Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing

Emily Kuang
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, USA
ek8093@rit.edu

Minghao Li
School of Computer Science and Engineering
Nanyang Technological University
Singapore
minghao002@e.ntu.edu.sg

Mingming Fan*
Computational Media and Arts Thrust
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science and Technology
Hong Kong SAR, China
mingmingfan@ust.hk

Kristen Shinohara*
School of Information
Rochester Institute of Technology
Rochester, New York, USA
kristen.shinohara@rit.edu

ABSTRACT

Usability testing is vital for enhancing the user experience (UX) of interactive systems. However, analyzing test videos is complex and resource-intensive. Recent AI advancements have spurred exploration into human-AI collaboration for UX analysis, particularly through natural language. Unlike user-initiated dialogue, our study investigated the potential of proactive conversational assistants to aid UX evaluators through automatic suggestions at three distinct times: before, in sync with, and after potential usability problems. We conducted a hybrid Wizard-of-Oz study involving 24 UX evaluators, using ChatGPT to generate automatic problem suggestions and a human actor to respond to impromptu questions. While timing did not significantly impact analytic performance, suggestions appearing after potential problems were preferred, enhancing trust and efficiency. Participants found the automatic suggestions useful, but they collectively identified more than twice as many problems, underscoring the irreplaceable role of human expertise. Our findings also offer insights into future human-AI collaborative tools for UX evaluation.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Natural language interfaces**; **Usability testing**.

*Corresponding authors

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05.
https://doi.org/10.1145/3613904.3642168

KEYWORDS

User experience; Usability testing; Human-AI collaboration; Proactive conversational assistants

ACM Reference Format:
Emily Kuang, Minghao Li, Mingming Fan, and Kristen Shinohara. 2024. Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642168

1 INTRODUCTION

Usability testing is a widely embraced user-centered design approach to detect usability issues in interactive systems [13, 22, 27, 67]. However, analyzing usability test recordings is challenging due to its labor-intensive, time-consuming, and intricate nature. UX evaluators must interpret behavioral signals from visual and audio channels while rapidly assessing multiple tasks to identify usability issues [13]. Constraints in time and resources within industrial settings may result in overlooked information or misinterpretation of usability problems [27, 35, 53, 67]. Despite the potential for collaborative analysis between multiple human evaluators to enhance reliability and completeness, practical resource constraints often discourage UX evaluators from adopting this approach [26, 27, 53].

Recent AI advancements have led researchers to investigate how to employ AI-driven analysis to provide complementary perspectives to UX evaluators [20, 24, 51, 52, 78]. Responding to a call by usability pioneer Jakob Nielsen to incorporate AI into UX research [66], we sought to explore a form of human-AI collaborative usability analysis via conversational assistants (CAs) given its growing popularity. CAs are a promising interface paradigm for usability analysis, as they can act as an assistant to UX evaluators by providing desired information about usability videos [52]. Past systems utilized *user-directed dialogue*, where the CA only responded to user questions [52]. However, the implications of

proactive or system-directed dialogue, in which messages from the CA appear without user initiation, remain unknown. Our research addresses this limitation by exploring how a proactive CA may enhance the analysis process by automatically providing suggestions of usability problems. Since the CA initiates interaction with the user, understanding the *optimal timing* for when these suggestions should appear is critical [56]. Prior work indicated that synchronous suggestions—appearing at the *beginning* of usability problems—were more trusted and better received than asynchronous suggestions that were shown continuously on a timeline [24]. Participants also proposed an alternative: presenting suggestions *after* the occurrence of a usability problem, since synchronous suggestions may lead to confusion and confirmation bias as participants did not independently assess the video segment first [24]. Thus, it remains unclear which of the timing factors is best during the analysis of usability videos. We investigate the efficacy of proactive dialogue in UX evaluation tools, with a specific focus on determining the optimal timing for automatic suggestions—whether it be before, synchronously, or after the occurrence of potential usability problems. Understanding the optimal timing will inform the design and functionality of future tools, ultimately enhancing AI-assisted decision-making in UX evaluations.

To investigate how the timing of automatic suggestions may impact UX evaluators' analysis behaviors, we first needed to generate these usability problem suggestions from usability test videos. The emergence of ChatGPT, a versatile generative AI tool driven by a large language model (LLM) [70], has garnered substantial interest across diverse domains [83]. According to recent surveys of UX researchers, ChatGPT is already being used in aspects of UX research like creating user flows and developing usability test protocols [7, 12, 47, 82]. Considering ecological validity, we also utilized ChatGPT to identify potential usability problems from the transcripts of usability test videos. Additionally, we assessed the quality of ChatGPT's suggestions with manual analysis conducted by three UX experts to ensure that they were sufficiently reasonable to be used in the study. At the time of designing this study, ChatGPT had text input limitations and lacked access to video content, making it potentially incapable of addressing queries related to users' interactions with the interface or product information—topics of interest to UX evaluators [52]. To account for this limitation, we adopted a hybrid Wizard of Oz approach, wherein ChatGPT generated usability problem suggestions, and a moderator addressed impromptu questions.

Inspired by prior work showing that AI suggestions appearing at the start of a problem led to enhanced user engagement and higher acceptance compared to asynchronous AI suggestions [24], we were motivated to investigate when suggestions appeared with respect to the potential usability problem (e.g., before, during, or after). In understanding the influence of timing, we used ChatGPT to generate suggestions. Further, to position the ecological validity of this approach, we analyzed the quality of the suggestions produced. Specifically, our study was guided by the following research questions (RQs):

- RQ1 - How does the *timing* of automatic AI-generated suggestions impact UX evaluators'
  - A) analytic performance (e.g., number of problems)?
  - B) subjective perceptions (e.g., efficiency, user trust)?
- RQ2 - After receiving automatic AI-generated suggestions, how do UX evaluators
  - A) respond to these suggestions (e.g., agreement or disagreement)?
  - B) assess the quality of these suggestions (e.g., level of agreement, completeness)?

Our within-subjects study with 24 UX evaluators revealed that the timing of automatic suggestions did not significantly influence the number of identified problems. However, most participants favored suggestions appearing after potential usability problems, which significantly enhanced trust and efficiency. Following an automatic suggestion, participants responded in one of four ways: 1) agreement, 2) correction, 3) request for clarification, or 4) disagreement or disregard of the suggestion. Participants accepted 77.6% of all suggestions, irrespective of timing considerations. Participants felt that the usability problem suggestions from ChatGPT acted as warnings and validation of their analysis, but they still relied on their expertise to identify a more comprehensive catalog of usability problems. In light of our findings, we deliberate on the timing and functionality of automatic suggestions, and the perceptions of AI capability. In summary, our contributions encompass:

- Demonstrating that efficiency, trust, and user preference are enhanced when providing suggestions after potential usability problems;
- Presenting the various responses of UX evaluators and their agreement levels with automatic suggestions;
- Highlighting the constraints of transcript-based UX analysis approaches and suggesting ways to enhance future CAs.

2 RELATED WORK

Our work is informed by prior research on using AI to detect usability problems and human-AI collaboration via interactive conversational assistants.

2.1 Using AI to Detect Usability Problems

Usability testing is the most common method for identifying usability issues in digital products [22]. UX evaluators simultaneously observe user actions and make notes while assessing both the visual and audio aspects of videos [13]. Analyzing usability test videos manually is time-consuming and challenging due to limited time and resources, which can result in missed information or misinterpreted problems [23, 27, 35, 67]. To address this challenge, researchers have explored two main approaches: automated methods without human involvement and human-AI collaborative methods for detecting usability problems.

2.1.1 Automated Methods for Usability Problem Detection. Researchers have employed several automatic detection methods for usability problems, including machine learning and pattern recognition, audio and visual analysis, and natural language processing with sentiment analysis.

Machine learning and pattern recognition: Researchers have developed machine learning (ML) classifiers using user interaction data from websites and mobile applications [31, 34, 43,

72, 73, 87, 90]. For instance, researchers employed mouse movements and interaction patterns like the "action repetition pattern" to classify sessions as having usability issues or not [30, 65, 76]. Additional studies demonstrated the applicability of eye tracking and gaze analysis for detecting usability problems, particularly minor navigational and comprehension issues [18, 69].

Audio and visual analysis: In addition to analyzing interaction logs, researchers have investigated the direct detection of usability problems from the video and audio of usability tests [20, 21, 79]. Specific speech patterns were discovered to indicate usability problems such as abnormally low speech rate [20], and temporal video segmentation could automatically detect segments where users encountered difficulties operating the product [79].

Natural language processing and sentiment analysis: Natural language processing techniques are applicable for extracting usability problem indicators from usability test transcripts, particularly when the user's verbalizations include negative sentiments, questions, and verbal fillers [20, 21]. Beyond traditional text-based methods, the emergence of ChatGPT [70] and other large language models (LLMs) has expanded the horizons of research approaches [83], demonstrating versatility in summarizing literature, drafting papers, coding, and passing medical licensing exams [54, 83]. Specifically, it has demonstrated utility in sentiment analysis [44, 84], as well as the analysis of questionnaire responses, interview data, and think-aloud data [47, 80]. Meta-summaries it generated showed substantial but not entirely identical alignment with analyses conducted by an independent researcher, suggesting that ChatGPT can be a reliable tool for text data analysis when used cautiously [80]. While ChatGPT has been employed in various aspects of UX research [12, 47], there remains a gap in understanding its effectiveness in identifying usability problems. Consequently, our strategy of utilizing ChatGPT to generate suggestions for usability problems required an assessment of the quality of its outputs.

2.1.2 Human-AI Collaborative Methods for Usability Problem Detection. Due to the limitations of automated methods, there is a growing interest in human-AI collaboration, where AI supplements human decision-making [55]. AI outcomes can inform individuals through an "algorithm-in-the-loop" process, recognizing that AI systems should support rather than replace domain workers' decisions and tasks [29, 88]. The 10 Levels of Automation framework by Mackeprang et al. describes various AI involvement levels, ranging from providing no assistance to offering suggestions, executing suggestions with human approval, informing humans after decision execution, and finally, acting autonomously [61]. Recent research on human-AI collaborative UX evaluation tools primarily focuses on the lower end of this framework, where AI provides usability problem indicators, leaving the final determination of usability problems to UX evaluators [23, 24, 78]. However, these tools offer non-interactive visualizations (e.g., icons and line charts), limiting UX evaluators' ability to ask questions about usability test videos or seek explanations for AI results. Therefore, there is an opportunity for human-AI collaborative tools to enhance interactivity and provide explanations on-demand.

2.2 Human-AI Collaboration Via Interactive Conversational Assistants

Conversational assistants, including chatbots, have gained prominence in various aspects of our daily lives [59]. Their utilization has witnessed significant growth, with chatbots emerging as one of the fastest-growing communication channels [64]. In 2022, surveys found that 88% [25] and 80% [81] of respondents had interacted with a chatbot, demonstrating their growing popularity. Researchers have explored their application across diverse domains, spanning from business documents [39] and collaborative games [4, 6] to customer services [5], journaling, and productivity applications [33, 48]. Our recent research has extended the application of CAs to the field of UX analysis. This exploration involves understanding how CAs can aid in UX analysis by addressing questions raised during analysis [52]. Building upon our efforts, this study focuses on text-based conversational assistants, chosen for their demonstrated efficiency over voice assistants, and leverages the dataset of potential questions to inform the design of our UX assistant [52].

2.2.1 Benefits of Proactive Dialogue. Conversational interactions can be categorized into three types: (1) **user-directed dialogue**, where the user initiates a question and the system reacts; (2) **system-directed dialogue**, where the system initiates an interaction; and (3) **mixed-initiative dialogue**, where both the system and the user have control over the dialogue flow [57, 62]. In our previous study, the Q&A dynamic was user-directed, where the CA only responded when prompted [52]. In contrast, system-directed or proactive dialogue has shown promise in providing timely assistance and enhancing user trust [49]. Older adults reported feeling significantly less lonely and more satisfied when using a proactive CA than the passive version since it started conversations and provided a sense of companionship [74]. Similarly, a proactive agent designed to enhance learning for undergraduate students led to a significantly more positive impact on recall than the passive agent since it actively provided information about the lesson [46]. Participants also appreciated proactive interactions to improve health regimen adherence since they acted as reminders to engage in a healthy activity [9]. Furthermore, providing recommendations through proactive dialogue could facilitate user's item selection in large decision spaces [15]. In sum, proactive interactions can alert users of information that they may have missed otherwise and provide recommendations to inform decisions, which can be useful when applied in UX evaluation to actively guide UX evaluators in identifying usability problems.

Despite the advantages of proactive interactions, they also pose concerns, such as being perceived as disruptive if initiated at an inappropriate time, leading to reduced transparency and distrust in the system [10, 41, 56, 68]. For instance, in a 17-day field study where users interacted with a personal AI agent for work, some participants exhibited an aversion to unsolicited proactive interactions [56]. Similarly, when UX evaluators are focused on reviewing a usability video, they may not want to be interrupted by messages from a CA. To mitigate this issue, we should: 1) *reduce their interruption cost by investigating the optimal timing of the interactions*; and 2) *increase their value by ensuring the content of the messages is relevant* [50, 56, 86].

2.2.2 Investigating the Timing Factor. To determine when proactive interactions occur, two possibilities exist: variable timing, inferred from human interaction data and task context [32, 58], and fixed timing, initiated at specific task steps [24, 49]. Previous research on UX evaluation adopted fixed timing, displaying AI results at the beginning of each usability problem [24]. This approach led to enhanced engagement, higher acceptance of AI suggestions, improved problem identification, and increased satisfaction [24]. Participants also suggested alternative timings, such as immediately after identifying a problem [24]. Given these findings, fixed timing, aligned with specific problem occurrences, was deemed suitable for our study, offering three possibilities: before, in sync with, and immediately after a potential usability problem.

2.2.3 Determining Dialogue Content. Dialogue content for proactive interactions varies across different proactivity levels: none, notification, suggestion, and intervention. Research indicates that notification and suggestion levels tend to foster higher perceived reliability and user trust [49]. Notifications inform users that the system has found a solution, while suggestions provide recommendations along with explanations [49]. Previous UX evaluation studies underscored the value of explanations, enhancing UX evaluators' analysis support and perception of AI [24]. Focusing dialogue content on the specific task of identifying usability problems is essential, as straying from this purpose could lead to negative user experiences [45]. Hence, our study adopts the "suggestion" strategy, wherein the proactive UX assistant provides focused recommendations of usability problems and explanations for what occurred.

3 DESIGNING THE PROACTIVE UX ASSISTANT

Building upon the notion that proactive agents can offer timely assistance [49], our primary interest lay in investigating how the timing of proactive interactions can enhance the UX analysis process. The proactive UX assistant could provide two types of messages:

(1) **Automatic suggestions** either before, in sync, or after the occurrence of a usability problem.
(2) **Reactive responses** to various questions posed by UX evaluators.

We adopted a hybrid approach, in which we used ChatGPT to generate automatic suggestions, while impromptu responses were provided by a human moderator, following a Wizard of Oz design. This approach has historically been employed to circumvent technical limitations in prior research (e.g., [60, 63, 77]).

3.1 Automatic Suggestion Generation with ChatGPT

To explore the most effective timing for usability problem suggestions, our first task involved generating these suggestions from usability test videos. In this process, we employed ChatGPT to simulate the results obtained by having a CA analyze the transcript, as it is the closest instantiation of one. Zoom's automatic transcription feature was used to initially transcribe the verbal content of usability videos, which included the instructions from the usability test moderator and the verbalizations from the user engaged in the tasks. Subsequently, a researcher reviewed these transcripts to correct errors, add punctuation, and edit timestamps to denote natural speech breaks. To preserve the originality of the dialogue, no additional alterations were made to the transcripts. In early 2023, we used ChatGPT (version 3.5) [71] to generate suggestions of usability problems by entering a prompt that contained the usability tasks and transcript format, then copy and pasting the raw transcript. Below is a prompt used for one of the study videos:

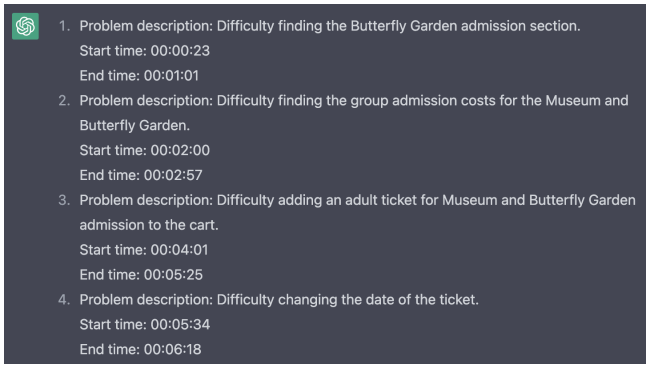

Figure 1: Screenshot of ChatGPT's response, which contains four usability problem descriptions and the start and end times of each problem.

> "The following is the transcript of a user study where a participant used the think-aloud protocol to complete the following tasks on a museum website: 1. Find the last entry time for the Butterfly Garden on a Friday, 2. Find group admission costs to the museum, 3: Add an adult ticket for Museum and Butterfly Garden admission to the cart, and 4: Update the ticket to a different date. The transcript contains the start and end timestamps and the words spoken by the participant. Based on the transcript, can you identify which usability problems the participant may have encountered and when these problems occurred? Provide your response in the format: Problem description, Start time, End time."

In response to this example prompt, ChatGPT provided a list of four usability problems and the associated timestamps (Fig. 1). This process was repeated for the three usability test videos used in the study. Table 5 in the Appendix shows the list of 14 problem suggestions from ChatGPT. To evaluate these suggestions, three UX experts with an average of 5 years of UX experience first manually analyzed the three videos independently, then engaged in a group discussion to consolidate their results and resolve any disagreements, following the recommended practice [36]. They identified 17 usability problems, of which 12 coincided with those proposed by ChatGPT. Thus, ChatGPT's precision (i.e., the proportion of correct problems among all identified problems) is 0.86, and recall (i.e., the proportion of correct problems among all correct problems) is 0.71[1].

[1]Detailed calculations are provided in Table 6 in the Appendix

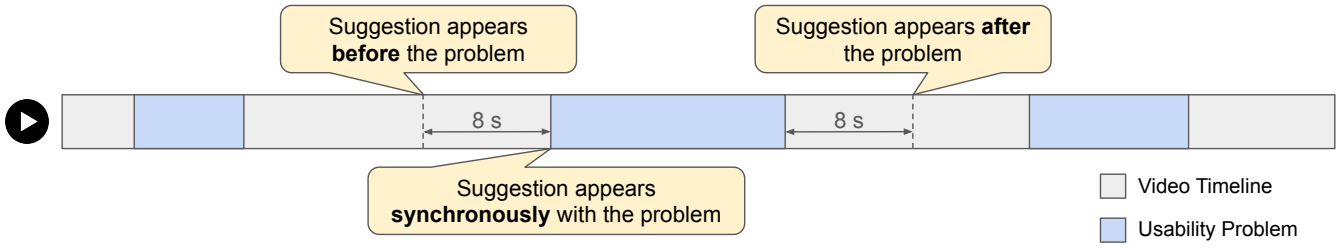

Figure 2: Video timeline illustrating the three timing conditions: 1) suggestion appears before the problem, 2) suggestion appears synchronously with the problem, and 3) suggestion appears after the problem.

This implies that, while not every problem was identified, the ones that were detected were relevant.

In our study, we opted not to directly rely on results generated by UX experts, recognizing that perfection in AI suggestions is an unrealistic expectation. For instance, previous researchers deliberately introduced false positive and false negative usability problems to make their Wizard of Oz AI system seem more realistic [24]. Compared to previous work, our approach of employing ChatGPT holds ecological validity, given that this AI tool can identify, albeit imperfectly, some usability problems based on input transcripts. Moreover, ChatGPT is commercially available and already being used for UX research according to recent surveys of UX researchers [12] and courses on AI in UI/UX design [82]. Given our evaluation, which demonstrated reasonable precision and recall, we utilized ChatGPT's suggestions in this study to investigate the timing aspects of the collaboration between human evaluators and AI. However, since ChatGPT only had access to the user's spoken words and not video actions, the start and end times of usability problem suggestions were occasionally inaccurate. To rectify this, one author reviewed and adjusted the problem timings based on video actions, ensuring more accurate timing across each condition.

3.2 Timing of the Conditions

To reduce the interruption cost of automatic suggestions, we designed three timing conditions relative to the occurrence of usability problems: **before**, **synchronous**, and **after**. Fig. 2 shows a video timeline with annotations of suggestions that appear before, synchronously, and after a usability problem. We employed a heuristics-based approach to determine the time gap between the suggestion and the problem. In the "synchronous" condition, automatic suggestions were displayed exactly at the start of the current problem. In the "before" condition, the gap was calculated as the minimum of 8 seconds or the time interval between the end of the previous problem and the start of the current one ($\min(8, start_{current} - end_{previous})$). Similarly, in the "after" condition, the gap was determined as the minimum of 8 seconds or the time interval between the end of the current problem and the start of the next one ($\min(8, end_{current} - start_{following})$). The use of 8 seconds aligns with the average human attention span [16]. This approach also ensured that suggestions did not appear while another problem was ongoing.

The tense of the automatic suggestions varied depending on the timing condition. For example, the first problem identified by ChatGPT was shown as:

(1) **Before** condition: *"I think a problem **will occur** because the user **will have** difficulty finding the Butterfly Garden admission section."*
(2) **Synchronous** condition: *"I think a problem **is occurring** because the user **is having** difficulty finding the Butterfly Garden admission section."*
(3) **After** condition: *"I think a problem **just occurred** because the user **had difficulty** finding the Butterfly Garden admission section."*

3.3 Generating Responses using Wizard of Oz

In addition to proactively suggesting potential usability problems, the UX assistant was designed to provide responses to questions from UX evaluators. This approach aligns with user expectations based on common CAs (e.g., Siri, Google Assistant) [59]. Providing responses to impromptu questions did not interfere with the investigation of the timing of automatic suggestions since the quality of responses remained consistent across all conditions. To anticipate the types of questions expected from participants, we referred to prior research that categorized the questions UX evaluators typically ask a CA [52]. Examples of such questions include *"How many clicks did the user have to go through to reach the target page?"* and *"What is the ideal path to complete this task?"* Thus, we extracted information related to user actions, the user's mental model, help from the AI assistant, product and task details, and user demographics from the videos, and used these notes during the study to answer any questions that arose.

3.4 User Interface of the UX Evaluation Tool

The user interface (UI) maintains simplicity and informativeness, consistent with previous tools employing conversational agents for UX analysis [52]. It comprises two main components: a video player for UX evaluators to review usability test videos (Fig. 3-A), and a chat window displaying the conversation between UX evaluators and the UX assistant (Fig. 3-B). The video player includes standard playback controls, such as play/pause, volume adjustments, a progress bar, and speed options. The progress bar segments tasks in each video, displaying the current task at the bottom of the video player (Fig. 3-a1). The chat window is positioned to the right of the

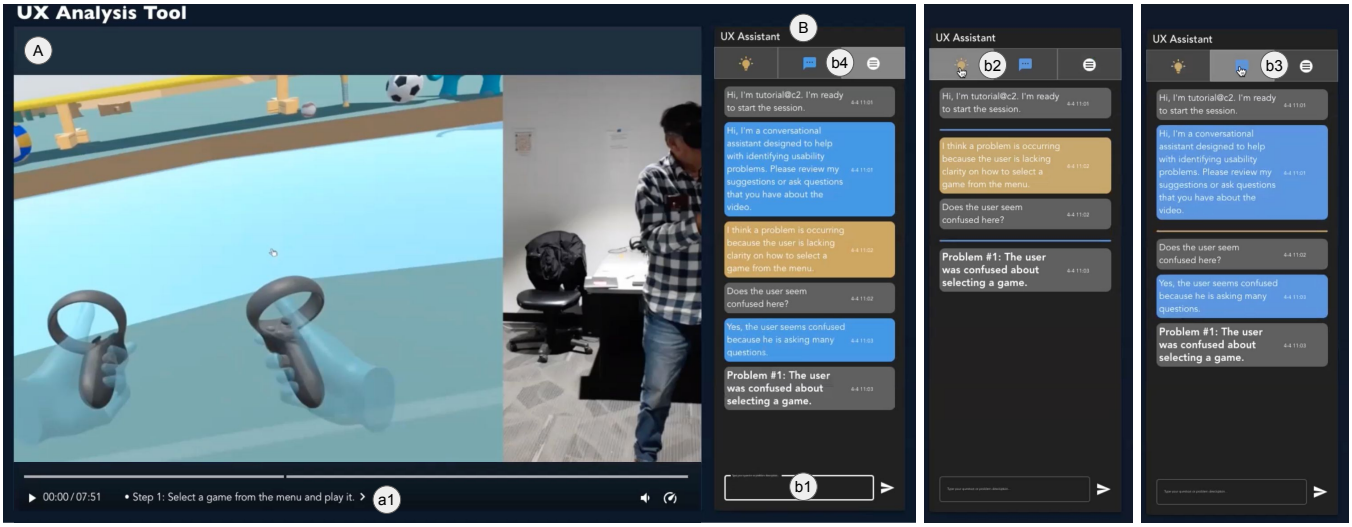

Figure 3: User interface of the UX Analysis Tool: (A) Video player, (a1) Progress bar, (B) Chat thread, (b1) Chatbox, (b2) Show suggestions, (b3) Show messages, and (b3) Show all.

video player and contains a chatbox at the bottom (Fig. 3-b1), where UX evaluators can type responses to automatic suggestions, descriptions of usability problems, and questions to the UX assistant. The moderator receives these messages in another chat window accessible only via an administrator account and password. Messages from the UX assistant appear in two formats: 1) Automatic suggestions are displayed in blue and UX evaluators can click on these speech bubbles to navigate the video player back to the timestamp when the suggestion first appeared, and 2) Responses to questions appear in yellow. UX evaluators can focus on reviewing one message type at a time by collapsing the other type using the respective icons (e.g., "lightbulb" for automatic suggestions (Fig. 3-b2), "message" for responses (Fig. 3-b3)). To return to the default mode with all bubbles expanded, they can click the "hamburger" icon Fig. 3-b4).

4 USER STUDY

We conducted an IRB-approved within-subjects experiment in April and May 2023. A single moderator served as the UX assistant to respond to impromptu questions.

4.1 Participants and Apparatus

We recruited 24 participants aged 22 to 43 ($M = 27.1, SD = 4.8$) with self-reported UX experience ranging from 2 to 11 years ($M = 3.9, SD = 2.2$). Participants indicated their familiarity with usability analysis, with most selecting "4 - very familiar" on a 5-point scale. All had prior experience with conversational assistants (e.g., Siri, Google Assistant, ChatGPT), and their median trust in AI algorithms or AI-powered conversational assistants was "3 - moderately trustful" on a 5-point scale. Participants completed the study remotely, using computers to access a web application and communicating with the moderator via Zoom.

4.2 Study Videos

Since there is currently no established taxonomy of products or tests that need to be covered during studies of UX analysis tools, we selected some examples to prompt analysis, which follows prior work (e.g., [23, 24, 78]). Although three videos can not be representative of all usability tests or tasks, we covered common digital interfaces (desktop website, smartphone app, and VR headset). Table 1 provides video details, including length, tasks, and the number of usability problem suggestions. The tasks were designed around the central functions of the product (e.g., ordering food in a food delivery app), and the number of tasks was chosen so that all three videos would have similar lengths. Before recording usability tests with users, the researchers tried to complete the tasks to ensure that usability issues existed in these products and to gain an estimate of the time required. Each resulting video contained at least four usability problem suggestions, enabling participants to engage in analysis and receive suggestions from the UX assistant.

4.3 Procedure

Fig. 4 illustrates the study session procedure, where participants initially received a brief introduction to the study and a tutorial on the web application, including information about the two types of messages from the UX assistant: automatic suggestions and responses to questions. Any questions regarding tasks or the web application were addressed before proceeding. As a within-subjects study, each participant analyzed all three videos with the order and conditions being counterbalanced. For each video, the moderator explained the scenario and tasks. Participants then engaged in the analysis session, during which suggestions automatically appeared when their video player reached predetermined timestamps based on the timing condition. All participant interactions occurred in the chatbox (Fig. 3-b1), where they typed *"I agree"* or *"I disagree"* in response to a suggestion, or ignored it completely. To record a usability problem, they prefaced their description with *"Problem*

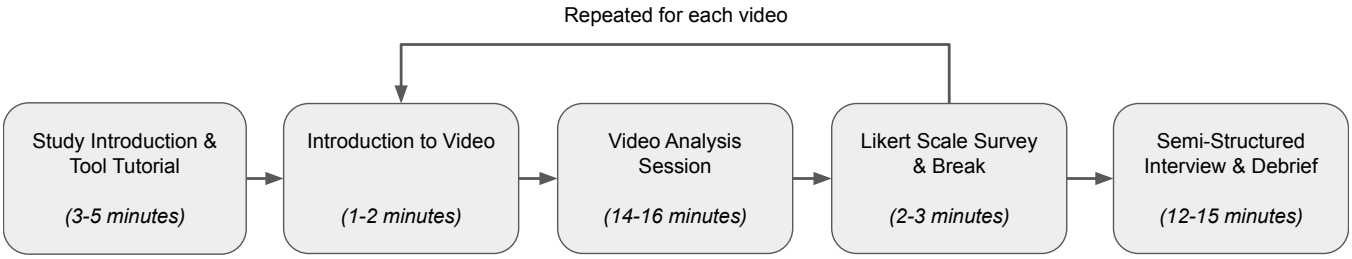

Figure 4: Flow chart showing the study procedure.

Table 1: Information on the videos used in the study

Product	Tasks	Length (m:ss)	Number of Suggestions
Museum Website	1: Find the last entry time for the Butterfly Garden. 2: Find group admission costs to the museum. 3: Add an adult ticket for the museum and Butterfly Garden admission to the cart. 4: Update the ticket to a different date.	6:54	4
Food Delivery App	1: Buy Coke, Sprite, and pizza within a $100 budget. 2: Change from pick-up to delivery.	8:01	5
VR Game	1: Select a game from the menu and play it. 2: Select a different game and play it.	7:51	5

#:" (Fig. 3-B). After analyzing each video, they completed a Likert scale survey, assessing efficiency, trust, preference, cognitive effort, satisfaction, and helpfulness, in line with prior work on CAs [52, 89]. Following each survey, participants had a short break before the moderator explained the scenario and tasks in the next video. After analyzing all three videos, participants underwent a semi-structured interview covering their experience, insights into suggestion timing, and ChatGPT's potential for usability analysis. Sessions, lasting 60-80 minutes, were video-recorded, and participants were compensated for their time.

4.4 Data Analysis

Each usability problem description was assigned a label to identify the number of unique problems and determine whether they coincided with ChatGPT's results. The questions posed by participants were coded and divided into categories. The semi-structured interview responses were transcribed using automatic speech-to-text software and then corrected by a researcher. Two researchers analyzed the transcripts using inductive coding and then grouped the codes into themes through iterative discussions.

Table 2: Number of problems identified by participants in each condition

Condition	Mean (SD) per video	Total
Before	4.4 (1.6)	105
Synchronous	4.6 (1.7)	111
After	5.3 (1.6)	126

For the quantitative measures collected in the study (e.g., number of problems identified, survey responses, percent agreement with ChatGPT suggestions), we used the Shapiro-Wilk test to check the normality of the data. Then we conducted a one-way repeated measures ANOVA to determine whether the means for the three timing conditions were significantly different. We also report the effect size with partial eta squared (η_p^2) and post-hoc pairwise comparisons with Bonferroni correction.

5 RESULTS

This section presents the findings regarding the impact of timing on participants' analytical performance and subjective perceptions (RQ1) and explores their responses to automatic suggestions and the comparison of analysis results (RQ2).

5.1 Impact of Timing on Analytic Performance and Subjective Perceptions (RQ 1)

This section describes the number of usability problems identified and participants' subjective feedback for each timing condition.

5.1.1 Analytic Performance (RQ 1.A). In total, participants recorded 342 usability problems. Table 2 shows the mean and standard deviation of problems per video and the total number of problems by condition. On average, participants identified 4.4 problems per video in the "before" condition, 4.6 problems in the "synchronous" condition, and 5.3 problems in the "after" condition. Although the "after" condition yielded the highest average number of identified problems, the differences between the various conditions were not statistically significant. These findings indicate that the timing of automatic suggestions did not significantly influence participants' analytic performance.

5.1.2 Subjective Feedback about Timing Conditions (RQ 1.B). As illustrated in Fig. 5, efficiency, trust, and preference demonstrated

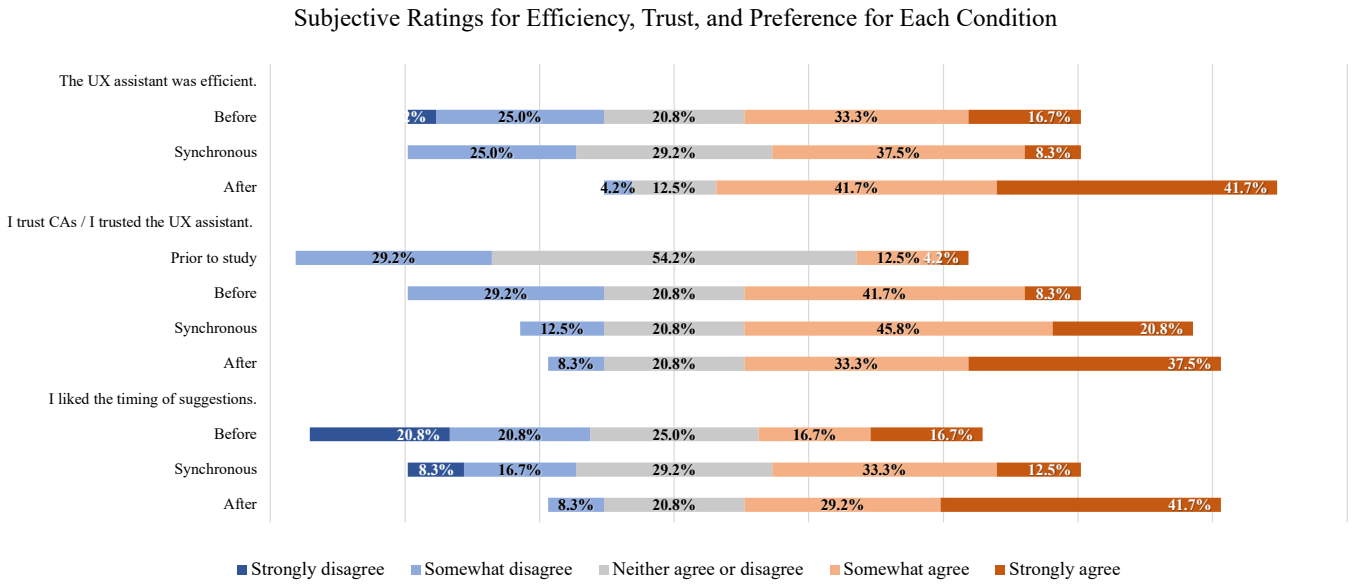

Figure 5: Diverging stacked bar chart that shows the participant ratings for efficiency, trust, and preference for each condition.

significant differences due to timing, while cognitive effort, satisfaction, and helpfulness (Fig. 6) did not exhibit any significant effects of timing. Among the participants, 14 out of 24 (58.3%) favored the "after" condition, 8 (33.3%) preferred the "synchronous" condition, and only 2 (8.3%) chose the "before" condition as their most preferred.

In the survey, participants responded to the phrase *"I liked that the UX assistant gave me suggestions [prior to/in sync with/after] the occurrence of usability problems."* Ratings were highest for the "after" condition ($Md = 4, IQR = 2$), followed by "synchronous" ($Md = 3, IQR = 1.25$), and then "before" ($Md = 3, IQR = 2$). ANOVA showed a main effect of timing ($F_{2,46} = 7.0, p < .01, \eta_p^2 = 0.2$), and pairwise comparisons with Bonferroni correction revealed that the "after" condition was significantly higher than the "before" condition ($p < .01$). The following sections expound on the advantages and disadvantages associated with each timing condition.

"Before" Timing Condition: Participants who preferred the "before" timing condition felt that the suggestions acted as warnings of upcoming problems. For instance, P16 stated, *"The suggestions made me more focused and alert, I paid more attention to the following video segment after I saw the message."* However, some participants believed that seeing suggestions before the video segment could introduce bias. P5 commented, *"I have to develop my own understanding of the context first and then check my understanding, seeing the suggestion before could bias my opinion."* Moreover, participants found it inefficient when they clicked on a suggestion to navigate to the corresponding timestamp, which took them back before the problem occurred, forcing them to watch extra seconds.

Additionally, some participants felt confused, thinking that the UX assistant was predicting problems in the future. P23 questioned, *"I was confused because the UX assistant hasn't even seen that part of the video yet, how did it know a problem will occur?"* This misconception likely contributed to lower trust ratings for the "before" condition. Notably, participants gave the highest ratings for trust in the "after" condition, followed by the "synchronous" and "before" conditions. ANOVA demonstrated a main effect of timing ($F_{3,69} = 8.4, p < .0001, \eta_p^2 = 0.3$), and pairwise comparisons with Bonferroni correction revealed that the "before" condition was significantly lower than the "after" condition ($p < .05$).

"Synchronous" Timing Condition: Participants who favored the "synchronous" timing condition believed that seeing the suggestions concurrent with the occurrence of the problem instilled confidence that the UX assistant was detecting issues at the same time they did. P7 expressed, *"I preferred this since it felt the most synchronous, and I was more trusting that it was actually noticing something."* This condition was also perceived as efficient, with P13 mentioning, *"it felt the smoothest and I wrote "I agree" the most out of all three conditions, which saved time from typing problem descriptions."* For participants who crafted their own descriptions, the suggestions aided in phrasing, as noted by P19, who said, *"the suggestion was better than how I would phrase it so seeing it at the beginning helped me write a better problem description."*. However, some participants reported that the suggestions were distracting and diverted their attention from the video. P18 stated, *"I was more focused on reading the text, which made me distracted and thought I missed something in the video."*

"After" Timing Condition: The "after" condition was preferred by most participants for three primary reasons: validation, similarity to current processes, and improved efficiency. Many participants mentioned that seeing suggestions after the problem occurred *"created validation of my own analysis and reinforced my confidence"* (P9) and that *"the timing was aligned with when I reached the conclusion"* (P4). Furthermore, it felt the most intuitive because *"it is most*

similar to working with a colleague: both of you do an individual analysis first, then combine your opinions to create validation" (P21). Efficiency was also a key factor since ANOVA showed a main effect of timing ($F_{2,46} = 8.4, p < .001, \eta_p^2 = 0.3$). Pairwise comparisons with Bonferroni correction revealed that the "after" condition was rated as significantly more efficient than both the "before" ($p < .01$) and "synchronous" ($p < .01$) conditions. A substantial 83.3% of participants somewhat or strongly agreed that the "after" condition was efficient for their analysis, compared to 50.0% and 45.8% in the "before" and "synchronous" conditions, respectively.

While some participants felt the timing of the suggestions matched their analysis, others felt that it occurred later than they preferred. For example, P10 said *"the suggestions appeared too late for me, after reading it, I have to rewind the video if I missed that problem."* This suggests that individual differences and level of UX expertise may influence the timing of analysis practices, as discussed in Section 6.1.1. Interestingly, P13 mentioned *"since I already identified the problem, it doesn't feel very helpful."* This suggests that they were confident in their analysis and were not open to validation.

5.2 Participants' Responses to and Agreement with Automatic Suggestions (RQ 2)

In this section, we first describe participants' responses to automatic suggestions and subsequently calculate their level of agreement with the automatic suggestions generated by ChatGPT.

5.2.1 Participants' Responses to Automatic Suggestions (RQ 2.A). Participants presented four distinct response types to the automatic suggestions: 1) agreement, 2) correction, 3) clarification, and 4) disagreement or disregard.

1) Agreeing with the Suggestion: When participants identified the same usability problem as suggested by the UX assistant, the predominant response was affirmation, expressed through messages such as *"Yeah, I agree"* (P1) and *"I agree, make that problem #2"* (P3). Participants felt that the suggestion served as a confirmation of their analysis. For example, P17 noted that *"when I saw the suggestion, it acted as a confirmation of my analysis, and I thought, 'Yeah, it got me!' So I wrote 'I agree' to support the suggestion."*

2) Correcting the Suggestion: Participants occasionally concurred that a usability problem existed but believed that the suggestion did not comprehensively or accurately describe the issue. For instance, a suggestion indicated that the user had trouble finding the "drinks" category, but P10 responded, *"I don't think the user had difficulty finding the drinks section, he was able to find it but had trouble finding the Coke inside this section."* In such cases, participants acknowledged the presence of a problem but amended the problem description to align with their assessment.

3) Seeking Clarification: Another approach taken by participants was to request additional clarification. For instance, P11 inquired, *"Why do you think this is a usability problem?"* When asked about this, P11 explained, *"I think the suggestion was correct, but it was incomplete. So I asked a follow-up question about why, and the assistant did give me valid reasons."* In such instances, participants sought further information about an automatic suggestion to confirm the existence of a usability problem.

4) Disagreeing or Disregarding the Suggestion: When participants disagreed with a suggestion, they either responded with dissenting comments (e.g., *"No, this is not a problem."*) or chose to disregard the message in the chat window. Participants cited two main reasons for not seeking further clarification: a stronger trust in their own intuition, and a belief that the UX assistant lacked the intelligence to provide meaningful answers.

Table 3: Percentage of usability problem suggestions that participants agreed with based on timing condition

Condition	Percentage Agreement	SD
Before	72.9%	21.7%
Synchronous	71.9%	24.0%
After	88.1%	15.7%
Total	77.6%	21.8%

5.2.2 Participants' Assessment of Usability Problem Suggestions (RQ 2.B). In this subsection, we describe the extent to which participants agreed with the usability problem suggestions generated by ChatGPT. We also explore the characteristics of suggestions that resulted in the highest and lowest agreement, as well as the types of problems that were missed.

Percentage Agreement Calculation: The percentage agreement was calculated as the number of suggestions that participants agreed with out of the 14 suggestions presented across three videos. For instance, if a participant acknowledged 11 out of 14 suggestions as actual usability problems, the percentage agreement would be calculated as $11/14 \times 100 = 78.6\%$. Table 3 shows the percentage agreement for each condition, while Table 5 in the Appendix shows the percentage agreement for each suggestion.

Overall Agreement: Collectively, participants agreed with 77.6% ($SD = 21.8\%$) of the 14 suggestions. The highest agreement occurred in the "after" condition (88.1%), followed by the "before" condition (72.9%), with the "synchronous" condition trailing at 71.9%. ANOVA exhibited a significant main effect of timing ($F_{2,46} = 4.6, p < .05, \eta_p^2 = 0.2$). Further pairwise comparisons with Bonferroni correction indicated that the "after" condition demonstrated significantly higher agreement compared to both the "before" ($p < .05$) and "synchronous" ($p < .05$) conditions, while the latter two did not exhibit significant differences.

Variability in Agreement Among Suggestions: Apart from variations in agreement based on timing conditions, we were curious if any differences existed between individual suggestions. We observed that ten out of the fourteen suggestions garnered acceptance rates ranging from 75% to 100% among participants, while one suggestion achieved a 66.7% acceptance rate. Three suggestions, however, had lower acceptance rates, ranging from 33.3% to 41.7% of participants. Upon analyzing suggestions with high agreement, it became evident that these problems were typically straightforward and task-oriented. Examples include users encountering difficulties in locating a specific element within the interface (e.g., selecting a pizza) or completing a particular action (e.g., changing the ticket date). Conversely, the three suggestions that garnered the lowest agreement tended to feature shorter descriptions and shorter durations in the usability video, which might have led some participants to overlook them. Notably, for suggestions with less than 75% agreement, the "after" condition displayed significantly higher

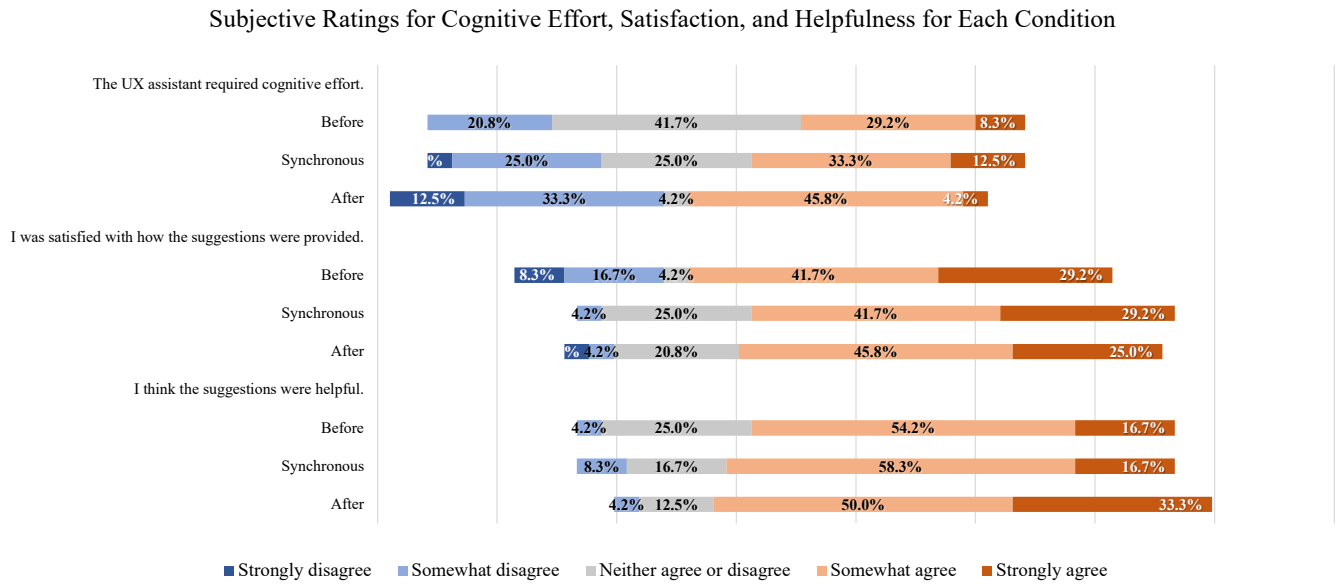

Figure 6: Diverging stacked bar chart that shows the participant ratings for cognitive effort, satisfaction, and helpfulness for each condition.

agreement. This suggests that participants were more inclined to concur with an ambiguous suggestion when it was presented after they had observed the corresponding video segment.

Analysis of Usability Problems Missed by the UX Assistant: Table 4 provides a summary of the total, unique, and average number of problems reported by participants. To identify unique problems, researchers labeled the 342 usability problem descriptions from participants. In total, 24 participants identified 34 unique problems across three videos while ChatGPT identified only 14 from the transcript alone, indicating that it may have missed 20 problems (58.8%). However, its performance was on par with the average number of problems identified by participants. For example, in the website video, each participant identified an average of 4.5 problems, and ChatGPT also found 4 problems. The problems that ChatGPT missed fell into several categories:

- **User interface issues**: These encompassed problems that relied on visual information from the videos, such as issues with low legibility due to small text, ambiguous error messages, missing details on certain pages, and excessive, confusing animations.
- **Interaction-based issues**: These problems pertained to users' tapping and scrolling behaviors on touchscreens and the operation of VR controllers. Such issues could stem from users' unfamiliarity with these input devices, low touchscreen sensitivity, and a lack of visual feedback.
- **Mismatch of user expectations**: This category encompassed issues where participants had to infer user expectations. For instance, P23 inferred that the user expected the group discount to be on the checkout page while that information was on a different page, which led to difficulty completing the task.

Table 4: Number of usability problems identified by participants and by ChatGPT for each video

Video	Total Problems	Unique Problems	Average (SD) per Participant	Problems from ChatGPT
Website	107	12	4.5 (1.0)	4
App	104	15	4.3 (1.7)	5
VR	131	17	5.5 (1.9)	5

- **Navigation issues**: Participants noted problems related to the unclear navigation architecture of the website and VR game. For instance, selecting the logo on the top left corner did not return users to the home page of the website, and choosing a VR game from the menu was confusing.
- **Inefficient designs**: Participants identified inefficiencies that, while not preventing task completion, could enhance the overall user experience. For example, users had to tap the '+' sign ten times to add ten bottles to their cart because there was no option to input the desired quantity directly, which could be frustrating for larger orders.

These categories of problems either relied on information only available in the video recordings or demanded an understanding of the user's mental model and best design practices, aspects that ChatGPT could not infer. Participants also acknowledged these limitations, with P7 noting that it was better than expected but still did not catch all the problems. Interestingly, some participants expressed a competitive spirit with the UX assistant, racing to identify more problems than it did (e.g., *"I was racing against it and found that I identified more problems it was identifying." -P4*)

Furthermore, participants expressed a desire for the UX assistant to provide explanations for identified problems. P13, for instance, stated that *"the suggestions were on the surface level, I already know that the user had difficulty, the point is why—what's the reason for that difficulty."* However, others appreciated the absence of reasons, as it encouraged them to think deeply about the problems instead of relying solely on the suggestions. For example, P17 mentioned *"I liked that I still needed to think about why the user was having a particular issue, we have to remind ourselves not to be carried away and not only rely on the suggestions."* The implications of this feedback and the limitations of using a transcript-based approach for usability analysis are discussed further in Section 6.2.1.

6 DISCUSSION

We conducted a user study to assess the effectiveness of proactive CAs in assisting UX evaluators in identifying usability issues efficiently. Our findings indicate that when the UX assistant provided suggestions after the appearance of potential issues, there was a significant improvement in efficiency, trust, and overall user preference. Additionally, we analyzed the evaluators' responses and their level of agreement with the automatic suggestions. In the following section, we discuss the implications of our findings regarding the timing and functionality of automatic suggestions, and the perception of AI capability.

6.1 Timing and Functionality of Automatic Suggestions (RQ 1.A & RQ 1.B)

Our investigation on the timing of automatic suggestions showed that they played different roles in supporting UX evaluators depending on when they were presented. If the suggestions appeared before or in synchronous with the video segments featuring a potential usability problem, participants perceived them as a **warning** to pay more attention to the upcoming segment. On the other hand, when the suggestions were presented after the video segment, participants viewed them as **validation** of their analysis. These strategies aligned with participants' prior experiences using AI-powered tools for usability analysis, such as tools featuring timelines of suggested problems [23, 24], and a Feature Panel containing icons to indicate potential issues [78]).

The preference for the "after" condition indicates the evaluators' inclination to use their own analytical skills before considering AI suggestions. Participants expressed a desire to initially tackle the task unaided, drawing on their observational skills to identify issues, and subsequently verify whether the AI's analysis corresponded with their conclusions. This approach mirrors the recommended practice of independently evaluating the usability video before engaging in collaborative analysis with another UX evaluator to enhance the reliability of their results [27]. Moreover, this preference underscores participants' engagement with their roles and confidence in their UX knowledge, not allowing the AI to constrain their thinking, which echoes prior findings [78]. Consequently, the development of future human-AI collaborative decision-making tools should prioritize empowering users to apply their knowledge and cultivate an unbiased perspective, reinforcing the value of human expertise in the evaluative process. Given a study showing that AI increased overall work productivity, but in particular for novice workers with minimal impact on experienced workers [11], it would also be interesting to explore whether the level of UX expertise would impact evaluators' preference for certain timing conditions and perceived usefulness of these AI suggestions.

6.1.1 Differences in Timing Preference and Implications for Personalized Timing of Automatic Suggestions. We observed a clear preference among most participants for displaying automatic suggestions after the occurrence of a potential problem, which significantly improved efficiency and trust ratings. However, it is worth noting that the ideal delay duration may not be universally applicable. For example, one participant (P4) felt that the timing of the suggestion coincided with their own conclusions, while another (P10) considered it too late. Individual differences, including analysis speed, prior experience in UX evaluation, and expectations regarding the response time of CAs, likely contributed to this variation. A swift response time is associated with higher system quality [85], suggesting that some participants might prefer to see suggestions immediately after the video segment or with a delay shorter than the 8-second interval used in our study. Prior research on CAs emphasized that a higher level of personalized messages enhances user experience and adoption [1, 19, 42]. While previous work primarily focused on content personalization, there is an intriguing opportunity to customize the timing of suggestions for each UX evaluator. For example, future tools could incorporate a slider that allows UX evaluators to adjust the delay duration, enabling them to determine the most efficient workflow based on their preferences. It would be valuable to explore how evaluators adapt to different delay settings over time to optimize their usability analysis processes.

6.2 Perception of AI Capability (RQ 2.B)

In the "before" condition, we observed an interesting phenomenon where some participants held misconceptions about the capabilities of the UX assistant. They believed that the AI could only detect issues up to the current timestamp and were puzzled by suggestions regarding potential "future" problems. The tool unintentionally gave the impression that the AI was working synchronously with the participants, whereas, in reality, all potential problems were asynchronously detected and revealed when participants reached specific segments of the video. Although this approach led participants to feel like they were working with a colleague in real-time, it had unintended consequences of eroding trust in the UX assistant in the "before" condition and fostering a sense of competition with the AI. Some participants even felt they were in competition with the UX assistant to identify problems faster and more accurately. This finding is consistent with a previous study that integrated machine learning-inferred usability problems into a visual analytics tool [23]. While competition is a common element of gamification, widely used in education contexts to encourage learning and engagement [8, 75], losing a competition can have adverse effects on user engagement and satisfaction [75]. To ensure robust results in usability analysis, it is imperative to foster collaboration rather than competition. AI-assisted systems designed for this purpose should prioritize collaboration and clearly communicate to human evaluators that they should consider the AI's suggestions while maintaining their own judgment and critical insights [3].

An alternative approach for the UX assistant's design could involve informing participants that the AI conducts analysis asynchronously and only flags decisions it is uncertain about for evaluators to review, as suggested by participants in a prior study [24]. However, it is worth noting that relying on AI when it is confident and flagging only uncertain issues reduces evaluator autonomy, as per the 10 Levels of Automation framework [61]. The current design of providing suggestions for all usability problems falls on the lower end of the automation framework, while recording confident problems and flagging uncertain ones would be on the higher end. There is an open question regarding which level of automation human evaluators would find comfortable and what impact different levels would have on analytical performance and trust. Regardless of the level chosen, it is essential to explore ways to facilitate the understanding of AI capabilities effectively.

Besides the level of automation, the level of abstraction of the usability problem suggestions also influenced participants' perceived usefulness. Some found the suggestions to be superficial, merely indicating that the user had difficulty with a task. While participants agreed that the problem occurred, they desired more detailed explanations of the underlying reasons. However, adjusting the level of abstraction might impact agreement with the suggestions. As mentioned in Sec 5.2.2, suggestions with low agreement often had short descriptions, causing confusion among participants. Including more low-level and detailed descriptions could enhance clarity by specifying which video segment the suggestion referred to, potentially increasing agreement. On the flip side, prompting ChatGPT to generate more detailed descriptions, given its known limitations in hallucination [12, 47], raises the risk of false information since it only had access to the transcript. This, in turn, might lower participants' agreement. While beyond the scope of our current focus on suggestion timing, further research is needed to explore finer options for suggestion content and their associated impacts.

6.2.1 Limitations of ChatGPT and Implications for Future AI Tools for UX Analysis. In this study, we simulated a CA that could provide usability problem suggestions by providing raw transcripts to ChatGPT and asking it to identify issues that the user may have encountered from their verbalizations. However, many participants expressed that the list of suggestions was not comprehensive. Our comparison found that ChatGPT missed 58.8% of the unique usability problems collectively identified by 24 participants. This aligns with prior work showing that ChatGPT presented challenges when used for complex tasks like event extraction, achieving only 51.0% of the performance of other task-specific models [28]. Furthermore, when used to generate personas, simulate interviews and usage scenarios, and evaluate user experience, ChatGPT sometimes forgot information, offered partial responses, and lacked output diversity [47]. A recent survey of 1093 researchers revealed that their primary concern with using AI is the potential for inaccurate or incomplete analysis [12]. Another study found that when collaborating with ChatGPT, UX designers reported feeling less actively engaged in the task and displayed lower ownership of the results [17]. These findings highlight the risks associated with using ChatGPT for high-level UX evaluations, which may be exacerbated for less experienced researchers, potentially resulting in misleading conclusions and biased decision-making [12]. Thus, despite its advantages, AI tools should be viewed as complements to human creativity and expertise rather than replacements [2]. Our findings underscore the importance of human expertise in the usability analysis process, which involves observing user actions and verbalizations, understanding context, and synthesizing information to determine the presence of usability problems [13].

While human expertise is indispensable, future AI tools can be enhanced to offer better support to UX evaluators, such as generating more comprehensive lists of usability problems and providing guidelines for usage. For example, instead of relying on text-based information, future AI-powered assistants should strive to incorporate multimodal information from usability test videos and improve contextual understanding. This could involve incorporating capabilities such as emotion detection based on facial expressions (e.g., [14, 40]) or speech analysis (e.g., [37, 38]). These tools could include a disclaimer for its known limitations and provide guidelines for usage, such as stating that it is best suited as an assistant that can speed up some research activities but the information provided by the tool should be validated by human evaluators [7].

6.3 Limitations and Future Work

Our research has effectively employed a proactive CA to augment UX analysis. However, we employed a Wizard of Oz design for handling participant questions, with a human moderator acting as the UX assistant. While this design introduced noticeable delays, it also enabled responses that circumvented the constraints associated with transcript-based analysis. In forthcoming research, it would be valuable to explore alternative methods that can create more realistic scenarios involving CAs.

Throughout the study, participants analyzed three videos, each with a duration of 6 to 8 minutes. Longer videos would naturally generate more automatic suggestions and messages, potentially making chat window scrolling and navigation cumbersome, particularly when clicking on a suggestion to access a specific timestamp. Thus, future studies should prioritize scalable designs capable of accommodating longer usability videos. Additionally, a longitudinal study would provide valuable insights into how UX evaluators' behavior and attitudes change over time. Such a study could yield insights into how their responses to suggestions, types of questions posed, timing preferences, and levels of trust evolve after certain periods of interaction with the UX assistant.

While the majority of participants expressed a preference for the "after" condition, a few conveyed that they did not find it particularly beneficial. Their reasoning centered around having already identified the problem and not perceiving the practice of double-checking analyses as common in their workplace. This observation aligns with prior research that indicates limited adoption of collaborative practices among evaluators due to the associated costs in terms of time, resources, and effort [22, 27, 53]. Therefore, future endeavors should aim to strike a balance between efficiency and robustness. It is worth exploring whether the additional time invested in confirming the presence of a problem, as opposed to reviewing a video once, is justified.

7 CONCLUSION

In this study, we employed a hybrid Wizard of Oz approach to gain insights into the dynamics of UX evaluators' interactions with a proactive CA and the nuanced influence of suggestion timing on their analytic performance and subjective perceptions. We discovered that the timing of suggestions did not impact the number of identified usability problems, but when suggestions appeared after potential problems, participants reported significantly higher levels of trust and efficiency. Participants preferred to rely on their observational and analytical skills while using AI suggestions as a form of validation after the problem occurred in the video. Moreover, our inquiry revealed that 77.6% of ChatGPT-generated suggestions were accepted, but participants noted that these suggestions lacked completeness. In fact, ChatGPT missed 58.8% of the total unique problems identified by the 24 participants, highlighting the irreplaceable role of human reasoning and expertise in usability analysis. Building upon these observations, we proposed design considerations that include providing personalized suggestion timing and harnessing multimodal data from usability videos. In sum, this study contributes insights into the collaborative interplay between human evaluators and AI-driven assistance in usability analysis.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Award No. DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work is partially supported by 1) 2024 Guangzhou Science and Technology Program City-University Joint Funding Project (PI: Mingming Fan); 2) 2023 Guangzhou Science and Technology Program City-University Joint Funding Project (Project No. 2023A03J0001); 3) Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007). We also thank Caluã de Lacerda Pataca and other members of the CAIR lab at Rochester Institute of Technology for reviewing earlier versions of the paper.

REFERENCES

[1] Mubashra Akhtar, Julia Neidhardt, and Hannes Werthner. 2019. The Potential of Chatbots: Analysis of Chatbot Conversations. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, Vol. 01. 397–404. https://doi.org/10.1109/CBI.2019.00052 ISSN: 2378-1971.

[2] Ahmed Al-sa'di and Dave Miller. 2023. Exploring the Impact of Artificial Intelligence language model ChatGPT on the User Experience. *International Journal of Technology, Innovation and Management (IJTIM)* 3, 1 (May 2023), 1–8. https://doi.org/10.54489/ijtim.v3i1.195 Number: 1.

[3] Ali. 2023. How to compete with AI and win in the job market - Beyond. https://beyond.ubc.ca/how-to-compete-with-ai-and-win-in-the-job-market/

[4] Zahra Ashktorab, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2021. Effects of Communication Directionality and AI Agent Differences in Human-AI Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA) *(CHI '21)*. Association for Computing Machinery, 1–15. https://doi.org/10.1145/3411764.3445256

[5] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA) *(CHI '19)*. Association for Computing Machinery, 1–12. https://doi.org/10.1145/3290605.3300484

[6] Zahra Ashktorab, Q. Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020. Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. 4 (2020), 96:1–96:20. Issue CSCW2. https://doi.org/10.1145/3415167

[7] Nick Babich. 2023. Using ChatGPT for User Research. https://uxplanet.org/using-chatgpt-for-user-research-5c3bdf7e26af

[8] Gabriel Barata, Sandra Gama, Joaquim Jorge, and Daniel Gonçalves. 2013. Improving participation and learning with gamification. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications (Gamification '13)*. Association for Computing Machinery, New York, NY, USA, 10–17. https://doi.org/10.1145/2583008.2583010

[9] Timothy Bickmore, Daniel Mauer, Francisco Crespo, and Thomas Brown. 2007. Persuasion, Task Interruption and Health Regimen Adherence. In *Persuasive Technology* (Berlin, Heidelberg) *(Lecture Notes in Computer Science)*, Yvonne de Kort, Wijnand IJsselsteijn, Cees Midden, Berry Eggen, and B. J. Fogg (Eds.). Springer, 1–11. https://doi.org/10.1007/978-3-540-77006-0_1

[10] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction* 12, 2 (June 2005), 293–327. https://doi.org/10.1145/1067860.1067867

[11] Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. 2023. Generative AI at Work *(Working Paper Series)*. National Bureau of Economic Research. https://doi.org/10.3386/w31161

[12] Lizzy Burnam. 2023. We Surveyed 1093 Researchers About How They Use AI—Here's What We Learned. https://www.userinterviews.com/blog/ai-in-ux-research-report

[13] Parmit K. Chilana, Jacob O. Wobbrock, and Andrew J. Ko. 2010. Understanding Usability Practices in Complex Domains. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*. ACM Press, Atlanta, Georgia, USA, 2337–2346. https://doi.org/10.1145/1753326.1753678

[14] M. Kalpana Chowdary, Tu N. Nguyen, and D. Jude Hemanth. 2021. Deep learning-based facial emotion recognition for human–computer interaction applications. (2021). https://doi.org/10.1007/s00521-021-06012-8

[15] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 815–824. https://doi.org/10.1145/2939672.2939746

[16] Microsoft Canada Consumer Insights. 2015. Attention spans. , 52 pages. https://dl.motamem.org/microsoft-attention-spans-research-report.pdf

[17] Hubert Ekvall and Patrik Winnberg. 2023. *Integrating ChatGPT into the UX Design Process : Ideation and Prototyping with LLMs*. https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-211251

[18] Fatma Elbabour, Obead Alhadreti, and Pam Mayhew. 2017. Eye Tracking in Retrospective Think-Aloud Usability Testing: Is There Added Value? - JUX. https://uxpajournal.org/value-eye-tracking-think-aloud-usability-testing/

[19] Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring Language Style in Chatbots to Increase Perceived Product Value and User Engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 301–305. https://doi.org/10.1145/3295750.3298956

[20] Mingming Fan, Yue Li, and Khai N. Truong. 2020. Automatic Detection of Usability Problem Encounters in Think-Aloud Sessions. *ACM Transactions on Interactive Intelligent Systems* 10, 2 (June 2020), 1–24. https://doi.org/10.1145/3385732

[21] Mingming Fan, Jinglan Lin, Christina Chung, and Khai N. Truong. 2019. Concurrent Think-Aloud Verbalizations and Usability Problems. *ACM Transactions on Computer-Human Interaction* 26, 5 (Sept. 2019), 1–35. https://doi.org/10.1145/3325281

[22] Mingming Fan, Serina Shi, and Khai N Truong. 2020. Practices and Challenges of Using Think-Aloud Protocols in Industry: An International Survey. *Journal of Usability Studies* 15, 2 (2020), 85–102.

[23] Mingming Fan, Ke Wu, Jian Zhao, Yue Li, Winter Wei, and Khai N. Truong. 2020. VisTA: Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan. 2020), 343–352. https://doi.org/10.1109/TVCG.2019.2934797

[24] Mingming Fan, Xianyou Yang, TszTung Yu, Q. Vera Liao, and Jian Zhao. 2022. Human-AI Collaboration for UX Evaluation: Effects of Explanation and Synchronization. 6 (2022), 96:1–96:32. Issue CSCW1. https://doi.org/10.1145/3512943

[25] Maryia Fokina. 2023. The Future of Chatbots: 80+ Chatbot Statistics for 2022. https://www.tidio.com/blog/chatbot-statistics/

[26] Asbjørn Følstad, Effie Lai-Chong Law, and Kasper Hornbæk. 2010. Analysis in Usability Evaluations: An Exploratory Study. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries (NordiCHI '10)*. Association for Computing Machinery, New York, NY, USA, 647–650. https://doi.org/10.1145/1868914.1868995

[27] Asbjørn Følstad, Effie Lai-Chong Law, and Kasper Hornbæk. 2012. Analysis in Practical Usability Evaluation: A Survey Study. In *Proceedings of the 30th SIGCHI Conference on Human Factors in Computing Systems - CHI '12*. ACM Press, Austin, Texas, 2127–2136. https://doi.org/10.1145/2207676.2208365

[28] Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the Feasibility of ChatGPT for Event Extraction. https://doi.org/10.48550/arXiv.2303.03836 arXiv:2303.03836 [cs].

[29] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 50:1–50:24. https://doi.org/10.1145/3359152

[30] J. Grigera, Alejandra Garrido, J. Rivero, and G. Rossi. 2017. Automatic Detection of Usability Smells in Web Applications. *Int. J. Hum. Comput. Stud.* (2017). https://doi.org/10.1016/j.ijhcs.2016.09.009

[31] Julián Grigera, Alejandra Garrido, José Matías Rivero, and Gustavo Rossi. 2017. Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies* 97 (2017), 129–148.

[32] Jasmin Grosinger, Federico Pecora, and Alessandro Saffiotti. 2016. Making Robots Proactive through Equilibrium Maintenance. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*.

[33] Ted Grover, Kael Rowan, Jina Suh, Daniel McDuff, and Mary Czerwinski. 2020. Design and evaluation of intelligent agent prototypes for assistance with focus and productivity at work. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (New York, NY, USA) *(IUI '20)*. Association for Computing Machinery, 390–400. https://doi.org/10.1145/3377325.3377507

[34] Patrick Harms. 2019. Automated usability evaluation of virtual reality applications. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 3 (2019), 1–36.

[35] Morten Hertzum and Niels Ebbe Jacobsen. 2001. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction* 15, 1 (2001), 183–204. https://doi.org/10.1207/S15327590IJHC1501_14

[36] Morten Hertzum, Niels Ebbe Jacobsen, and Bonnie E. John. 1998. The Evaluator Effect in Usability Tests. In *CHI 98 Conference Summary on Human Factors in Computing Systems (CHI '98)*. Association for Computing Machinery, New York, NY, USA, 255–256. https://doi.org/10.1145/286498.286737

[37] Maryam Imani and Gholam Ali Montazer. 2019. A survey of emotion recognition methods with emphasis on E-Learning environments. 147 (2019), 102423. https://doi.org/10.1016/j.jnca.2019.102423

[38] Dias Issa, M. Fatih Demirci, and Adnan Yazici. 2020. Speech emotion recognition with deep convolutional neural networks. 59 (2020), 101894. https://doi.org/10.1016/j.bspc.2020.101894

[39] Farnaz Jahanbakhsh, Elnaz Nouri, Robert Sim, Ryen W. White, and Adam Fourney. 2022. Understanding Questions that Arise When Working with Business Documents. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 6. 1–24. Issue CSCW2. https://doi.org/10.1145/3555761

[40] Deepak Kumar Jain, Pourya Shamsolmoali, and Paramjit Sehdev. 2019. Extended deep neural network for facial emotion recognition. 120 (2019), 69–74. https://doi.org/10.1016/j.patrec.2019.01.008

[41] Jeffrey Jenkins, Bonnie Anderson, Anthony Vance, Brock Kirwan, and David Eargle. 2016. More Harm Than Good? How Messages That Interrupt Can Make Us Vulnerable. *Information Systems Research* 27 (Aug. 2016). https://doi.org/10.1287/isre.2016.0644

[42] Liss Jenneboer, Carolina Herrando, and Efthymios Constantinides. 2022. The Impact of Chatbots on Customer Loyalty: A Systematic Literature Review. *Journal of Theoretical and Applied Electronic Commerce Research* 17, 1 (March 2022), 212–229. https://doi.org/10.3390/jtaer17010011 Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[43] JongWook Jeong, NeungHoe Kim, and Hoh Peter In. 2020. Detecting usability problems in mobile applications on the basis of dissimilarity in user behavior. *International Journal of Human-Computer Studies* 139 (2020), 102364.

[44] Indri Tri Julianto, Dede Kurniadi, Yosep Septiana, and Ade Sutedi. 2023. Alternative Text Pre-Processing using Chat GPT Open AI. *Jurnal Nasional Pendidikan Teknik Informatika : JANAPATI* 12, 1 (2023), 67–77. https://doi.org/10.23887/janapati.v12i1.59746 Number: 1.

[45] Hyeji Kim, Inchan Jung, and Youn-kyung Lim. 2022. Understanding the Negative Aspects of User Experience in Human-likeness of Voice-based Conversational Agents. In *Designing Interactive Systems Conference* (New York, NY, USA) *(DIS '22)*. Association for Computing Machinery, 1418–1427. https://doi.org/10.1145/3532106.3533528

[46] Yanghee Kim, Amy L. Baylor, and PALS Group. 2006. Pedagogical Agents as Learning Companions: The Role of Agent Competency and Type of Interaction. 54, 3 (2006), 223–243. https://doi.org/10.1007/s11423-006-8805-z

[47] A. Baki Kocaballi. 2023. Conversational AI-Powered Design: ChatGPT as Designer, User, and Product. (Feb. 2023). https://doi.org/10.48550/arXiv.2302.07406 arXiv:2302.07406 [cs].

[48] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. In *Proceedings of the 2018 Designing Interactive Systems Conference* (New York, NY, USA) *(DIS '18)*. Association for Computing Machinery, 881–894. https://doi.org/10.1145/3196709.3196784

[49] Matthias Kraus, Nicolas Wagner, Zoraida Callejas, and Wolfgang Minker. 2021. The Role of Trust in Proactive Conversational Assistants. 9 (2021), 112821–112836. https://doi.org/10.1109/ACCESS.2021.3103893 Conference Name: IEEE Access.

[50] Veronika Krauß, Alexander Boden, Leif Oppermann, and René Reiners. 2021. Current Practices, Challenges, and Design Implications for Collaborative AR/VR Application Development. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Number 454. Association for Computing Machinery, 1–15. https://doi.org/10.1145/3411764.3445335

[51] Emily Kuang. 2023. Crafting Human-AI Collaborative Analysis for User Experience Evaluation. In *In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3544549.3577042

[52] Emily Kuang, Ehsan Jahangirzadeh Soure, Mingming Fan, Jian Zhao, and Kristen Shinohara. 2023. Collaboration with Conversational AI Assistants for UX Evaluation: Questions and How to Ask them (Voice vs. Text). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3544548.3581247

[53] Emily Kuang, Xiaofu Jin, and Mingming Fan. 2022. "Merging Results Is No Easy Task": An International Survey Study of Collaborative Data Analysis Practices Among UX Practitioners. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3491102.3517647

[54] Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health* 2, 2 (Feb. 2023), e0000198. https://doi.org/10.1371/journal.pdig.0000198 Publisher: Public Library of Science.

[55] Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. arXiv:2112.11471 [cs] http://arxiv.org/abs/2112.11471

[56] Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. 2016. What Can You Do? Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (New York, NY, USA) *(DIS '16)*. Association for Computing Machinery, 264–275. https://doi.org/10.1145/2901790.2901842

[57] Diane J. Litman and Shimei Pan. 2002. Designing and Evaluating an Adaptive Spoken Dialogue System. *User Modeling and User-Adapted Interaction* 12, 2 (2002), 111–137. https://doi.org/10.1023/A:1015036910358

[58] Bingjie Liu. 2021. In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human–AI Interaction. (2021). Issue zmab013. https://doi.org/10.1093/jcmc/zmab013

[59] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA) *(CHI '16)*. Association for Computing Machinery, 5286–5297. https://doi.org/10.1145/2858036.2858288

[60] Michal Luria, Guy Hoffman, and Oren Zuckerman. 2017. Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA) *(CHI '17)*. Association for Computing Machinery, 580–628. https://doi.org/10.1145/3025453.3025786

[61] Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. 2019. Discovering the Sweet Spot of Human-Computer Configurations: A Case Study in Information Extraction. 3 (2019), 195:1–195:30. Issue CSCW. https://doi.org/10.1145/3359297

[62] Michael McTear. 2020. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst).

[63] Indrani Medhi Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'Neill. 2017. How Do You Want Your Chatbot? An Exploratory Wizard-of-Oz Study with Young, Urban Indians. In *Human-Computer Interaction - INTERACT 2017* (Cham) *(Lecture Notes in Computer Science)*, Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler (Eds.). Springer International Publishing, 441–459. https://doi.org/10.1007/978-3-319-67744-6_28

[64] Matt Moran. 2022. *25+ Top Chatbot Statistics: Usage, Demographics, Trends*. https://startupbonsai.com/chatbot-statistics/

[65] Noboru Nakamichi, Kazuyuki Shima, Makoto Sakai, and Ken-ichi Matsumoto. 2006. Detecting low usability web pages using quantitative data of users' behavior. In *Proceedings of the 28th international conference on Software engineering (ICSE '06)*. Association for Computing Machinery, New York, NY, USA, 569–576. https://doi.org/10.1145/1134285.1134365

[66] Jakob Nielsen. 2023. UX Needs a Sense of Urgency About AI. https://www.uxtigers.com/post/ux-urgency-ai

[67] Mie Nørgaard and Kasper Hornbæk. 2006. What Do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing. In *Proceedings of the 6th Conference on Designing Interactive Systems (DIS '06)*. Association for Computing Machinery, New York, NY, USA, 209–218. https://doi.org/10.1145/1142405.1142439

[68] Florian Nothdurft, Stefan Ultes, and Wolfgang Minker. 2015. Finding Appropriate Interaction Strategies for Proactive Dialogue Systems—An Open Quest. In *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication*, Vol. 110. Tartu, Estonia, 73–80.

[69] Karina Oertel and Oliver Hein. 2003. Identification of Web Usability Problems and Interaction Patterns with the RealEYES-iAnalyzer. In *Interactive Systems. Design, Specification, and Verification (Lecture Notes in Computer Science)*, Joaquim A. Jorge, Nuno Jardim Nunes, and João Falcão e Cunha (Eds.). Springer, Berlin, Heidelberg, 77–91. https://doi.org/10.1007/978-3-540-39929-2_6

[70] OpenAI. 2023. ChatGPT. https://chat.openai.com

[71] OpenAI. 2023. ChatGPT — Release Notes. https://help.openai.com/en/articles/6825453-chatgpt-release-notes

[72] Asil Oztekin, Dursun Delen, Ali Turkyilmaz, and Selim Zaim. 2013. A Machine Learning-Based Usability Evaluation Method for eLearning Systems. *Decision Support Systems* 56 (Dec. 2013), 63–73. https://doi.org/10.1016/j.dss.2013.05.003

[73] Fabio Paternò, Antonio Giovanni Schiavone, and Antonio Conti. 2017. Customizable automatic detection of bad usability smells in mobile accessed web applications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–11.

[74] Lazlo Ring, Barbara Barry, Kathleen Totzke, and Timothy Bickmore. 2013. Addressing Loneliness and Isolation in Older Adults: Proactive Affective Agents Provide Better Support. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 61–66. https://doi.org/10.1109/ACII.2013.17 ISSN: 2156-8111.

[75] Sepandar Sepehr and Milena Head. 2013. Competition as an element of gamification for learning: an exploratory longitudinal investigation. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications (Gamification '13)*. Association for Computing Machinery, New York, NY, USA, 2–9. https://doi.org/10.1145/2583008.2583009

[76] I. Shah. 2008. Event Patterns as Indicators of Usability Problems. *Journal of King Saud University - Computer and Information Sciences* 20 (2008), 31–43. https://doi.org/10.1016/S1319-1578(08)80003-1

[77] Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel K. E. Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the Effects of Embodiment for a Group Facilitation Agent. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA) *(CHI '18)*. Association for Computing Machinery, 1–13. https://doi.org/10.1145/3173574.3173965

[78] Ehsan Jahangirzadeh Soure, Emily Kuang, Mingming Fan, and Jian Zhao. 2021. CoUX: Collaborative Visual Analysis of Think-Aloud Usability Test Videos for Digital Interfaces. *IEEE Transactions on Visualization and Computer Graphics* (2021), 1–11. https://doi.org/10.1109/TVCG.2021.3114822

[79] Babak Taati, Jasper Snoek, and Alex Mihailidis. 2013. Video analysis for identifying human operation difficulties and faucet usability assessment. *Neurocomputing* 100 (Jan. 2013), 163–169. https://doi.org/10.1016/j.neucom.2011.10.041

[80] Wilbert Tabone and Joost de Winter. 2023. *Using ChatGPT for Human-Computer Interaction Research: A Primer.*

[81] UserLike. 2022. What Do Your Customers Actually Think About Chatbots? [Research Study]. https://www.userlike.com/en/blog/consumer-chatbot-perceptions

[82] Uxcel. 2023. AI in UI/UX Design Course. https://app.uxcel.com/courses/ai-in-ux-ui-design

[83] Eva A. M. van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L. Bockting. 2023. ChatGPT: five priorities for research. *Nature* 614, 7947 (Feb. 2023), 224–226. https://doi.org/10.1038/d41586-023-00288-7

[84] Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study. https://arxiv.org/abs/2304.04339v1

[85] Jeong-Bin Whang, Ji Hee Song, Jong-Ho Lee, and Boreum Choi. 2022. Interacting with Chatbots: Message type and consumers' control. *Journal of Business Research* 153 (Dec. 2022), 309–318. https://doi.org/10.1016/j.jbusres.2022.08.012

[86] Jun Xiao, John Stasko, and Richard Catrambone. 2004. An Empirical Study of the Effect of Agent Competence on User Performance and Perception. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1* (USA) *(AAMAS '04)*. IEEE Computer Society, 178–185.

[87] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[88] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[89] Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 2022. UX Research on Conversational Human-AI Interaction: A Literature Review of the ACM Digital Library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–24. https://doi.org/10.1145/3491102.3501855

[90] John Zimmerman, Changhoon Oh, Nur Yildirim, Alex Kass, Teresa Tung, and Jodi Forlizzi. 2020. UX Designers Pushing AI in the Enterprise: A Case for Adaptive UIs. *Interactions* 28, 1 (Dec. 2020), 72–77. https://doi.org/10.1145/3436954

A APPENDIX

Table 5: List of usability problem suggestions generated by ChatGPT and its corresponding percentage agreement from participants

Product	Start	End	Problem Suggestions	% Agreement
Museum Website	0:44	1:25	The user had difficulty finding the Butterfly Garden admission section.	100%
	2:14	2:52	The user had difficulty finding the group admission costs for the Museum and Butterfly Garden.	87.5%
	4:26	5:27	The user had difficulty adding an adult ticket for Museum and Butterfly Garden admission to the cart.	100%
	5:42	6:29	The user had difficulty changing the date of the ticket.	87.5%
Food Delivery App	0:55	1:20	The user had difficulty finding the "drinks" category to select classic Coke and Sprite.	91.7%
	2:26	3:01	The user had difficulty finding the option to order full-sheet pizzas.	75%
	4:44	4:58	The user was uncertain about how to mark that they were done with the pizza customization.	41.7%
	6:11	6:42	The user had difficulty finding how to switch from carryout to delivery and update the delivery address.	66.7%
	7:18	7:42	The user had difficulty entering the delivery address.	37.5%
VR Game	0:01	0:30	The user was lacking clarity on how to select a game from the menu.	75%
	0:36	1:11	The user had difficulty in understanding the objective of the game due to lack of instructions.	87.5%
	1:13	1:58	The user had difficulty in controlling the game.	33.3%
	4:11	5:32	The user had difficulty in navigating the menu.	87.5%
	5:55	7:00	The user had difficulty in understanding game controls.	100%

Table 6: Calculations of precision and recall of ChatGPT's problem suggestions compared to the ground truth (Note: T_p = True positives, F_p = False positives, F_n = False negatives)

Metric	Calculations
Precision	$\frac{T_p}{T_p+F_p} = \frac{12}{12+2} = 0.857$
Recall	$\frac{T_p}{T_p+F_n} = \frac{12}{12+5} = 0.706$