS scores and the number of heuristic violations. For example, where Google Bard and Bing Chat garnered lower SUS scores, the heuristic evaluation also uncovered a higher number of heuristic problems than ChatGPT within these chatbots. The results shown in Table 8 illustrate this correlation where ChatGPT was found to have violated a total of five heuristics, while Google Bard and Bing Chat were associated with seven and six heuristic violations, respectively.

Moreover, a closer examination of the specific heuristics that were violated reveals interesting trends. From Table 8, it becomes evident that the majority of problems stem from the violation of heuristic 3, which relates to user control and freedom, and heuristic 7, which relates to the flexibility and efficiency of system use. Hence it can be said that users need to be given control and freedom over the system and it should be flexible and efficient to use.

Moving to the severity ratings assigned to these problems, a noteworthy difference emerges among the three chatbots. Google Bard and Bing Chat each exhibit one and two catastrophic problems, indicating critical usability issues that demand immediate attention. On the other hand, ChatGPT presents no catastrophic problems, suggesting a relatively smoother user experience. Google Bard and Bing Chat also face major and minor usability problems.

Bing Chat, in particular, is associated with a higher number of major usability problems, a total of six, while Google Bard presents seven such issues. Based on minor, major, and catastrophic issues, it becomes apparent that Bard and Bing Chat have more severe usability issues than ChatGPT. Thus, Bard and Bing Chat have comparatively poorer usability. However, it is worth noting that ChatGPT has significant major and minor usability issues that must be addressed to improve overall usability performance.

4. CONCLUSION

This study employed both heuristic evaluation and the System Usability Scale (SUS) assessment to gain a comprehensive understanding of the chatbots' usability performance. These evaluation methodologies provided valuable insights into the extent to which the chatbots adhered to usability standards. In this research, a comprehensive investigation using heuristic evaluation enabled us to pinpoint specific areas where the chatbots' usability was deficient, encompassing design characteristics, interaction flow, and user engagement. Each identified usability issue provided a clear directive for improvement.

Concurrently, the System Usability Scale (SUS) evaluation furnished a detailed and quantitative perspective on user

satisfaction and usability. The SUS scores for each chatbot highlighted their relative strengths and limitations concerning ease of use, learnability, and overall user satisfaction.

These numerical scores were instrumental in assessing the effectiveness of the chatbot systems in addressing user needs. By integrating the findings from both the heuristic evaluation and the SUS, we developed an exhaustive assessment of the chatbots' usability. This methodology not only identified areas requiring enhancement but also prioritized the severity of each usability issue. Some issues necessitated immediate intervention due to their significant impact on user experience, while others warranted gradual improvement.

To further support the SUS analysis, a one-way ANOVA was conducted to determine whether the differences in the mean SUS scores across ChatGPT, Google Bard, and Bing Chat were statistically significant. The analysis yielded an F-statistic and a p-value that indicated no significant variance among the three platforms, despite minor variations in specific interface features and interactions.

The primary contribution of this research lies in its comprehensive evaluation of the usability of three widely used AI chatbots: ChatGPT, Google Bard, and Bing Chat. This study uniquely combines both user-centered evaluation through the System Usability Scale (SUS) and expert evaluation via Heuristic Evaluation (HE). Such an integrative approach is relatively uncommon in the existing literature, offering a robust framework for understanding and improving chatbot usability. By identifying specific design flaws, interaction issues, and user engagement problems, the research provides actionable insights for enhancing the overall user experience of these AI systems. The findings indicate that while all chatbots have usability challenges, ChatGPT shows comparatively better performance, thus offering a valuable benchmark for future improvements in chatbot design and functionality.

Our study does have certain limitations that should be acknowledged. One of these limitations is the relatively small number of participants involved in the System Usability Scale (SUS) survey. Engaging a larger and more diverse group of users with varying levels of exposure to these chatbots could have provided more comprehensive insights. Additionally, we only examined three chatbots and, whereas including a broader range of chatbots could have enriched the study. Another limitation was the inaccessibility of premium editions of the chatbots. For instance, in our study, we only explored the free edition of ChatGPT. This restricted access may have limited our ability to thoroughly investigate specific operations, features, or characteristics of the chatbot systems, which would have been beneficial for a more in-depth analysis. Future research could focus on evaluating other widely used chatbots (with paid and trial versions) to further generalize the findings to enhance the paper's contribution to usability research within the field of Human-Computer Interaction.

In this study, we have thoroughly examined the chatbots from a usability perspective, identified areas needing improvement, assessed the severity of issues, and provided guidance on potential solutions. Our study is unique as it combines both user-centered evaluation (SUS) and expert evaluation (heuristic evaluation), which is a relatively uncommon approach in previous related works. By adopting this comprehensive approach, we aimed to contribute to developing chatbot systems that offer an optimal and satisfying user experience. However, in this study, we conducted a thorough examination of chatbots from a usability perspective, identifying areas needing improvement, assessing the severity of issues, and providing guidance on potential solutions. Our study is unique in combining both user-centered evaluation (System Usability Scale, SUS) and expert evaluation (heuristic evaluation), an approach that is relatively uncommon in previous related works. By adopting this comprehensive methodology, we aim to contribute to the development of chatbot systems that offer an optimal and satisfying user experience.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to all the participants who took part in the System Usability Scale (SUS) evaluation and contributed their valuable feedback in this research.

AUTHOR DECLERATION

The authors declare that there is no conflict of interest.

FUNDING INFORMATION

This research did not receive any fund.

REFERENCES

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. Machine Learning with Applications, 2, 100006.

Brandtzaeg, P. B., & Folstad, A. (2017). Why people use chatbots. In Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22–24, 2017, Proceedings (Vol. 4, pp. 377–392). Springer.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), Usability evaluation in industry (pp. 189–194). Taylor & Francis.

Folmer, E., & Bosch, J. (2004). Architecting for usability: A survey. Journal of Systems and Software, 70(1–2), 61–78.

Fuchs, K. (2023). Exploring the opportunities and challenges of NLP models in higher education: Is ChatGPT a blessing or a curse? Frontiers in Education, 8, 1166682.

Hill-Yardin, E. L., Hutchinson, M. R., Laycock, R., & Spencer, S. J. (2023). A Chat (GPT) about the future of scientific publishing. Brain, Behavior, and Immunity, 110, 152–154.

Hossain, T., Mohiuddin, T., Hasan, A. S., Islam, M. N., & Hossain, S. A. (2020). Designing and developing graphical user interface for the multichain blockchain:

Towards incorporating HCI in blockchain. In International Conference on Intelligent Systems Design and Applications (pp. 446–456). Springer.

HubSpot, (2018). What's the system usability scale (SUS) and how can you use it? HubSpot Blog.

Hvannberg, E. T., Law, E. L.-C., & Lérusdóttir, M. K. (2007). Heuristic evaluation: Comparing ways of finding and reporting usability problems. Interacting with Computers, 19(2), 225–240.

International Organization for Standardization. (2018). ISO 9241-11:2018 – Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts. ttps://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed2:v1:en

Islam, M. N., Bouwman, H., & Islam, A. K. M. N. (2020). Evaluating web and mobile user interfaces with semiotics: An empirical study. IEEE Access, 8, 84396–84414. https://doi.org/10.1109/ACCESS.2020.2991840

Jain, M., Kumar, P., Kota, R., & Patel, S. (2018). Evaluating and informing the design of chatbots. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 895–906. https://doi.org/10.1145/3196709.3196735

Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., ... & Jahangir, E. (2023). Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the ChatGPT model.

Kasinathan, G. (2023). Musk's Twitter acquisition. Economic & Political Weekly, 58(2), 21.

Khairat, M. I. S. B., Priyadi, Y., & Adrian, M. (2022). Usability measurement in user interface design using heuristic evaluation & severity rating (case study: Mobile TA application based on MVVM). In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 974–979). IEEE.

Kundu, S., Kabir, A., & Islam, M. N. (2020). Evaluating usability of pregnancy tracker applications in Bangladesh: A heuristic and semiotic evaluation. In 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC) (pp. 1–6). IEEE.

Langevin, R., Lordon, R. J., Avrahami, T., Cowan, B. R., Hirsch, T., & Hsieh, G. (2021, May). Heuristic evaluation of conversational agents. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–15).

Liu, X., Wu, C., Lai, R., Lin, H., Xu, Y., Lin, Y., & Zhang, W. (2023). ChatGPT: When the artificial intelligence meets standardized patients in clinical training. Journal of Translational Medicine, 21(1), 447.

Mack, R., & Nielsen, J. (1993). Usability inspection methods: Report on a workshop held at CHI'92, Monterey, CA, May 3–4, 1992. ACM SIGCHI Bulletin, 25(1), 28–33.

Martindale, J. (2023). What is Google Bard? Here's how to use this ChatGPT rival. Digital Trends. https://www.digitaltrends.com/computing/how-to-use-google-bard

McLellan, S., Muddimer, A., & Peres, S. C. (2012). The effect of experience on system usability scale ratings. Journal of Usability Studies, 7(2), 56–67.

Muaz, M. H., Islam, K. A., & Islam, M. N. (2021). Assessing the usability of truck hiring mobile applications in Bangladesh using heuristic and semiotic evaluation. In Advances in Design and Digital Communication (pp. 90–101). Springer.

Munim, K. M., Islam, I., Rahman, M. M., & Islam, M. N. (2020). Adopting HCI and usability for developing Industry 4.0 applications: A case study. In 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI) (pp. 1–6). IEEE.

Nielsen, J. (1995). How to conduct a heuristic evaluation. Nielsen Norman Group, 1(1), 8.

Nielsen, J. (1995). 10 usability heuristics for user interface design (Vol. 1). Nielsen Norman Group.

Nguyen, P., Trng, H., Nguyen, P., Bruneau, P., Cao, L., & Wang, J. (2023). Evaluation of Google Bard on Vietnamese high school biology examination. ResearchGate. https://www.researchgate.net/

Public Affairs, A. S., (2023). (n.d.). System Usability Scale (SUS). Usability.gov. https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html [Accessed August 11, 2023]

Rahaman, M. S., Ahsan, M., Anjum, N., Rahman, M. M., & Rahman, M. N. (2023). The AI race is on! Google's Bard and OpenAI's ChatGPT head-to-head: An opinion article. SSRN.

Ram, B., & Verma, P. (2023). Artificial intelligence AI-based chatbot: Study of ChatGPT, Google AI Bard and Baidu AI. World Journal of Advanced Engineering Technology and Sciences, 8(01), 258–261.

Rane, N. (2023). Roles and challenges of ChatGPT and similar generative artificial intelligence for achieving the sustainable development goals (SDGs). SSRN. https://ssrn.com/abstract=4603244

Rane, N. L., Tawde, A., Choudhary, S. P., & Rane, J. (2023). Contribution and performance of ChatGPT and other large language models (LLM) for scientific and research advancements: A double-edged sword. International Research Journal of Modernization in Engineering Technology and Science, 5(10), 875–899.

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. Journal of Applied Learning and Teaching, 6(1).

Shidiq, M. (2023). The use of artificial intelligence-based ChatGPT and its challenges for the world of education: From the viewpoint of the development of creative writing skills. In Proceedings of the International Conference on Education, Society and Humanity (Vol. 1, pp. 353–357).

Tasfia, S., Islam, M. N., Nusrat, S. A., & Jahan, N. (2023). Evaluating usability of AR-based learning applications for children using SUS and heuristic evaluation. In Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering: TCCE 2022 (pp. 87–98). Springer.

Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on social media. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 3506–3510).

Zdnet. (2023). What is Bing Chat? Here's everything you need to know. https://www.zdnet.com/article/what-is-the-new-bing-heres-everything-you-need-to-know [Accessed August 11, 2023]

Table 3: Revealed usability problems of ChatGPT (Figure 5)

No	Problems (In detail)	Evidence (Violated heuristics)	E1	E2	E3	Average Severity ating (0-4)	Possible solution
P1	There is no clear link for user profile. Clicking on the user's name doesn't show any info about the profile or no option of editing user info	H3	3	3	4	3.33	User info can be added/ made editable.
P2	Help and documentation segment is not visible. The system can provide easily accessible and comprehensive documentation to assist users in under- standing and using the interface effectively.	H10	3	2	2	2.33	Attach help and documentation section in profile.
P3	Response generation of a chat cannot be resumed after stopping it.	H3, H7	3	3	3	3	Include resume button/option
P4	Sharing a chat link does not show that the link is being shared anonymously in the initial stage.	H7	2	2	1	1.67	Anonymous and naming both options can be visible when sharing.
P5	The edit and delete option is not visible in chat list. It only appears if the chat is clicked.	H3, H7	3	4	3	3.33	Make operation icons visible in chat list or it can at least be visible if mouse is hovered above the chat name.
P6	*Chat Exporting* option is present but hard to find for a user.	H7, H8	3	1	1	1.67	It may be made easily accessible.
P7	There is no instruction with the icon copy, like or dislike right after the ChatGPT response. Elderly or new user may not recognize the copy icon.	H1, H7, H3	3	4	3	3.33	Short instruction with icon can be given.
P8	While creating a new chat, the chat window only expands to a certain height and the example, capabilities and limitation section remains.	H8	1	2	2	1.67	If a chat grows then the chat window may grow and example, capabilities and limitation section may be vanished.
P9	After clicking on the edit icon on the chat window, the cursor is not visible in the text. Hence it is not clear whether the icon is let- ting the user to edit the chat message	H1	3	3	2	2.67	Cursor can be visible on the text.
P10	There is no way to delete a message in a chat	H3	4	3	3	3.33	This option can be more intuitive and easily understandable.

Table 4: Revealed usability problems of Google Bard (Figure 6)

No	Problems (In detail)	Evidence (Violated heuristics)	E1	E2	E3	Average Severity ating (0-4)	Possible solution
P1	There is nothing inside the main menu. It just slides the recent chat bar heading or widen the chat display by clicking the main memory.	H2	3	4	3	3.33	Inside main menu- multiple chat theme, media section, font section, private chat section, documentation and logout section can be introduced.
P2	Bard only uses google account. If a user wants to sign different account in google and want to access different bard account it will not be possible. There is no sign out option from Bard.	H3	2	1	2	1.67	It may use other accounts rather than only Google, Like Yahoo, Microsoft account for signing in with password. So that a user has the flexibility to use this chatbot without having a google account.
P3	No document, PDF, PowerPoint can be share through bard. It can access only JPEG, PNG and Webp.	H4, H7	3	2	3	2.67	Bard may be given the access to docs or PDF or PowerPoint and can work on" find a key-word", "search a meaning", "suggest related books on that topic".
P4	There is no stop response function.	H3	4	4	4	4	There may be a *stop* response *function* while generating the chatbot response.
P5	There is enter function not working for next line. Similarly, no visible next line function or instruction is avail- able.	H1, H7	3	3	3	3	There can be a visible instruction for next line.
P6	No beautiful or engaging picture, chat theme, icon or logo in Bard.	H8	2	1	2	1.67	There may be using some interesting pictures, icon, logo, theme in chat.
P7	Regenerate function is not visible. It is hidden in View all drafts in case of multiple response only.	H1	3	2	3	2.67	The regenerate function can be made easily accessible for user.
P8	The Modify Function used here is not familiar with real life example	H2	2	3	3	2.67	The Modify Icon may be changed.
P9	"Enter" a prompt here and the left "+" icon are conflicting. New user may think of it as a new chat bar.	H2, H4	3	4	3	3.33	The Plus (+) icon may be changed with image icon.
P10	The share icon and create link in the settings option both are same. It may look confusing to user.	H4	3	3	3	3	There is no need to be create public link in settings option.

Table 5: Revealed usability problems of Bing Chat (Figure 7)

No	Problems (In detail)	Evidence (Violated heuristics)	E1	E2	E3	Average Severity Rating (0-4)	Possible solution
P1	Help and documentation segment is not visible. The system may provide easily accessible and comprehensive documentation to assist users in under- standing and using the interface effectively.	H10	2	3	2	2.33	Include a help and documentation section.
P2	No Chat History available.	H6	4	4	4	4	Keep the record of the chats.
P3	It is not clear why 1 of 30 is written in the chat. It may be confusing for new users	H7	3	2	2	2.33	Some hints could be given if mouse pointer is hovered above it.
P4	Though there is a sign in option, but there is no option of signing out from Bing Chat	H3	4	4	3	3.67	Sign out may be added.
P5	Response generation of a chat cannot be resumed after stopping it.	H3, H7	2	2	1	1.67	Resume button may be added.
P6	There is no way of saving or deleting a chat	H3	3	3	3	3	Saving and deleting operation maybe implemented.
P7	The New topic icon is not consistent with the convention and with other systems	H4	2	3	2	2.33	Use the conventional icon.
P8	Edit icon in the right of Bing reply message is not workable.	H3, H7	3	4	3	3.33	Make edit functionality work or unnecessary function may not be introduced.
P9	In Insight tab, after pressing the copy button no status is shown whether the text is copied or not.	H1	3	3	3	3	Status message can be made visible.
P10	After clicking on the snooze and unsnoozed button, no message is shown. It is not clear what these two options do.	H1, H7	4	3	3	3.33	Status message can be made visible and clearer instructions needed for the feature.
P11	The function of tab button in text box in Insight window is not consistent with the Chat window. In chat window, pressing tab auto writes the suggestion shown in the chat box. But in insight window, pressing tab doesn't write the suggestion shown, rather it moves to another button	H4	2	3	3	2.67	The operations can be more consistent.
P12	The pin icon doesn't work properly or its functionality isn't clear	H7	3	2	2	2.33	Functionality can be made clear.
P13	The Add to site button is not responsive, no status message is shown and its purpose is not clear	H1, H7	3	3	3	3	Status message can be made visible and more clear instructions needed for the feature.