Evaluating ChatGPT's User Interface Using Nielsen's Heuristics

Esra Özmen, Nurcan Alkış-Bayhan

Abstract— **Artificial intelligence applications are used across various domains, including education, commerce, science, language, law, and healthcare. With the rapid advancement of technology, new artificical applications continue to emerge. Among these, chatbots are the most widely discussed, with ChatGPT being the most prominent one. Although several studies have evaluated ChatGPT, few offer specific recommendations regarding its usability, particularly in terms of its user interface. In this study, the design of ChatGPT's web user interface is evaluated through expert review and system usability assessment. The findings indicate that the most significant issue is the "match between system and the real world," based on Nielsen's heuristics. In contrast, "aesthetic and minimalist design" received the most positive feedback from participants. While the system achieved a usability score of 80, the results also revealed a few inadequately integrated functions.**

Keywords—interface evaluation, ChatGPT, Nielsen Heuristics, user experience

I. INTRODUCTION

ChatGPT is a language model developed by OpenAI, based on the Generative Pre-trained Transformer (GPT) architecture. It is designed to understand and generate human-like text based on the input it receives. The model is pre-trained on a wide range of internet text, allowing it to craft coherent and contextually fitting responses to user inputs. ChatGPT lacks real-time data access; it relies solely on patterns and information retrieved from its training data up to its last update in January 2022 [1]. Users interact with ChatGPT by presenting prompts or questions, and it generates text responses tailored to the input. It has various applications, including natural language understanding, text completion, dialogue systems, problem-solving, and coding. ChatGPT's usage extends to a multitude of domains, including rapid literature browsing, educational sciences, mathematics, and language learning [2, 3, 4, 5]. It is crucial to note that while ChatGPT offers valuable assistance, it may not always provide accurate or up-to-date information. Users should verify their responses in critical situations.

ChatGPT can achieve a wide array of capabilities as a language model [6]. Some of its primary abilities include:

- Answering Text-Based Questions: ChatGPT can generate coherent and logical responses to text-based queries, offering general information, clarification, and answers across diverse topics.
- Text Completion: Users can ask ChatGPT to complete text passages by entering partial sentences or paragraphs. The model completes unfinished sentences and texts.
- Language Understanding and Generation: ChatGPT has the capability to comprehend natural language expressions and provide appropriate responses, making it applicable to a variety of language-related tasks.
- Text-based chat: Users can engage in text-based conversations with ChatGPT. They can ask questions or start discussions on diverse topics.
- Storytelling: Once provided with initial data, the model can assist users in creating a story or scenario about a specific topic.
- Providing Information: ChatGPT can answer questions on general information topics. As mentioned above, it is important to note that the knowledge may not be up to date due to its training data.

In addition to its capabilities, ChatGPT also has several limitations. The model lacks access to real-time updated data and does not guarantee accuracy. Additionally, it cannot offer professional advice on specialized topics such as health, law, or finance. Hence, the information obtained from ChatGPT should be verified before using [7, 8, 9].

In addition to factors such as effectiveness, verifiability, and ethical considerations regarding the information retrieved from the ChatGPT platform, the user interface also plays a crucial role in its overall usability. In the literature, various usability approaches and tests are employed to evaluate user interfaces, as usability assessment is essential for enhancing user effectiveness, efficiency, and satisfaction. However, there is a lack of studies specifically focused on evaluating the user interface of ChatGPT, particularly its web-based interface.

A study conducted by Nathania et al. [10] focused on the user experience approach and evaluated the use of ChatGPT in higher education. Their research revealed that most of the participants were satisfied with ChatGPT's existing features. Participants gave positive evaluations regarding its ease of access and its ability to generate responses consistent with user expectations by understanding the context. They mentioned that they felt comfortable using ChatGPT and were willing to continue using it. However, the study also emphasized the need for improvements in the user interface, citing usability issues

Nurcan Alkış-Bayhan, Technology and Knowledge Management Department, Başkent University, Ankara, Turkey, nalkis@baskent.edu.tr)

Esra Özmen, Ankara Hacı Bayram Veli University, Ankara, Turkey esra.ozmen@hbv.edu.tr (corresponding author)

and functional deficiencies that disrupted the user experience." [10].

In a study evaluating the platform of ChatGPT, the System Usability Scale was applied [11]. This study revealed that ChatGPT's text creation feature received high satisfaction ratings in the survey. Users gave positive feedback on the system's ability to provide short and clear information in a very short time. Additionally, the platform's accessibility at any time and place, error correction, and question-answering capabilities were evaluated positively. However, it was noted that the study mainly focused on text-based usability analysis, highlighting the need for an evaluation of other features of ChatGPT and a comprehensive usability assessment [11].

On the other hand, there is research based on the use of ChatGPT and its role in user experience. Emirhan and Jieun [13] emphasized the importance and effectiveness of ChatGPT in user experience research. In their study, the role of ChatGPT was examined in evaluating user experience. They evaluated design standards of online shopping platforms based on Nielsen heuristics and the effectiveness of AI in user experiences research [13].

Another research on usability analysis and evaluations was performed both by ChatGPT and individuals. The findings of this research revealed that ChatGPT missed the assessment carried out by the participants, and they concluded that human judgment and expertise in usability analysis are irreplaceable [14].

This study aims to assess the usability of the ChatGPT interface design, considering the limited studies in literature and the recommendations provided by these studies. In this context, ChatGPT was evaluated through both expert assessments based on Nielsen's Heuristics and the System Usability Scale.

II. METHODOLOGY

A. Aim of the Study

Usability, as defined by various researchers [15], is the user's capability to execute tasks effectively, efficiently, and with a sense of satisfaction within a specific context by utilizing the relevant tool or interfaces. This fundamental concept of usability has significant importance for both web and mobile interfaces, as it directly impacts user experience. In line with this understanding, our research aims to evaluate the usability of ChatGPT's web interface. Since the wider use of web-based platforms for communication, information retrieval, and various other tasks, a user-friendly interface is important. By scrutinizing the usability of ChatGPT's web interface, this study aims to identify strength areas and potential areas for improvement of ChatGPT to enhance user interaction and satisfaction.

B. Sample

To determine the sample size, we followed the suggestion of Nielsen [16]. Nielsen [16] stated that usability evaluation can be done with at least five participants. So, a total of 5 participants completed the evaluation and the survey in the study. The demographics of the participants are given in Table 1. All the participants have expertise in artificial intelligence, interface design, and human-computer interaction, ensuring the reliability of evaluating the usability of ChatGPT.

TABLE I. DEMOGRAPHIC INFORMATION OF THE SAMPLE

	Education	Work	Age	Gender
Participant1	PhD, Management Information System	Academician	32	Female
Participant2	PhD, Information Systems	Academician	41	Female
Participant3	PhD, Information Systems	Academician	40	Male
Participant4	PhD, Business Administration	Academician	34	Male
Participant5	Bachelor, Computer Engineering	Software Engineer	36	Male

C. Data Collection Tools and Process

To evaluate the user interface of ChatGPT, Nielsen's "10 Usability Heuristics" were used [17]. During the expert evaluation, each participant individually assessed ChatGPT's user interface based on these ten heuristics. Participants categorized the usability issues identified for each heuristic into three levels: minor, moderately important, and major problems. The evaluations were then compared and analyzed to provide design suggestions for improving ChatGPT's interface. Nielsen's "10 Usability Heuristics" [17] and their descriptions are presented below.

1. ***Visibility of system status:*** The system should inform the users by giving feedback about what is happening.
2. ***Match between system and the real world:*** The system should be designed to speak the users' language by using familiar words, phrases, and concepts, making the system more intuitive.
3. ***User control and freedom:*** System functions are usually selected by mistake by the users, so the systems should provide an emergency exit or undo function.
4. ***Consistency and standards:*** The design elements and their behaviors should be consistent, so the users do not have to think about whether different words and actions mean the same thing.
5. ***Error prevention:*** The system design should minimize the errors and problems. If an error occurs, its impact should be prevented, and good error messages should be presented.
6. ***Recognition rather than recall:*** The system design should minimize the memory load of the users. Users should not have to memorize interface details. Information, actions, and options should be visible.

7. ***Flexibility and efficiency of use:*** The system design should be appropriate for both novice and expert users. Shortcuts and similar components can be used to speed up experienced users.
8. ***Aesthetic and minimalist design:*** System components should not include irrelevant information and features in order not to prevent visibility of useful information.
9. ***Help users recognize, diagnose, and recover from errors:*** Error messages should be simple and indicate the problem and its solution.
10. ***Help and documentation:*** The system should not need documentation to be used. However, it should provide help and documentation.

Following the heuristics evaluation, the System Usability Scale (SUS), developed by Brooke [18], was administered to the same participants. The items included in the scale are presented in Table 2. The scale uses a five-point Likert format, ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

TABLE II. SYSTEM USABILITY SCALE

Items
I think that I would like to use this system frequently.
I found the system unnecessarily complex.
I thought the system was easy to use.
I think that I would need the support of a technical person to be able to use this system.
I found the various functions in this system were well integrated.
I thought there was too much inconsistency in this system.
I would imagine that most people would learn to use this system very quickly.
I found the system very cumbersome to use.
I felt very confident using the system.
I needed to learn a lot of things before I could get going with this system.

The evaluation score is computed by applying a specific scoring method to the participants' responses to the System Usability Scale. This includes subtracting -1 from the scale values for items 1, 3, 5, 7, and 9 while subtracting -5 from the scale values for items 2, 4, 6, 8, and 10. The resulting scores are then summed for each participant and multiplied by a coefficient of 2.5. The total score ranges from 0 to 100. In this study, the scale scores were calculated by summing them and dividing by the number of participants, 5.

III. FINDINGS AND DISCUSSION

To assess the web-based user interface of ChatGPT, the participants initially evaluated it based on "10 Usability Heuristics." The evaluations were categorized into three categories based on problematic design: minor, moderate, and major. A frequency analysis was conducted to determine the distribution of these categories. Figure 1 presents the evaluation scenarios of the participants based on the ten heuristics. As seen from the figure, the main element identified as having major problem is the heuristic "Match between system and the real world." This issue can be caused by the misalignment of the icons in the systems with their real-world meaning. It seems that the icons are not consistent with the users' expectations. When we look at the other heuristics "Consistency and standards" were evaluated as a moderate problem by all participants. On the other hand, "Aesthetic and minimalist design," "Visibility of system status," and "Recognition rather than recall" were evaluated as the most problem-free heuristics.

Based on the findings of the study, while we have found the more moderate problems concerning the "10 Usability Heuristics" items, there are also major issues in certain elements of ChatGPT. This observation aligns with Nathania's [10] findings in her study, suggesting a necessity for improvements in the user interface due to disruptive problems and identified deficiencies. However, Nathania [10] did not give details about the causes behind these shortcomings, distinguishing the findings of our study.

Usability scores, retrieved from participants' responses on the System Usability Scale, are presented in Figure 2. According to the results, Participant 4 and Participant 5 scored the highest (92.5%) while most of the participants evaluated the usability score at over 60%. On average, the participants' ChatGPT system usability evaluation yielded a score of 80. This finding implies that users perceive the system as generally usable, although they have issues with functional icons.

Mulia [11] stated that ChatGPT demonstrated satisfactory performance in terms of ease of use and accessibility at any time and location. This finding is parallel with the satisfaction levels obtained on the System Usability Scale in our study.

When assessing the study through both heuristic evaluation and usability scale, the studies conducted by Nathania [10], Mulia [11], and Sakirin [12] overlap with the emphasized requirements for carrying out different usability evaluations, thus highlighting the novelty of the study.

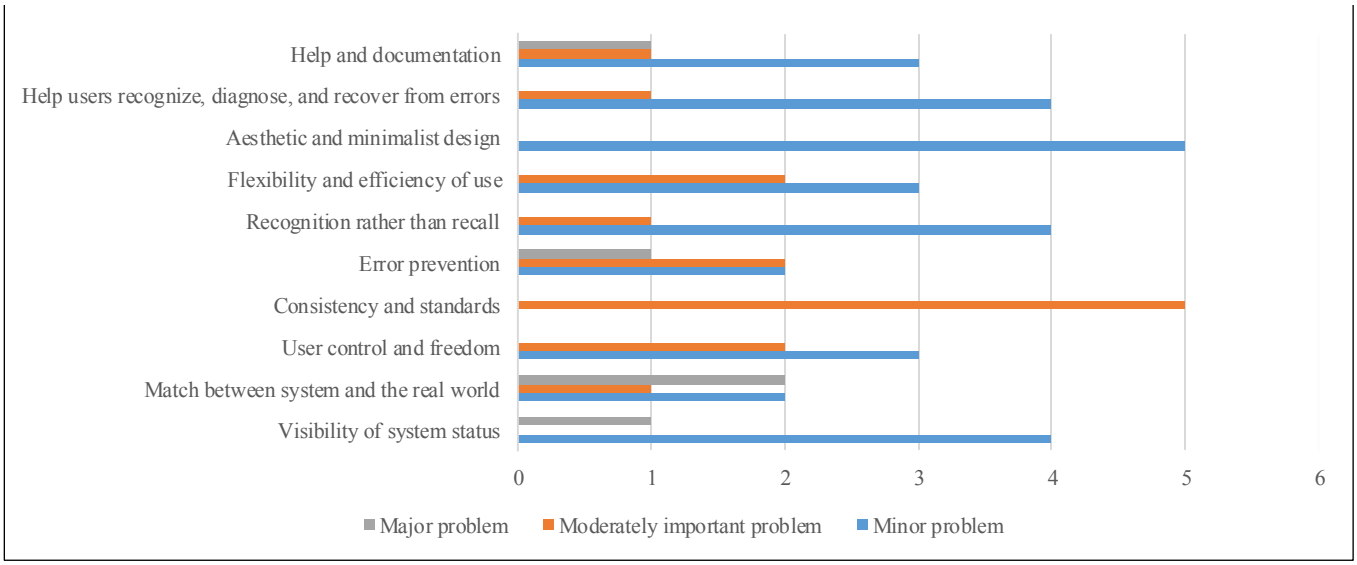

Figure 1. Participants' Evaluation of ChatGPT Interface According to "10 Usability Heuristics.

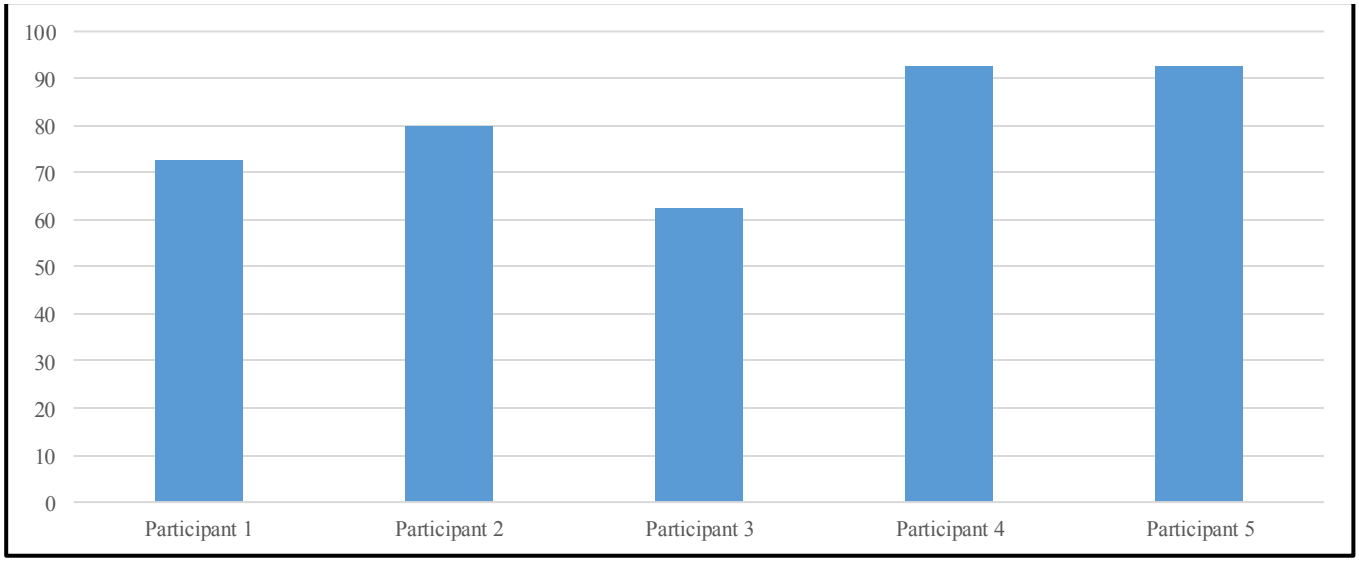

Figure 2 ChatGPT System Usability Scale Scores of Participants

IV. CONCLUSION AND FUTURE WORKS

In the scope of this study, a usability evaluation of the ChatGPT interface design was conducted. For evaluation, an expert evaluation was initially conducted, adhering to the 10 usability heuristics suggested by Nielsen [17]. The results of this expert evaluation revealed problems related to the alignment of icons or symbols utilized within the system with real-world counterparts. Illustrating this finding, a recommendation was proposed to rearrange the icons of “redo the query”, “copy the text”, and “delete the query”. Additionally, major problems were identified in assistance, documentation, and error prevention. Consequently, it is recommended to implement error messages for errors occurring during the query process and to enhance the accessibility of essential documentation for any potential assistance request. In terms of heuristics, no problems were found in the aesthetic and minimalist design.

After the expert evaluation, the participants assessed usability using the "System Usability Scale." The results of this evaluation revealed an average usability score of 80 out of 100 points. At the usability level, it was determined that users expressed a willingness to use the system. Additionally, they found the system easy to use and did not require technical support. The system was perceived as straightforward, with high user engagement. However, the evaluation also identified some inconsistencies within the system, and certain functions were not well integrated.

The study could be repeated to reassess the system by revisiting the methodology and expanding the participant pool to include a more diverse range of demographic characteristics. Conducting the study with a larger and more varied sample size could enhance its reliability and robustness. Also, the current

study focused only on text-based features of ChatGPT, the evaluation framework can be extended to newer features, including voice interaction and multimodal inputs and outputs. Additionally, replicating the study with alternative applications beyond the ChatGPT chatbot could provide valuable insights. Utilizing different usability evaluation methodologies may yield distinct findings, further enriching the understanding of usability across various contexts.

REFERENCES

[1] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[2] C. K. Lo, "What is the impact of ChatGPT on education? A rapid review of the literature," *Education Sciences*, vol. 13, no. 4, pp. 410, 2023.

[3] S. Frieder, L. Pinchetti, R. R. Griffiths, T. Salvatori, T. Lukasiewicz, P. Petersen, and J. Berner, "Mathematical capabilities of ChatGPT," in *Advances in Neural Information Processing Systems*, vol. 36, pp. 27699–27744, 2023.

[4] A. Haleem, M. Javaid, and R. P. Singh, "An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 2, no. 4, pp. 100089, 2022.

[5] I. Kostka and R. Toncelli, "Exploring applications of ChatGPT to English language teaching: Opportunities, challenges, and recommendations," *TESL-EJ*, vol. 27, no. 3, 2023.

[6] A. Koubaa *et al.*, "Exploring ChatGPT Capabilities and Limitations: A Survey," *IEEE Access*, vol. 11, pp. 118698-118721, 2023.

[7] V. Aşkun, "Sosyal Bilimler Araştırmaları İçin Chatgpt Potansiyelinin Açığa Çıkarılması: Uygulamalar, Zorluklar Ve Gelecek Yönelimler," *Erciyes Akademi*, vol. 37, no. 2, pp. 622-656, 2023.

[8] D. Gunning and D. W. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Magazine*, vol. 40, no. 2, pp. 44-58, 2019.

[9] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1-38, 2019.

[10] C. A. Nathania et al., "Analisis User Experience Penggunaan ChatGPT pada Lingkungan Pendidikan Tinggi," *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, vol. 3, no. 2, pp. 307-316, 2023.

[11] A. P. Mulia, P. R. Piri, and C. Tho, "Usability Analysis of Text Generation by ChatGPT OpenAI Using System Usability Scale Method," *Procedia Computer Science*, vol. 227, pp. 381-388, 2023.

[12] T. Sakirin and R. B. Said, "User preferences for ChatGPT-powered conversational interfaces versus traditional methods," *Mesopotamian Journal of Computer Science*, 2023, pp. 24-31.

[13] E. A. Emirhan and H. Jieun, "AI-Based UX Assessment: The Role of GPT-4 Vision in UX/UI Comparison and Heuristic Evaluation," in Proc. *HCI Korea Conf.*, 2024, pp. 272–276.

[14] E. Kuang, M. Li, M. Fan, and K. Shinohara, "Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2024, pp. 1–16.

[15] K. Çağıltay, İnsan Bilgisayar Etkileşimi ve Kullanılabilirlik Mühendisliği: Teoriden Pratiğe (1. b.). Ankara: ODTÜ Yayıncılık, 2011.

[16] J. Nielsen, *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann, 1993.

[17] Nielsen, J. "Ten Usability Heuristics," Semantic Scholar, [Online]. Available: https://pdfs.semanticscholar.org/5f03/b251093aee730ab9772db2e1a8a7eb8522cb.pdf. [Accessed: May 12, 2025].

[18] J. Brooke, "SUS: a 'quick and dirty' usability," *Usability Evaluation in Industry*, vol. 189, no. 194, pp. 4-7, 1996.