"It Became My Buddy, But I'm Not Afraid to Disagree": A Multi-Session Study of UX Evaluators Collaborating with Conversational AI Assistants

Emily Kuang
Electrical Engineering and Computer
Science
York University
Toronto, Ontario, Canada
ekuang@yorku.ca

Ehsan Jahangirzadeh Soure
School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
ejahangi@uwaterloo.ca

Luyao Shen
Computational Media and Arts Thrust
The Hong Kong University of Science
and Technology (Guangzhou)
Guangzhou, China
lshen595@connect.hkust-gz.edu.cn

Nitesh Goyal
Google DeepMind
New York, New York, USA
teshgoyal@acm.org

Mingming Fan
The Hong Kong University of Science
and Technology (Guangzhou)
Guangzhou, China
mingmingfan@ust.hk

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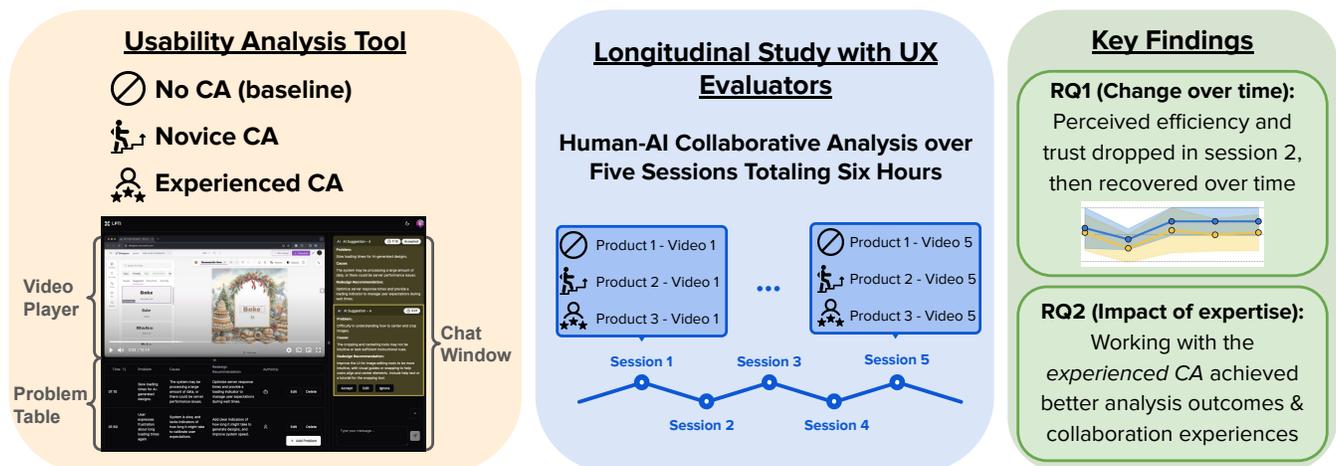Rochester, New York, USA
kristen.shinohara@rit.edu

Figure 1: Visual abstract summarizing the study. We developed a Usability Analysis Tool with three conditions: (1) No Conversational Assistant (CA), (2) a Novice CA, and (3) an Experienced CA. In a within-subjects study spanning five sessions, twelve UX evaluators analyzed 15 usability videos across all three conditions. We investigated how their analysis behaviors and perceptions of the CAs changed over time (RQ1), and how the perceived level of CA expertise influenced these outcomes (RQ2).

Abstract

AI-assisted usability analysis can potentially reduce the time and effort of finding usability problems, yet little is known about how AI's perceived expertise influences evaluators' analytic strategies and perceptions over time. We ran a within-subjects, five-session study (six hours per participant) with 12 professional UX evaluators who worked with two conversational assistants designed to appear novice- or expert-like (differing in suggestion quantity and response accuracy). We logged behavioral measures (number of passes, suggestion acceptance rate), collected subjective ratings (trust, perceived efficiency), and conducted semi-structured interviews. Participants experienced an initial novelty effect and a subsequent dip in trust that recovered over time. Their efficiency improved as they shifted from a two-pass to a one-pass video inspection approach. Evaluators ultimately rated the experienced CA as significantly more efficient, trustworthy, and comprehensive, despite not perceiving expertise differences early on. We conclude with design implications for adapting AI expertise to enable calibrated human-AI collaboration.

CCS Concepts

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

Keywords

User experience; Usability analysis; Human-AI collaboration; Proactive conversational assistants; Longitudinal study; Novelty effect

1 Introduction

Technology is now embedded in nearly every aspect of daily life. As digital interfaces grow more complex, delivering a meaningful and seamless user experience (UX) has become increasingly challenging [83]. Without good usability, even powerful tools risk going unused. Usability reflects how easily people can use a system, assessed across learnability, efficiency, memorability, error rates, and satisfaction [79]. Usability testing is central to identifying obstacles to users' goals and generating redesign recommendations. It typically involves observing users performing tasks, with sessions often video-recorded for later review [24, 81, 129]. These videos help evaluators identify problems by examining users' actions and context, yet reviewing them remains demanding and time-consuming [13, 82]. To improve the completeness and reliability of usability evaluations, researchers and practitioners have applied artificial intelligence (AI) and machine learning (ML) techniques to analyze usability recordings [23, 36, 43, 67, 88, 90, 136, 140]. Commercial platforms now offer AI-powered features such as emotion detection and sentiment analysis [124, 125], since frustration and negative sentiment often indicate usability problems [23]. Academic work similarly analyzes anomalies in interaction logs (e.g., clicks, navigation paths, time on page) to detect issues [36, 43, 88, 90, 136, 140]. Despite these advances, fully automated approaches remain incomplete, often missing complex or subtle usability problems that require contextual understanding of user expectations and emotional responses [23, 33, 64]. This underscores the continued need for human expertise to interpret user behaviors and feedback.

The limitations of fully automated approaches for detecting usability problems have led to increasing interest in *human-AI collaborative analysis*, where AI augments rather than replaces human judgment [58]. Usability pioneer Jakob Nielsen describes AI as a "free colleague," suggesting that the combination of human expertise and AI's scalable processing can exceed the capabilities of either alone [80]. Prior work has begun to map the landscape of human-AI collaborative usability analysis [4, 25, 26, 55, 57], examining dimensions such as (1) *visualizations* of usability problem indicators [4, 25, 115], (2) *interaction modality* with conversational assistants (CAs) [55], (3) *explanations* for problem suggestions [26], and (4) the *timing* of those suggestions [57]. While these studies provide promising evidence for AI-supported evaluations, they leave open two foundational dimensions that shape real-world collaboration. The first is the **time** factor: how collaboration unfolds across repeated sessions as novelty diminishes, familiarity increases, and evaluators develop stable practices. Most existing

work employs short, single-session studies (35-100 minutes), offering limited insight into how evaluators adapt to AI tools over time. Such brief interactions make it difficult to distinguish genuine effects from novelty, as user perceptions and behaviors often shift with continued use [94, 116]. The second is **perceived expertise**, operationalized here as the precision and recall of AI-generated suggestions. Prior studies typically held the underlying accuracy of AI suggestions constant, whether through Wizard-of-Oz control [26, 55], by using a general-purpose LLM such as ChatGPT without tailored prompt engineering [57], or by employing a custom GPT trained on Nielsen's heuristics [120]. As a result, we know little about how evaluators respond when the AI's expertise varies, even though real-world systems inevitably differ in quality. Yet effective human-AI collaborative usability analysis depends on a **joint development of trust and expertise over time**. One can occur without the other, but it is unclear whether meaningful collaboration is possible without both. Because prior work has largely bypassed these dimensions, we still do not know whether human-AI collaborative UX evaluation remains effective when collaboration unfolds over time and when the AI's expertise varies.

To explore the **time** factor, we examine how collaboration between UX evaluators and CAs unfolds across multiple sessions. A longitudinal design allows us to capture how user behavior, trust, and reliance evolve as initial novelty diminishes and familiarity grows, thereby improving ecological validity by accounting for both novelty and learning effects [103]. The understudied evolution of human-AI collaborative analysis motivates our first research question (RQ):

- **RQ1**: How do UX evaluators' analytic behaviors and attitudes toward a CA change over time?

Meanwhile, advances in prompt engineering show that LLMs can adopt tailored personas that significantly change the accuracy of outputs and shift user interaction dynamics [53, 110]. This raises an important design question for usability analysis, a domain grounded in specialized knowledge of user behavior, heuristics, and design principles [77]: how does varying the **perceived expertise** of the AI system impact evaluators' behavior and attitudes? We foreground the *novice-expert framework* as the most theoretically and practically relevant lens for this inquiry. Extensive research shows that novices and experts differ systematically in domain knowledge [2, 126], skills [42, 99, 105], and analytic strategies [30, 45, 56, 102]. Because these differences directly shape how evaluators identify and interpret usability problems, expertise provides a meaningful and ecologically grounded design dimension compared to communication style (which is often superficial) or underlying model capability (which is commonly opaque and not tunable in commercial systems). This focus on expertise is also motivated by *real-world practice*. Commercial AI tools for usability analysis vary widely in their apparent expertise due to differences in underlying models (e.g., UserTesting combines proprietary and open-source models [123]) and the data they can access (e.g., Maze draws on accumulated transcripts [118]). As UX teams increasingly mix tools within their workflows, they will encounter AI systems with inconsistent levels of expertise, making variable AI expertise a practically relevant and emerging design consideration.

Building on this foundation, we investigate how prompting a CA to simulate either a novice or expert UX evaluator shapes human-AI collaboration in usability analysis. Prior work suggests that expert-like assistants and high-accuracy AI systems can enhance trust and insight generation [93, 137], making the design of an experienced CA that provides accurate and comprehensive usability problem descriptions a natural goal. However, over-reliance on AI is a well-documented concern in AI-assisted decision-making, as it can reduce critical engagement and impair human performance [9, 127]. Novice-like assistants, by occasionally making mistakes, may counteract this risk by prompting evaluators to reflect more critically on AI suggestions. This idea aligns conceptually with the *protégé effect* in education, where teaching others promotes deeper learning [35], and with the *rubber duck effect* in programming, where explaining code aloud to an object with no knowledge fosters critical reflection and better debugging outcomes [50]. Novices may also bring a fresher perspective, as they are less constrained by established knowledge frames that can anchor experts [30]. Since both novice and expert CAs may provide distinct benefits, and little is known about how perceived AI expertise influences evaluators over time, this motivates our second RQ:

- **RQ2**: How does the perceived expertise of a CA (simulated as either a novice or experienced UX evaluator) influence UX evaluators' analytic behaviors and attitudes over time?

To address these RQs, we developed a usability analysis tool that incorporated two different CAs. Advances in LLMs enable the creation of custom CAs tailored to specific tasks [84], and prior work has applied them to design contexts, including DesignGPT and Idea Generation GPT [21, 130]. Building on this capability, we designed two custom GPTs to simulate novice and experienced UX evaluators, using the novice-expert paradigm as a prompt-generating framework to elicit distinct yet valuable forms of feedback across the two conditions. The CAs were integrated into a video analysis tool that supported common usability evaluation tasks: identifying problems, diagnosing causes, suggesting redesigns, and answering user queries. To isolate the contribution of CA support, we implemented three experimental conditions: a baseline with no CA, a novice CA, and an experienced CA. We conducted a within-subjects study with twelve UX evaluators, each completing five analysis sessions over three weeks (approximately six hours of interaction with the tool). We analyzed their strategies by examining video playback behaviors (e.g., pauses, number of passes) and interactions with the CAs (e.g., accepting or ignoring problem suggestions). Analytic performance was measured by the total number of usability problems identified, the number of unique problems found, and inter-rater reliability. Finally, we assessed participants' attitudes towards the CAs through Likert scale ratings of perceived efficiency, completeness, trust, confidence, and UX expertise.

For **RQ1**, we identified two categories of analysis strategies: "two-pass" (e.g., reviewing a video twice) and "one-pass" (e.g., reviewing once). Two-pass strategies were more common during early sessions, while one-pass strategies became more prevalent over time. Participants exhibited a clear *novelty effect*, reporting higher perceived efficiency and trust in the first session, followed by a dip in the second, and a recovery in the third session, which sustained until the end. For **RQ2**, participants consistently preferred the experienced CA, rating it significantly higher in efficiency, trustworthiness, and the completeness of its suggestions. In contrast, the novice CA was seen as more useful for training, prompting participants to critically engage with its suggestions rather than accepting them at face value. Importantly, participants were unable to distinguish between the perceived expertise levels of the two CAs until they had more hands-on experience in the second session. Our findings across both RQs highlight the value of longitudinal studies in human-AI collaboration and point to different use cases for novice and experienced CAs. Furthermore, we provide implications for future longitudinal studies on human-AI collaboration. In summary, our contributions include:

- Development of an LLM-based analysis tool with novice and experienced CAs designed to identify usability problems and respond to user questions;
- Findings from a multi-session study illustrating how UX evaluators' analytic behaviors and perceptions during human-AI collaborative analysis progress over time;
- Comparative analysis of novice and experienced CAs, examining their differing impacts on usability analysis.

2 Related Work

Our work draws inspiration from related research in two key areas: 1) longitudinal studies on human-AI collaboration, and 2) collaboration with CAs featuring varying levels of expertise.

2.1 Longitudinal Studies on Human-AI Collaboration

Longitudinal studies involve gathering data over multiple time points, incorporating time as a dependent variable into the research design [31]. In HCI, these studies deepen our understanding of how human-technology interactions evolve over time [52].

2.1.1 Benefits of Longitudinal Studies. Longitudinal studies address the *novelty effect*, where users' initial responses to technology differ from long-term usage patterns [94]. Previous research shows that initial fascination with new technologies often diminishes, leading to discontinued use once the novelty effect subsides [46, 116, 117]. At the same time, extended use can also foster positive outcomes [75, 107, 128]. For example, over four months, older adults' perceptions of a smart speaker evolved from appreciating its simplicity to valuing it as a digital companion, showing a shift from superficial interaction to meaningful engagement [51]. Their initial challenges with the device diminished as they adapted their troubleshooting behaviors, like repeating or paraphrasing commands, instead of blaming the device [51]. These findings highlight how user behaviors and perceptions continue to evolve after the novelty effect fades, with adoption trajectories varying widely across contexts. In our study, we therefore focused on the usability context and how evaluators adapted their strategies over time as they developed a better understanding of the CAs' capabilities.

2.1.2 Research Gap in Longitudinal Studies on Human-AI Collaboration. Recent advances in AI have spurred the development of human-AI collaboration tools [68]. However, a review of 37 empirical studies on human-AI collaboration from 2018 to 2023 reveals a

heavy reliance on short-term, quantitative experiments, with relatively few studies employing longitudinal or qualitative methods [92]. This lack of methodological diversity may be due to the novelty of AI technologies and the practical challenges of conducting long-term research, such as participant retention and time investment [31, 47]. Yet relying solely on short-term experiments can produce misleading conclusions. These studies typically capture participants' initial reactions to AI tools, which may be shaped by curiosity and unfamiliarity, rather than sustained, stable patterns of use [116]. For instance, early interactions with CAs often evoke feelings of social connection or anthropomorphism [39], but it remains unclear whether these impressions endure with repeated exposure [3]. Some longitudinal studies of social or health-focused CAs have shown diminishing engagement over time [17], while others have found increasing trust with continued use [3]. However, these findings have not been extended to task-oriented domains like usability evaluation, where sustained performance and domain expertise are crucial.

To address this gap, our study takes a longitudinal approach to examine how evaluators' interactions with a CA evolve over time. Unlike short, one-off evaluations, multi-session studies enable us to observe shifts in trust, reliance, and analytic strategies that emerge through repeated use. For example, trust in AI is known to accumulate gradually, influenced by past experiences and evolving expectations [44]. By studying UX evaluators across five sessions over three weeks, we capture both their immediate impressions and how they adapt their evaluative practices, respond to AI suggestions, and form sustained working relationships with CAs. This approach provides an ecologically valid and comprehensive view of how human-AI collaboration unfolds in practice.

2.2 Collaborating with Novice and Experienced CAs

2.2.1 Existing Human-AI Collaborative Tools for Usability Analysis.
In exploring the role of AI as an *assistant* to UX evaluators, prior research has employed two approaches: representing AI as non-interactive visualizations [4, 25, 26, 115], and developing interactive conversational assistants (CAs) [54, 55, 57]. Visualizations, such as icons and line charts, effectively highlight usability problems like abnormal pitch and negative sentiment [23]. However, these static representations limit UX evaluators' ability to engage with the data and ask follow-up questions for AI-generated results. CAs address some of these limitations by enhancing interactivity; they offer usability problem suggestions and provide explanations on demand [55, 57]. Research has shown that UX evaluators prefer text-based interactions over voice when analyzing usability videos [55], and that proactive CAsthat automatically suggest usability problems at specific timestamps can further support evaluators [57].

However, while these advancements demonstrate the benefits of both visualization and interactivity, there remains an unaddressed gap: previous systems have not accounted for *variations in AI's expertise* [26, 57, 106, 115]. This omission raises concerns about the generalizability of the findings, as participants' reactions are likely influenced by the AI's perceived UX expertise. For instance, a recent study demonstrated that both trust and reliance on AI

advice were significantly higher when the model had high accuracy compared to low accuracy [44]. Prior studies on AI-assisted usability analysis have largely relied on generic versions of ChatGPT without prompt customization. For example, some researchers pasted usability test transcripts directly into ChatGPT and asked it to identify usability problems [57], while others used GPT-4 to automate heuristic evaluations and detect UI violations without defining a specific role or expertise for the model [22]. Although these approaches demonstrated the potential of conversational AI to support analysis, they also revealed the limitations of using general-purpose models without UX-specific expertise. In contrast, a recent short paper systematically compared five prompting techniques: zero-shot prompting, role prompting, chain-of-thought prompting, self-refine prompting, and least-to-most prompting, for identifying usability problems on a shopping website. The results showed that role prompting was particularly effective in uncovering problems, providing contextual analyses, and generating actionable insights from a user perspective [110]. Building on these insights, our work investigates how perceived AI expertise (operationalized through role prompting) influences users' engagement with and utilization of the system over time.

2.2.2 Motivation for Simulating CAs as Novice and Experienced UX Evaluators.
The *novice-expert paradigm* explains that novices learn and think differently from experts [62], progressing through five stages of proficiency: novice, advanced beginner, competent, proficient, and expert [5]. Differences between novices and experts have been observed in various fields, including product and UI design [12, 15], pair programming [69], and aerospace engineering [18]. In the context of UX, expertise is defined as the ability to apply knowledge effectively in UX work, with previous research categorizing UX professionals as either "early career" or "later career" [105]. Given the varying terminologies used to describe levels of expertise [27], we adopt the general terms "**novice**" and "**experienced**" in this paper.

A key motivation for simulating CAs as novice and experienced UX evaluators is the significant differences in how UX evaluators approach usability analysis based on their level of expertise [56]. Experienced UX evaluators possess better-organized knowledge and more effective strategies for accessing and applying that knowledge, enabling them to select and implement the most appropriate research methods [93, 99]. In contrast, novice evaluators often have a narrower range of techniques and a more limited understanding [2]. However, they may also bring up-to-date training that is valuable for navigating ongoing digital transformation [42]. Furthermore, novices are less anchored to established "frames" (i.e., cognitive structures used to interpret events or data), allowing them to offer a fresher perspective [30]. This aligns with the "curse of knowledge," where experts may assume others have the same knowledge, introducing bias and constraining their viewpoints [121, 135]. When analyzing user feedback, experts are more likely than novices to question inconsistencies and seek critical information to deepen their understanding of design goals [30]. Additionally, experienced UX evaluators often customize usability problem descriptions and collaborate to ensure reliability [28, 56], whereas novice researchers may struggle with data analysis, such as identifying meaningful patterns from codes [45] and analyzing complex problems [102].

While designing an experienced CA that provides accurate and complete usability problem descriptions is a clear goal, over-reliance on AI remains a well-documented concern in AI-assisted decision-making, as it can impair human performance and reduce critical engagement [9, 127]. Various strategies have been proposed to mitigate over-reliance and increase human autonomy, such as fore-grounding choices and prompting reflection [14] or introducing cognitive forcing functions to encourage independent judgment [9]. In our study, we explore an alternative approach: simulating a novice CA that identifies fewer usability problems compared to the experienced CA. This design encourages users to stay alert, critically evaluate the AI's suggestions, and take a more active role in the analysis process [35]. Rather than intentionally weakening the AI, the novice CA is framed as a collaborative peer-in-training. This framing draws inspiration from the *protégé effect*, a phenomenon in which individuals tend to think more deeply and critically when they are responsible for teaching, supporting, or correcting someone else's understanding [11]. While the protégé effect has been primarily studied in educational settings, its core mechanism—heightened cognitive engagement driven by a sense of responsibility—may extend to collaborative work, including human-AI interaction. For instance, studies have found that when individuals felt accountable for teaching others, even fictional characters, they were more likely to engage more critically with the material and acknowledge errors [11, 16]. In our context, a novice CA may elicit a mentoring mindset in UX evaluators, leading them to examine the AI's suggestions more critically, reflect more deeply on the analysis, and develop a more nuanced understanding of usability problems. We investigate this as a possible mechanism through which novice CAs could offer distinct benefits.

3 Design of the Conversational Assistants

To answer our RQs on the long-term use of CAs in usability analysis and the impact of perceived UX expertise, we first needed to simulate a novice CA and an experienced CA. Building upon prior work on CAs for usability analysis, we designed two types of support:

(1) **Automatic suggestions** that appear *after* a usability problem occurred in the video. Findings in prior work indicated that presenting suggestions after potential usability problems significantly enhanced perceived efficiency, trust, and user preference [57]. Recognizing the participants' feedback for the CA to not only describe the problem but also identify its root cause, we have included information on the *cause of the problem*. Additionally, each identified problem is accompanied by a *redesign recommendation*, aligning with the overarching goal of usability testing to unearth opportunities for design improvement [74]. This inclusion of causes and redesign recommendations distinguishes our CAs from previous work [55, 57].

(2) **Reactive responses** to UX evaluators' questions. Building on prior research showing that UX evaluators seek to ask questions during analysis, and using a dataset of common queries [55], the CA is designed to provide answers to various questions posed by evaluators about usability videos.

The following subsections describe the process of simulating the CAs, generating automatic suggestions and reactive responses, and evaluating our approach.

3.1 Simulating Novice and Experienced CAs

We adopted the novice-expert paradigm as a prompt-engineering framework, informed by our literature review on differences in UX knowledge and analysis practices between novice and experienced evaluators (Section 2.2). Building on recent work showing that LLMs can simulate diverse user characteristics [53, 60, 109, 134] and be prompted for specific design tasks (e.g., DesignGPT, Idea Generation GPT, Keyword Extraction GPT, Nielsen-heuristics-trained GPT) [21, 120, 130], we created two custom GPTs to simulate novice and experienced UX evaluators, both powered by GPT-4V (see Table 1). Task instructions were defined using OpenAI's GPT Builder interface on the ChatGPT web platform, following the "Creating a GPT" guidelines [85] (see Appendix, Table 5). Prior studies have shown that LLMs can approximate domain experts in contexts such as research ideation and games, producing overlapping input across expertise groups [63, 122]. In contrast, our aim was not to test how faithfully GPT could replicate novice or experienced UX evaluators. Rather, we sought to generate two CAs with clearly distinguishable capabilities in usability analysis, ensuring a meaningful contrast in *perceived expertise*. To contextualize the effects of this tailoring, we also included the general version of GPT-4V[1] (without prompt customization) as a comparison point for evaluating the added value of expertise design.

Table 1: Prompts Used to Simulate Novice and Experienced CAs

Condition	Prompt [with references to relevant sources]
Novice	You are a novice UX evaluator with limited hands-on experience in conducting user research and a basic understanding of UX research methodologies and concepts [2]. Although you are trained in the latest tools [42], you may find it challenging to analyze data, such as identifying meaningful patterns from codes [45] and tackling complex problems [102].
Experienced	You are an experienced UX evaluator with extensive hands-on experience in conducting user research and a deep understanding of UX research methodologies and concepts [99]. You frequently customize usability problem descriptions and prioritize ensuring reliability in your analyses [28, 56]. Additionally, you actively question inconsistencies and seek critical information to deepen your understanding of design goals [30].

3.2 Generating Automatic Usability Problem Suggestions

We compiled a dataset comprising 15 usability videos covering three products, each having five users completing the same tasks. Details of the study videos are in Section 5.2. Initially, transcripts for these videos were automatically generated by Zoom, as the recordings

---

[1]GPT-4 with vision was the most advanced model available at the time of the study.

were made using this software. Subsequently, we thoroughly reviewed the transcripts to correct any errors, add punctuation, and adjust timestamps to reflect natural speech breaks.

Following recent work that demonstrated GPT's ability to understand UI page information [65], we first uploaded screenshots of the main pages of the product directly to the GPT webpage[2] so that the custom GPTs could develop an understanding of the UI elements available for users to interact with. For example, we uploaded a screenshot of the home page of a test website (Fig. 12-A in the Appendix). The "experienced GPT" provided a response that described the content of the screenshot, including a title bar, input field, example posters, poster thumbnails, footer link, preview pane, user profile icon, and closing option (Fig. 12-B in the Appendix). The detailed description shows that the GPTs could extract relevant information from the provided screenshots. We then provided the GPTs with the tasks completed by each user in the video, along with the transcript, and prompted them to generate usability problem descriptions, causes, and redesign recommendations in a prescribed format. Table 5 in the Appendix shows the specific prompts, which followed OpenAI's official prompt engineering guide (e.g., including relevant details in the query, asking the model to adopt a persona, and using delimiters to clearly separate distinct parts of the input) [87]. Prior work suggests that generating responses three times can improve the stability of ChatGPT's output performance [32, 98]. Thus, we prompted the GPTs three times for each video and used the union of all problems generated across these three rounds as the final list. To generate usability problem suggestions for the general version, we used the same process as for the custom GPTs, but directly through the ChatGPT interface without any prompt customization.

3.3 Generating Responses using the OpenAI API

To meet user expectations and in line with the behavior of common voice assistants like Siri and Google Assistant [68], the novice and experienced CAs were designed to respond to impromptu questions posed by UX evaluators. In contrast to prior work that utilized the Wizard of Oz method [55, 57], we leveraged the OpenAI API (using the "gpt-4-0613" model) to provide real-time responses during each study session [86]. To maintain consistency between real-time responses and pre-generated automatic suggestions, the same prompts used for simulating the respective CA were provided at the beginning of each session (Table 1). We also supplied the video background, transcript, and the pre-generated list of automatic suggestions so the CA understands the video content (Table 5). For the general version, we directly provided the video transcript and suggestions list without prompting to adopt a persona.

3.4 Evaluation of Suggestions and Responses

To assess differences in usability analysis between the two custom CAs and the general version, three authors with UX expertise independently analyzed the study videos. They then held a group discussion to consolidate their findings and resolve any disagreements, following established UX practices [38]. The finalized list of usability problems from this manual analysis served as the *ground truth* for evaluating the performance of both the custom GPT models and

the OpenAI API responses. This method of comparing AI-generated results to manual analysis is consistent with prior research on AI-powered usability analysis tools [57]. For automatic suggestions, we manually evaluated whether each suggestion matched a usability problem in the ground truth. Based on this assessment, we calculated precision (the proportion of correct problems among all identified problems) and recall (the proportion of correct problems among all correct problems) for each video. For responses, we asked one example question from each of the five main categories previously used by UX evaluators when interacting with CAs [55]. We then rewatched the corresponding video segments to determine whether the response was correct and calculated the accuracy as the proportion of correct responses per video. The mean and standard deviation across 15 videos are presented in Table 2.

Table 2: Suggestion precision and recall, and response accuracy for the three conditions, reported in "Mean (Standard Deviation)"

Condition	Suggestion Precision	Suggestion Recall	Response Accuracy
Novice	0.80 (0.13)	0.54 (0.13)	0.73 (0.16)
Experienced	0.93 (0.08)	0.76 (0.09)	0.88 (0.17)
General	0.84 (0.12)	0.59 (0.13)	0.76 (0.19)

We found that the precision of the novice CA was 0.80, while the general version had a precision of 0.84, meaning 20% and 16% of the identified problems were false positives, respectively. In contrast, the experienced CA had a precision of 93%, with only 7% being false positives. All conditions had lower recall than precision, with the novice CA identifying just 54% of all ground truth problems and the experienced CA identifying 76%. This pattern is comparable to the evaluation of ChatGPT's performance in prior work [57], where precision was higher than recall, suggesting that while not every problem was detected, the ones identified were relevant.

To compare the three conditions, we conducted a one-way ANOVA on suggestion precision ($F_{2,28} = 17.3, p < .0001, \eta_p^2 = 0.55$), recall ($F_{2,28} = 82.3, p < .0001, \eta_p^2 = 0.85$), and response accuracy ($F_{2,28} = 4.0, p < .05, \eta_p^2 = 0.22$), all of which showed significant differences. Posthoc pairwise comparisons revealed significant differences between the novice and experienced CAs in suggestion precision ($p < .01$), recall ($p < .0001$), and response accuracy ($p < .05$). However, the differences between the general version and the novice CA were not significant (all $p > .05$). Significant differences were only found between the general version and the experienced CA in precision ($p < .05$) and recall ($p < .001$), but not in response accuracy. Thus, the experienced CA achieved higher precision, recall, and response accuracy compared to both the novice and general versions, which performed similarly.

We also compared the suggestion content across the three conditions when they identified the same problem. To minimize confounding factors such as wording, semantics, or tone, our prompts instructed the CAs to provide responses in a consistent format, without assigning any personality or stylistic variations (Table 5). Table 3 shows representative examples of suggestions from each

---

[2]https://chatgpt.com/gpts

Table 3: Example usability problem suggestion for the three conditions, with differences bolded and underlined

Condition	Problem Description	Cause of Problem	Redesign Recommendation
Novice	Confusion about how to save the current design.	The platform did not provide clear visual feedback or confirmation when a design was saved.	Include a 'Save' button and visual confirmation that the design has been saved.
Experienced	The user is unsure if the design is automatically saved **and how to proceed with caption generation**.	There were no clear visual cues or confirmations indicating that work is saved, **nor a clear transition to the next steps for social media sharing.**	Introduce an auto-save feature and corresponding notifications, **and a clear 'Next Steps' feature that leads users to caption and hashtag generation after the poster is saved.**
General	The user is unsure if their work is being automatically saved.	Lack of clear visual feedback or a message indicating that the design was saved.	Introduce an auto-save feature with a clear indicator and offer a manual save option.

condition, demonstrating that the structure and tone of the outputs were comparable. In the example from Table 3, the experienced CA identified an additional issue (e.g., how to proceed with caption generation) and provided more specific causes and redesign recommendations. The novice and general versions were largely similar, although the general version recommended an auto-save feature in addition to the manual save option. Since the novice CA performed comparably to the general version in terms of precision, recall, accuracy, and content quality, we excluded the general version from the user study. We focused on the novice and experienced CAs, which aligns with Nelson and Stolterman's concept of the "ultimate particular," which emphasizes engaging with specific, well-defined design instances rather than generic or universal ones [76]. In our case, the novice CA served as a "floor" and the experienced CA as a "ceiling," creating the strongest contrast in perceived expertise. This contrast allowed us to examine how evaluators' analytic performance and perceptions evolved across distinct levels of CA capability. Furthermore, limiting the study to two CA conditions helped reduce participant workload and session length, thereby lowering the risk of dropout.

4 Usability Analysis Tool

Once we had created the two CAs, we needed to build a tool that allowed UX evaluators to interact with the CAs while analyzing usability videos. In line with prior human-AI collaborative analysis tools [55, 57], the usability analysis tool includes a **video player** for UX evaluators to review recordings (Fig. 2-A). Additionally, we introduced a **problem table** (Fig. 2-B) to facilitate easier management of problem descriptions by UX evaluators. In conditions where participants use a CA, a **chat window** displays automatic suggestions and the conversation thread (Fig. 2-C). The baseline condition lacks this chat window (Fig. 13 in the Appendix). The chat window is used exclusively for communicating with the CA, while the problem table serves to collect confirmed usability problems identified by UX evaluators.

When automatic suggestions appear for the first time, they come with three action buttons:

(1) **Accept**, which directly adds the suggestion to the problem table

(2) **Edit**, which opens a modal box for UX evaluators to modify the suggestion before adding it to the problem table

(3) **Ignore**, which does not add anything to the problem table

The design of these actions is based on research on AI-assisted decision-making tools, which outlines typical responses to AI recommendations: accept/trust, ignore/reject, or edit/negotiate [113]. Once the UX evaluator selects an action button, the suggestion is updated with a status badge (e.g., Accepted). The corresponding usability problem then appears in the Problem Table with a robot icon, indicating it was authored by the CA (🤖).

In addition to adding problems based on automatic suggestions, UX evaluators can click the "**+ Add Problem**" button (+ Add Problem). This action opens an empty modal form (Fig. 13), allowing UX evaluators to manually enter usability problem descriptions, causes, and redesign recommendations. These manually added problems then appear in the Problem Table with a user icon, indicating manual addition (👤). For problems edited by UX evaluators, both robot and user icons are displayed side by side.

When UX evaluators click on the timestamp at the top right corner of each suggestion (e.g., ⏱ 1:10), the video player jumps to the corresponding timestamp, allowing UX evaluators to easily navigate and rewatch specific video segments. Additionally, if UX evaluators have questions about a suggestion, they can type their messages into the text input field and receive responses. Responses appear in blue to distinguish them from automatic suggestions, which are shown in yellow. Finally, our interface supports both light and dark modes (Fig. 14 in the Appendix).

5 User Study

To answer our RQs on the long-term use of CAs in usability analysis and the impact of different UX experience levels, we conducted an IRB-approved within-subjects study with three conditions: 1) no CA (baseline), 2) novice CA, and 3) experienced CA.

5.1 Participants and Apparatus

We recruited 12 (10 females and 2 males) participants with an average of 4.6 years of UX experience (SD = 2.1). Eight participants were UX professionals, and four were HCI graduate students who had completed UX internships. Their average age was 28 years (SD
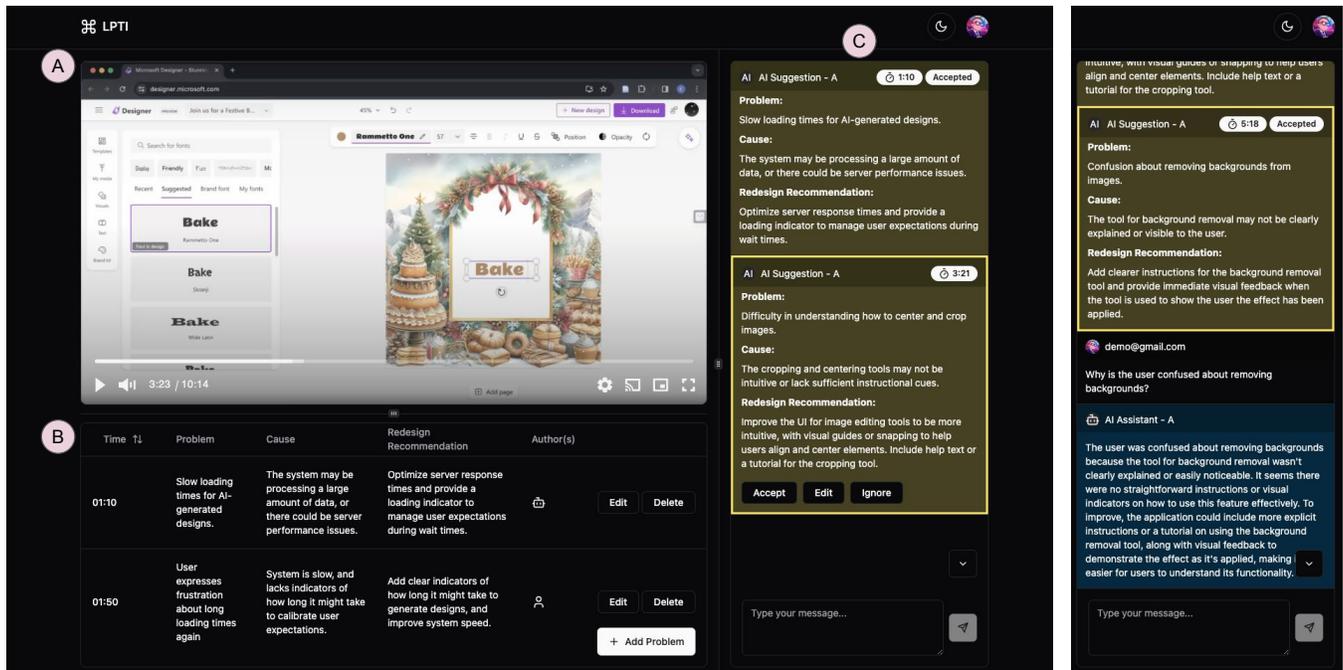
Figure 2: User interface of the usability analysis tool for the conditions with CAs containing: A) video player, B) table of usability problem descriptions, causes, and redesign recommendations with the corresponding timestamp, and C) chat thread.

= 4). Figure 3 shows participants' familiarity with usability analysis (Md = 4, IQR = 0.5), frequency of AI tool usage (Md = 4, IQR = 1.5), and understanding of how AI works, including how LLMs generate responses (Md = 3, IQR = 1.5). All participants had prior experience using AI tools, such as ChatGPT, for tasks like content generation, writing, and qualitative analysis (e.g., surveys, interviews). However, most lacked a deep understanding of how these tools work. The study sessions were conducted remotely through Zoom, and video recordings were captured for subsequent analysis.

5.2 Study Videos

Since there is no standardized taxonomy of products for evaluating usability analysis tools, we curated a set of usability videos to facilitate analysis, following previous research (e.g., [25, 26, 55, 115]). To ensure diversity in interface types, we selected three distinct platforms—a desktop website, a smartphone app, and a VR headset—with five users testing each product, yielding a total of 15 videos for analysis. The rationale behind involving five users for testing each product aligns with the optimal sample size for usability studies, grounded in a cost-benefit analysis [71, 78].

The tasks in the video focused on the fundamental functionalities of the respective products, such as seeking the weather forecast on a weather app or creating a poster using a generative AI (GenAI) design tool. The users in the videos were either recruited from the community or were students from the authors' institution. As they had no prior experience with these products, they encountered a wide range of usability issues, providing rich material for evaluating usability analysis approaches. Table 6 in the Appendix details each video, including its duration, associated tasks, and the number

of problem suggestions from each CA. The novice CA found an average of 6.1 usability problems (*SD* = 1.2) per video, while the experienced CA found an average of 7.7 problems (*SD* = 1.8), which aligned with our evaluation in Section 3.4. Paired Bonferroni-corrected t-tests revealed significant differences between the two conditions ($p < .01$). By ensuring that each video contained at least 4 usability problem suggestions, we aimed to provide participants with ample material for analysis and interactions with the CA.

5.3 Procedure

The study used a within-subject design in which all 12 participants completed five sessions, each reviewing three videos, one per condition. Across five sessions, each participant analyzed 15 unique videos (5 sessions × 3 videos/session = 15 videos) and spent over 6 hours with the tool. The three conditions (no CA, novice CA, and experienced CA) were counterbalanced using a balanced Latin square design [104]. The videos were similarly counterbalanced to ensure each appeared with different products across the conditions. With 15 videos and three conditions, there were 45 unique video-condition combinations (15 × 3 = 45). Thus, each video-condition combination was reviewed by four participants (12 participants × 15 videos ÷ 45 video-condition pairs = 4).

We did not disclose the differences between the two CA versions to participants, only referring to them as Version A or B. To ensure this, we reviewed all conversation logs between the participants and the two CA versions to verify that their roles were not inadvertently revealed during the study sessions.
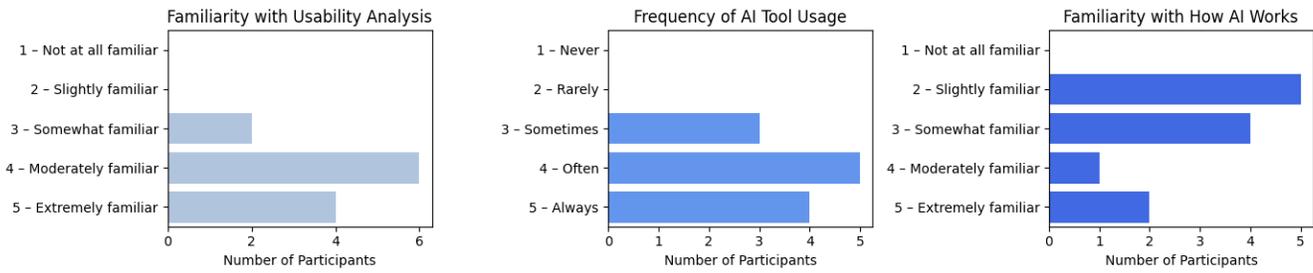
Figure 3: Bar charts showing participants' familiarity with usability analysis, frequency of AI tool usage, and familiarity with how AI works.
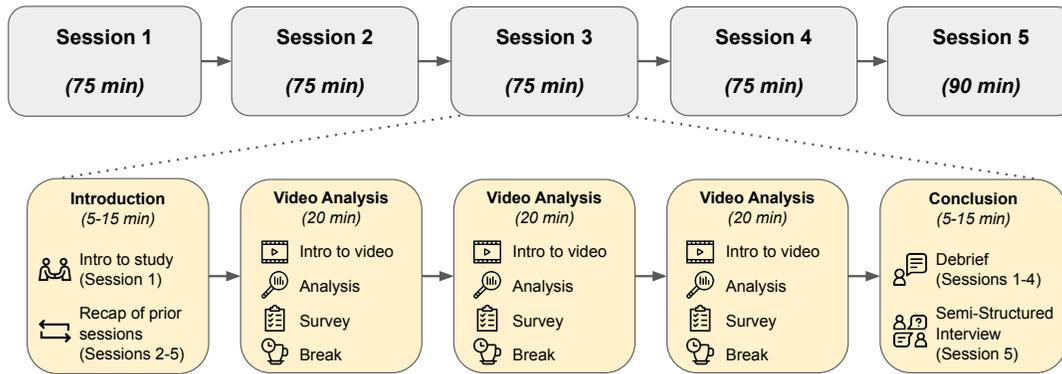

Figure 4: Flowchart of the longitudinal study containing five sessions in total. Each session included an introduction or recap, an analysis of three videos, and a debrief or semi-structured interview.

Figure 4 outlines the study procedure. The first session included an introduction to the study, a tutorial on the tool, and a Q&A session to address any questions. In each session, participants received background information on each of the three videos before starting their analysis, followed by a survey. The survey assessed efficiency, trust, suggestion completeness, confidence in analysis, and perceived UX expertise of the CA. Sessions 1-4 concluded with a brief debrief to gather feedback and confirm the next session. In session 5, we conducted a semi-structured interview to gather participants' overall experience, feedback, and suggestions for improvement.

5.4 Data Analysis

We collected the following data: (1) usability problem descriptions, causes, and redesign recommendations; (2) conversation logs between participants and the CAs; (3) interaction logs in the analysis tool; and (4) video recordings of each session. To analyze usability problem descriptions, we labeled and identified unique usability problems. We calculated Fleiss' kappa, a generalization of Cohen's kappa, to assess inter-rater reliability among multiple raters [73]. We coded the conversation logs to categorize messages by content. We graphed the interaction logs to identify video playback behaviors and subsequent actions after suggestions. Semi-structured interview responses were transcribed using speech-to-text software and manually corrected. We applied inductive coding to analyze transcripts individually, followed by thematic grouping through discussions [10].

For all quantitative data, we used the Shapiro-Wilk test to check the normality. To examine the impacts of long-term usage (RQ1) and CA expertise (RQ2), we conducted a two-way ANOVA with two within-subject factors to determine significant differences across *sessions* and *conditions*. Effect sizes are reported using partial eta squared (η_p^2), and post-hoc pairwise comparisons were made with Bonferroni correction. Statistical test results are summarized in Table 8 in the Appendix to avoid cluttering the main text.

6 Results

The aims of our data analysis are twofold: to understand how UX evaluators' analytic behaviors and attitudes toward a CA evolve (RQ1) and to explore the impact of novice and experienced CAs on UX evaluators' behaviors and attitudes (RQ2). We divided the results into three subsections: 6.1: usability video analysis behaviors, 6.2: analytic performance, and 6.3: subjective feedback on human-AI collaborative analysis, with a summary at the beginning of each.

6.1 Usability Video Analysis Behaviors

> **Summary for RQ1 (Change over time)**: *Participants' analysis strategies evolved from a two-pass to a one-pass approach, enhancing efficiency. A novelty effect emerged, as participants increasingly ignored problem suggestions from the CA after the first session. Moreover, their messages progressed from simple questions to more complex task requests.*
>
> **Summary for RQ2 (Impact of expertise)**: *Using CAs introduced the One-pass, No-Pause-Write strategy, which was absent in the baseline. Participants accepted significantly more and ignored fewer suggestions when using the experienced CA than the novice CA. Additionally, messages sent to the experienced CA were more evenly distributed across different categories, while those to the novice CA were primarily focused on problem identification and generation.*

6.1.1 Video Analysis Strategies when Working with the CAs. In our study, twelve participants reviewed 15 videos, resulting in 180 recorded sessions. We examined how many passes they made on a video and their playback behaviors when going through a pass. Fig. 5 shows six timeline plots of typical behaviors when analyzing one video. These examples were selected to represent the main analysis strategies observed, rather than to show an idealized progression. Fig. 6 displays the distribution of the five main analysis strategies across 180 sessions:

1) Two-pass, Overview-Write: This strategy involved first watching the video from beginning to end to gain an overview before going back to add problems in the second pass (Fig. 5a and Fig. 5c). It was employed for 10 videos (5.6%) by four participants and only appeared in the first session.

2) Two-pass, Write-Check: This strategy is similar to the previous one, but participants recorded problems or responded to suggestions during the first pass and then went back to check (Fig. 5b). It was employed for 20 videos by seven participants in the first two sessions, and with one video in session 3 (11.1% of all videos).

3) One-pass, Pause-Write: This was the most frequently employed strategy and was used by all participants, accounting for 82 videos (45.6%). Participants paused to record usability problems or respond to suggestions while reviewing the video once (Fig. 5d).

4) One-pass, Micro-Playback-Write: This strategy involved rewinding a small portion of the video before recording a problem and was employed for 36 videos (20%) by eight participants. For example, Fig. 5e shows P11 rewinding between 400 to 500 seconds to review and add a problem. Other instances of micro-playback occurred in response to an automatic suggestion, where participants rewound the video to check whether they agreed with the suggestion. Fig. 6 shows that this strategy was more frequently used in the first two sessions (17 videos) than the last two (12 videos).

5) One-pass, No-Pause-Write: This strategy involved minimal pausing, with participants directly accepting or editing suggestions as they appeared. Fig. 5f shows a straight line where P10 accepted all suggestions. This strategy was applied to 32 videos (17.8%) by eleven participants, but it only appeared in conditions with CAs, as the baseline condition required pausing to type problem descriptions. It emerged in the last three sessions, accounting for two-thirds of videos by session 5.

6.1.2 Responses to Automatic Suggestions from the CAs. In this section, we describe the three types of responses to automatic suggestions: accept, edit, or ignore.

Rate of Accepted Suggestions. We found that only the effect of *condition* was significant, meaning that using the CA over time did not significantly impact how much participants accepted the suggestions. For the comparison between CA conditions, participants accepted significantly more suggestions from the experienced CA ($M = 66.5\%, SD = 23.6\%$) than the novice CA ($M = 56.0\%, SD = 26.0\%$) throughout all sessions. Fig. 7a shows a box plot of the percentage of accepted suggestions separated by session and CA condition, where the boxes for experienced CA generally rested higher than those of the novice CA. Participants noted that the experienced CA provided more detailed and specific suggestions, while the novice CA offered general observations while providing superficial causes (e.g., stating that the user expressed frustration without identifying its underlying cause). For instance, although both CAs identified the same problem in the VR game, the novice CA's suggestion to "add a tutorial" was seen as too vague and failed to recognize that a tutorial already existed but was difficult for the user to find (P2). In contrast, the experienced CA offered specific recommendations, such as making the tutorial button more noticeable by increasing its size and contrast against the background and suggesting exact text edits. As a result, the experienced CA's suggestions were accepted more often than those of the novice CA.

Rate of Edited Suggestions. There were no significant patterns for the edit rate, suggesting that how much participants edited the suggestions did not change over time or for different conditions (shown in Fig. 15 in the Appendix).

Rate of Ignored Suggestions. Both the effect of *session* and *condition* were significant for the ignore rate. Post-hoc pairwise comparisons between session pairs showed that the ignore rates in session 2 ($M = 16.2\%, SD = 13.6\%$), session 3 ($M = 13.3\%, SD = 13.8\%$), and session 4 ($M = 17.8\%, SD = 19.0\%$) were significantly higher (all $p < .05$) than the ignore rates in session 1 ($M = 8.0\%, SD = 15.4\%$). The difference in ignore rates demonstrates a novelty effect where participants were less likely to ignore suggestions at the beginning. From session 2 onward, participants reported that they began to recognize the limitations of the CAs, and as a result, started ignoring more suggestions. Fig. 7b shows a box plot of the percentage of ignored suggestions separated by session and CA condition. The boxes for experienced CA are always lower than the novice CA since participants ignored 19.7% of all suggestions from the novice CA while only 6.3% of suggestions from the experienced CA. Many participants felt that the novice CA often provided irrelevant suggestions (false positives), which aligns with our evaluation in Section 3.4. P3 noted that it felt too sensitive, often flagging minor delays as problems. For instance, P8 believed that a user's difficulty completing a task was due to unfamiliarity with the app, not a problem with the app itself. Similarly, P5 interpreted a user's statement of "I do not like this" as a personal preference rather than an app problem. As a result, participants disregarded these suggestions, leading to higher ignore rates for the novice CA.
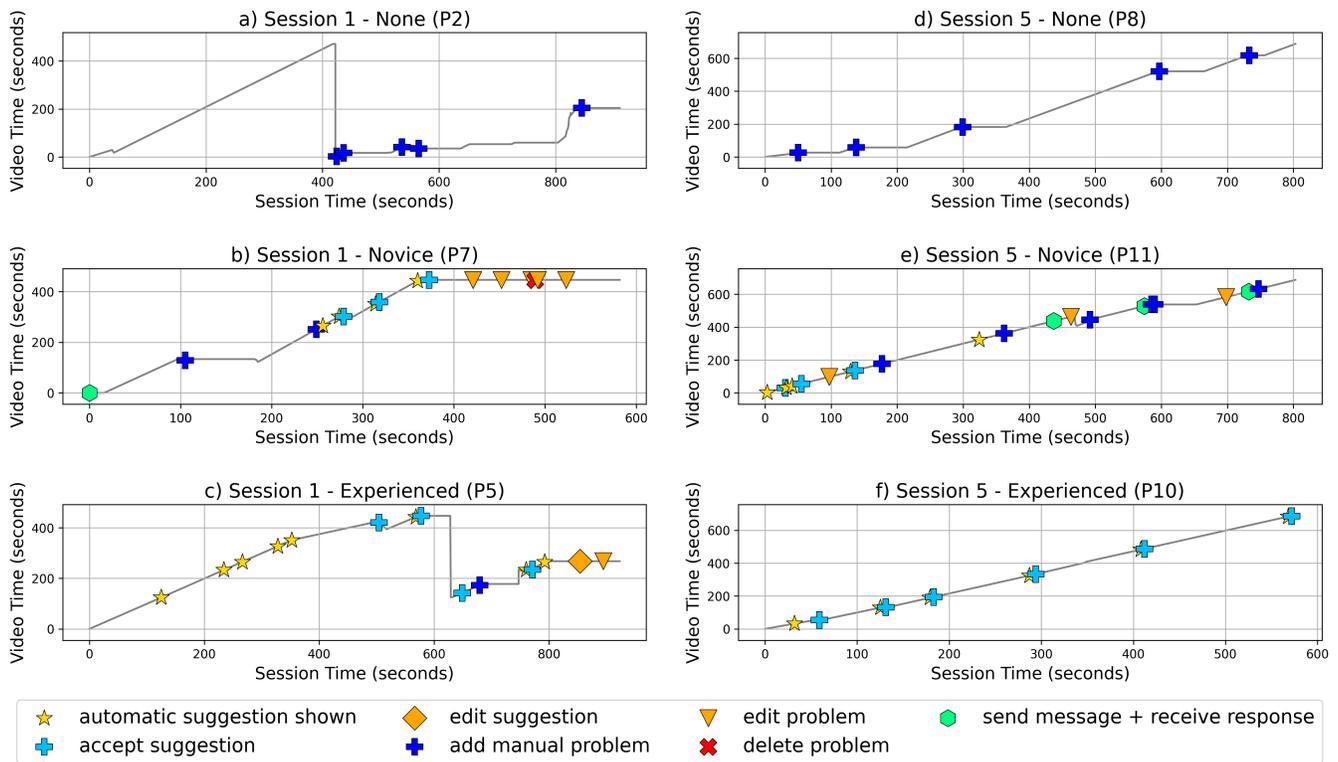
Figure 5: Timeline plots of analysis behaviors with session time on the x-axis and video time on the y-axis. The first column shows the three conditions (no CA, novice CA, and experienced CA) during session 1, while the second column presents these same conditions during session 5. These video timelines were chosen to illustrate the five main analysis strategies visually.
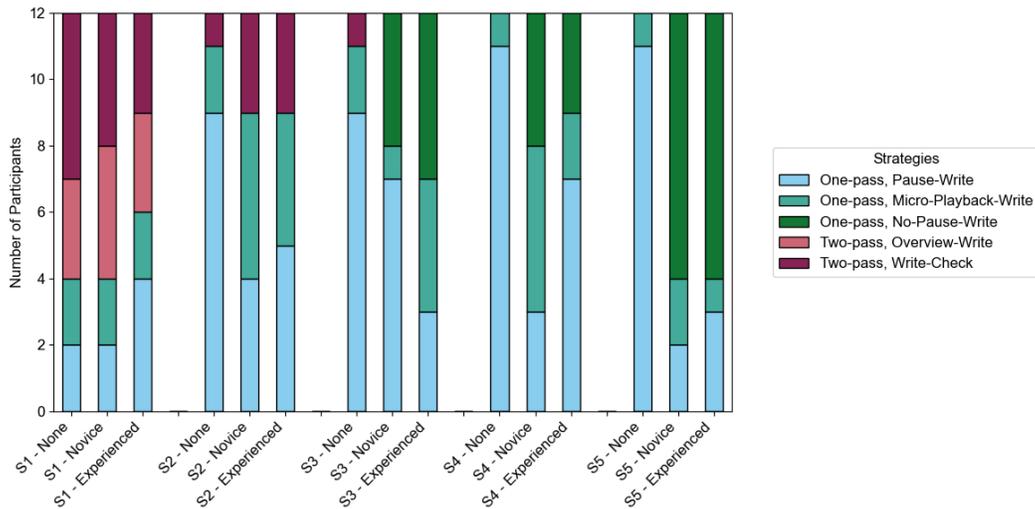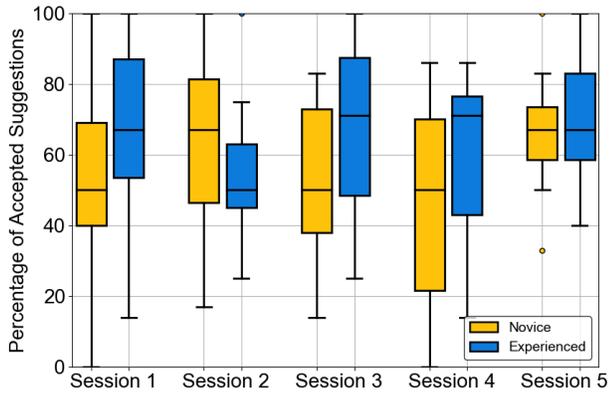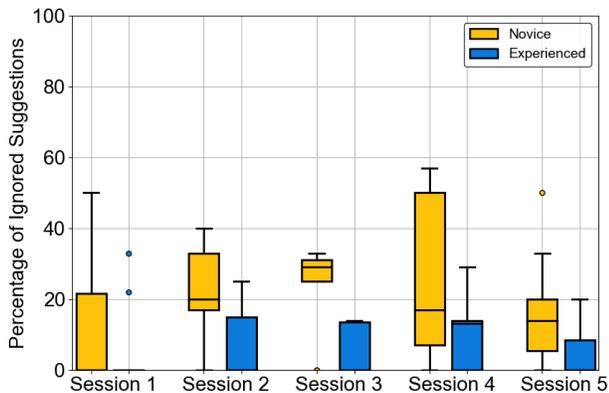

Figure 6: Stacked bar chart showing the distribution of five video analysis strategies by session and condition. Two-pass strategies were more frequent in the first two sessions, while the One-pass, No-Pause-Write strategy emerged in session 3 and was used by two-thirds of participants by session 5.

(a) Acceptance rates of novice and experienced CA's suggestions.

(b) Ignore rates of novice and experienced CA's suggestions.

Figure 7: Box plots displaying the percentage of (a) accepted suggestions and (b) ignored suggestions, separated by session and CA condition. Acceptance rates were generally higher for the experienced CA compared to the novice CA, while ignore rates were higher for the novice CA. Additionally, ignore rates for both CAs increased after session 1.

6.1.3 Categories of Messages Sent to the CAs. In total, participants sent 170 messages to the CAs, and we identified 17 different categories of messages. Table 7 in the Appendix displays the categories, frequency, and example messages.

Impact of Longitudinal Use. Since there were no significant differences in the number of messages between sessions, we focused on analyzing trends in message categories and their distribution across sessions (Fig. 8a). The category with the highest number of messages (39 in total) was **problem identification**. In this category, participants questioned whether the CA had missed any usability problems or if additional problems could be identified, indicating that this was a primary concern during the analysis. This was the only message category that appeared in all five sessions.

In the initial sessions, participants tended to ask simpler questions, such as requesting information on tasks, products, and users.

They also explored the CA's capabilities by inquiring whether it could generate a problem, identify causes, or provide redesign recommendations based on the information they supplied. In the later sessions, participants increasingly relied on the CA for more complex tasks, such as generating precise wording for identified problems and prioritizing the list of usability problems after analyzing the video. They also engaged in more dialogue turns, such as informing the CA of additional causes it had missed and seeking clarification on specific recommendations.

Impact of Varying UX Experience. Fig. 8b shows the categories and number of messages sent to the novice and experienced CAs. Participants sent 89 messages to the novice CA and 81 messages to the experienced CA, with no significant difference between the two conditions. However, the content of these messages were different. For the novice CA, the messages mainly concentrated on two categories: identifying usability problems (30.3%) and generating problem descriptions (15.7%), reflecting more task-oriented exchanges. In contrast, message to the experienced CA were more evenly distributed and included two categories that did not occur for the novice CA: redesign recommendations and user background, indicating more strategic and reflective interactions. Participants also engaged more in user perception and suggestion clarification and correction messages with the experienced CA, suggesting greater willingness to discuss higher-level design rationale and to negotiate the AI's output. This pattern implies that as perceived AI expertise increased, participants shifted from basic information extraction toward collaborative reasoning.

6.1.4 Individual Differences in Analysis Behaviors Among Participants. Since our sample was small, we did not compute correlations between participants' UX backgrounds or AI usage habits and their behaviors. However, we observed several qualitative trends. Participants who used the "Two-pass, Overview-Write" strategy tended to be more meticulous, had less UX experience (2-4 years), and were only somewhat familiar with AI. For example, P2, P5, and P7 carefully validated the CAs' suggestions and reviewed video context before writing, using this strategy across all three videos in their first session. As they became more familiar with the CAs, they shifted to faster approaches. In contrast, P10 and P12, who had more UX experience (7-8 years), higher familiarity with usability analysis (rated 4-5), and frequent AI use (rated 5), immediately adopted the "One-pass, Pause-Write" strategy and even sped up videos in the first session.

We also observed differences in message categories. Problem prioritization emerged only in the final session and was primarily used by P12, who described prioritization as part of her normal workflow when completing the analysis. Problem generation was common among P5, P6, and P11 once they discovered that CAs could help articulate findings. For instance, P11 prompted the novice CA with: *"Provide the problem, cause, and redesign recommendation for the user's confusion about reload."* In interviews, P11 explained that the CA sometimes missed issues but was faster at writing out the descriptions once she identified the problem.

Acceptance rates also varied significantly across participants for both the novice CA ($F_{11,48} = 2.62, p < .05, \eta_p^2 = 0.38$) and expert

(a) Message categories by session.
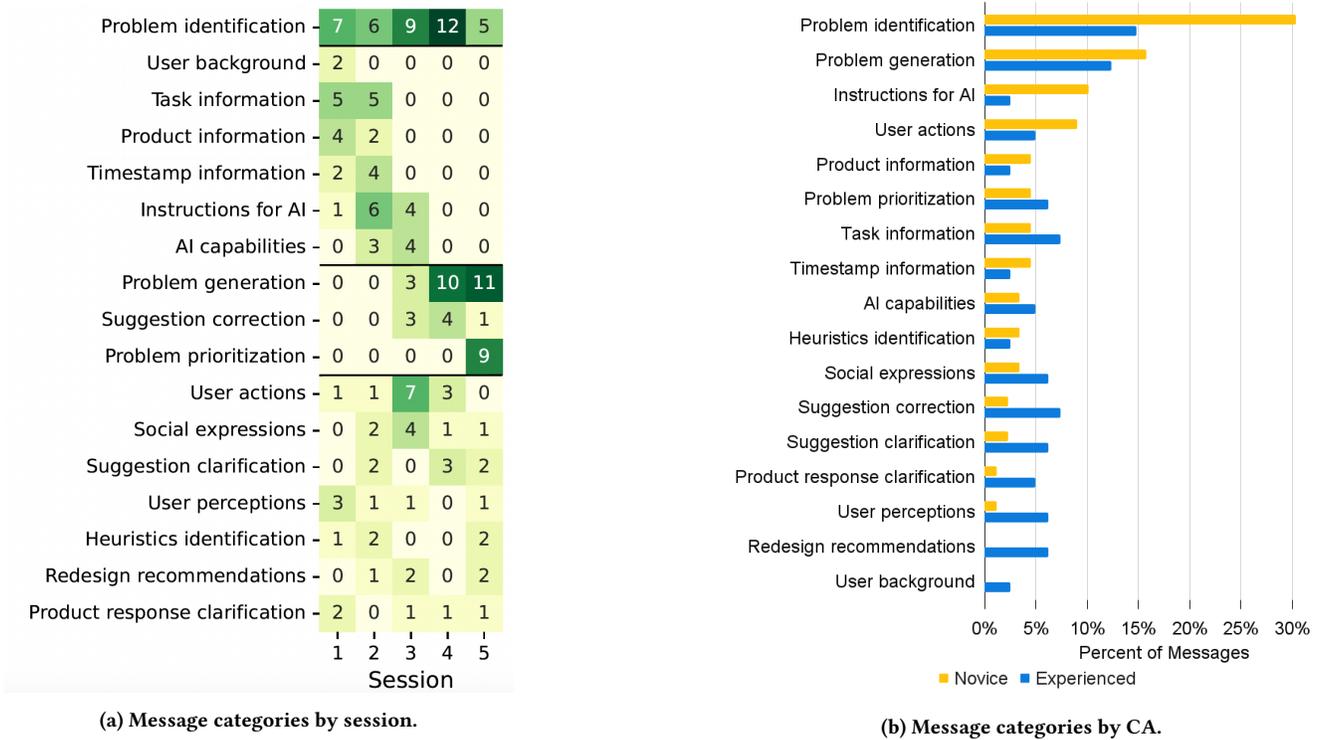

(b) Message categories by CA.

Figure 8: Visualizations of message categories: (a) Heatmap by session: Problem identification appeared in all sessions. The next six categories were present only in the first three sessions, and the next three categories only in the last three sessions. The remaining categories showed no clear pattern. (b) Bar chart by CA: Message categories were more evenly distributed for the experienced CA.

CA ($F_{11,48} = 2.81, p < .01, \eta_p^2 = 0.39$), based on one-way between-subjects ANOVA. Participants with the highest acceptance rates were P1-P3, who each had 3 years of UX experience and reported being only slightly familiar with AI. In contrast, participants with the lowest acceptance rates (P9-P11) reported higher AI familiarity; for example, P10 rated himself as "5 - extremely familiar." Prior work suggests that experts may be more hesitant to adopt AI because deeper technical understanding often heightens awareness of system limitations [112].

6.2 Analytic Performance

Summary for RQ1 (Change over time): *Participants identified more problems in the final session when using the novice and experienced CAs, indicating that familiarity with the videos and CAs improved their analysis.*
Summary for RQ2 (Impact of expertise): *Using CAs significantly increased the total number of problems, unique problems, problem coverage, and inter-rater reliability (Fleiss' kappa) compared to the baseline. The experienced CA produced significantly better results than the novice CA.*

6.2.1 Number of Usability Problems. In total, participants identified 1217 problems from the 15 usability videos: 330 on their own, 393 with the novice CA, and 494 with the experienced CA. Fig. 9a shows the total number of identified problems per video separated

by session and condition, while Table 8 shows the statistical test results. Both the effect of *session* and *condition* were significant. Post-hoc pairwise comparisons between session pairs showed that the number of identified problems in session 5 ($M = 7.5, SD = 3.1$) was significantly higher ($p < .01$) than in session 3 ($M = 6.3, SD = 2.0$). This difference could be attributed to participants being already familiar with the common usability problems encountered by session 5 and relying more on the CAs at the end.

Table 4: Analytic performance measures based on condition, reported in "Mean (Standard Deviation)"

Variable	No CA	Novice CA	Experienced CA
Total problems per participant per video	5.5 (2.4)	6.6 (1.4)	8.2 (2.5)
Unique problems per video	8.6 (1.6)	9.3 (1.6)	10.3 (2.3)
Problem coverage per video	70.9% (13.3%)	76.3% (7.8%)	84.3% (8.3%)
Fleiss' kappa per video	0.40 (0.17)	0.58 (0.18)	0.62 (0.15)

Table 4 shows the mean and standard deviation of the number of problems identified per participant based on condition. Post-hoc pairwise comparisons showed that the number of problems found by the participants using the novice and experienced CAs was significantly higher than no CA assistance ($p < .01$ and $p < .0001$, respectively). Furthermore, there were significant differences between CA conditions, where participants using the experienced CA identified significantly more problems than the novice CA ($p < .0001$). This shows that using the experienced CA led to the best results in the number of identified problems.
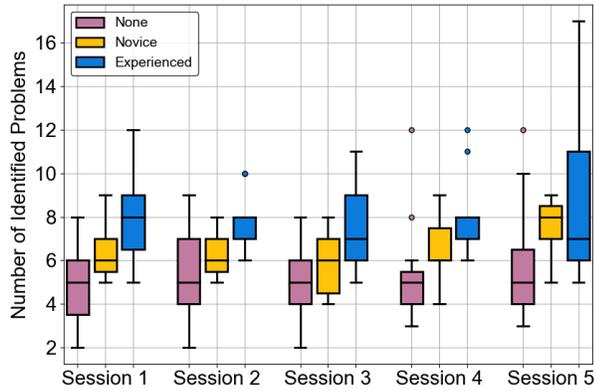
(a) Number of identified problems.
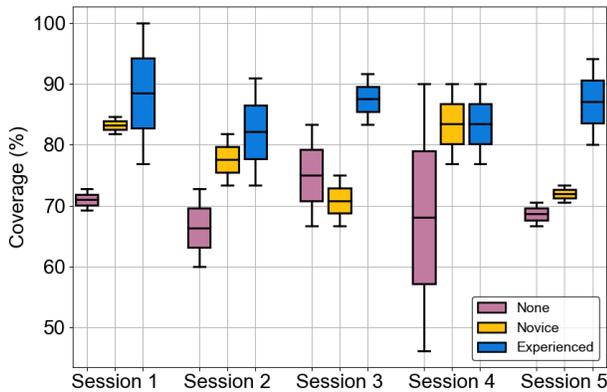
(b) Coverage of unique problems.

Figure 9: Box plots show (a) Total number of identified problems: Problems were highest in session 5, and were higher with the experienced CA, followed by the novice CA, and least with no CA. (b) Coverage of unique problems: There were no session-based patterns, but the experienced CA achieved higher coverage compared to both novice CA and no CA.

6.2.2 Number of Unique Problems and Coverage. In addition to counting the total number of problems, we also examined the number of unique problems for each video (Table 4). As mentioned in Section 5.3, four participants reviewed each video-condition

combination. The number of unique usability problems was then calculated as the total number of distinct problems identified across those four participants. Only the effect of *condition* was significant. Post-hoc pairwise comparisons showed that the number of unique problems with the experienced CA was significantly higher than without any CA assistance, with $p < .05$. The novice CA was not significantly different than the other two conditions.

We also calculated the coverage of each condition by dividing the number of unique problems by the total unique problems across all twelve participants for each video (Fig. 9b). Similar to unique problems, only the effect of *condition* was significant. Post-hoc pairwise comparisons showed that the coverage with the experienced CA was significantly higher than without any CA assistance ($p < .01$) and the novice CA ($p < .05$).
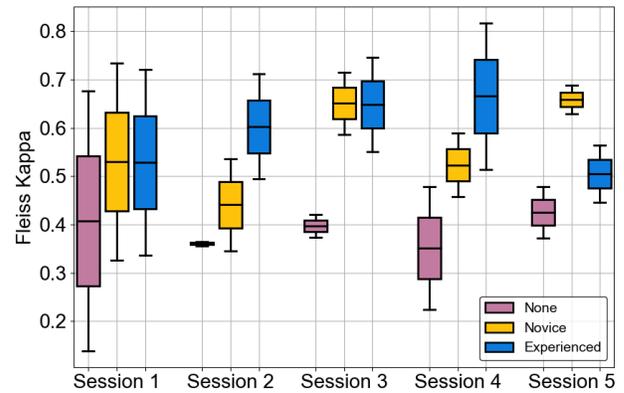
Figure 10: Box plot showing Fleiss' kappa separated by session and CA condition.

6.2.3 Inter-rater Reliability. In line with prior work using Fleiss' kappa as a measure of inter-reliability between ratings from Chat-GPT and humans [48], we also computed this value based on the results from the four participants in each condition for each video. Fig. 10 shows a box plot of Fleiss' kappa separated by session and condition. Only the effect of *condition* was statistically significant. Post-hoc pairwise comparisons showed that values in the novice and experienced CA were significantly higher than without any CA assistance ($p < .01$ and $p < .001$ respectively). Interpreting the kappa values from Landis and Koch's table [59], our findings indicate there was moderate agreement (0.4 - 0.6) for the no CA and novice CA conditions, while experienced CA showed substantial agreement (>0.6).

6.3 Subjective Feedback on Human-AI Collaborative Analysis

> **Summary for RQ1 (Change over time):** *For both CAs, the novelty effect caused a drop in perceived efficiency and trust in session 2 compared to session 1, which recovered from session 3 onward. Participants began to notice differences in UX expertise between the two CAs from session 2 onward.*
>
> **Summary for RQ2 (Impact of expertise):** *Participants rated the experienced CA as significantly more efficient, trustworthy, and complete in its suggestions, which increased their confidence. Overall, the experienced CA was preferred.*

Fig. 11 shows the ratings on a 5-point Likert scale for the efficiency of using the CA, trust in the CA, perceived completeness of suggestions, confidence in analysis, and perceived UX expertise of the CA.

Efficiency of Collaborating with the CA. Overall, participants felt that using the CAs was very efficient for their analysis ($Md = 5, IQR = 1$). P12 mentioned, *"it was very efficient and made me feel like I was getting through analysis much quicker without typing everything myself."* When comparing participants' ratings across sessions and conditions, both factors led to significant differences. Post-hoc pairwise comparisons between session pairs showed that the efficiency ratings in session 2 were significantly lower than the other four sessions. In session 2, participants began to notice the CA's limitations, indicated by asking more questions about its capabilities and starting to figure out how to best collaborate with the CA. For instance, P2 mentioned in session 2's interview: *"I noticed a problem and waited for the AI suggestion so I wouldn't have to type it, but when it didn't detect the problem, I had to type it manually and even rewatch the segment to frame the statements accurately."* Overall, participants rated the experienced CA as significantly more efficient than the novice CA since the novice CA often missed problems, requiring participants to add them manually. P9 noted, *"It didn't catch the things I was catching. Sometimes, multiple usability issues occurred in succession, but it only identified the first one, so I had to type the rest in myself, which took more time."* In contrast, in the last 2-3 sessions, participants using the experienced CA increased their efficiency by speeding up the videos. P1 explained, *"I actually sped up the video because I know it will point out the right things. I don't double check the suggestions from the [experienced CA], but I always check for the [novice CA]"*-P3.

Trust in the CA. Overall, participants somewhat agreed that they trusted the information provided by both CAs ($Md = 4, IQR = 1$). The trust ratings were significantly different depending on the *session* and "condition." Similar to the efficiency ratings, post-hoc pairwise comparisons between session pairs showed that the trust ratings in session 2 were significantly lower than the other four sessions (*all Md = 4, IQR = 1*). This difference shows an overall trend where both the perceived efficiency and trust were high in session 1, decreased in session 2, and then recovered from session 3 onwards. P4 explained the reasons behind this trend: *"I was very impressed in the first session, but I got more critical once I noticed mistakes in session 2. After session 3, I knew exactly what to expect from the videos and the assistant, so my trust and speed increased until the end."* Conversely, other participants felt skeptical during

the first session, then their trust gradually increased throughout the sessions because *"the more correct examples I saw, the more I trusted it"*-P6.

The levels of trust were also significantly different depending on the condition. P10 mentioned that their *"trust in [experienced CA] increased, but decreased for the [novice CA]."* Similarly, P12 noticed that novice CA did not always provide appropriate redesign recommendations, so they had to fill in these missing details, decreasing their overall reliance on this version.

Perceived Completeness of Suggestions. Participants agreed that the CAs provided a complete list of suggestions overall ($Md = 4, IQR = 1$). Only the effect of *condition* led to significantly different ratings, indicating that participants' opinions on completeness did not change over time. However, they noticed that the experienced CA provided significantly more complete suggestions than the novice CA. In particular, P7 found the novice CA's suggestions helpful when they appeared but felt they were often missing. She emphasized the importance of consistency, which lowered her trust in the novice CA, stating, *"I had to be more active and alert because I knew it wasn't reliable."*

Confidence in Analysis. Participants generally agreed that the CAs increased their confidence in usability analysis ($Md = 4, IQR = 1$). P11 described the CAs as *"a second opinion, which is always nice to have. I'm not afraid to disagree; I have the power to decide whether to accept it, but it could be especially helpful when I'm unsure."* Although participants' ratings for confidence remained consistent over time, they reported feeling significantly more confident when using the experienced CA compared to the novice CA.

Perceived UX Expertise. Participants rated the CA's level of UX expertise from 1 (none) to 5 (expert). The ratings were significantly different between sessions and conditions. Interestingly, the interaction effect was also significant, indicating that the two variables of session and condition depend on each other. For the novice CA, post-hoc pairwise comparisons between session pairs showed that ratings in session 2 were significantly lower than in session 1. For the experienced CA, the median ratings were consistent between 4 and 5, so there were no significant differences over time. Fig. 11e shows that the ratings for both CAs were nearly the same in session 1, but starting from session 2, the ratings for the novice CA became lower than those of the experienced CA, suggesting that participants started noticing the differences after the first session. In the final interview, participants described the novice CA as an *"intern"* or *"entry-level"* UX evaluator, while they felt that the experienced CA was *"entry-level," "mid-level,"* or *"expert."*

Overall Comparison of Novice and Experienced CAs. In the final interview, participants assigned an overall satisfaction rating from 1 to 10 to each CA. The novice CA received $Md = 7.5, IQR = 1.25$, while the experienced CA achieved $Md = 9, IQR = 0.25$. A one-way ANOVA showed that the difference was significant, demonstrating a clear preference for the experienced CA. P8 mentioned that the experienced CA *"made me more efficient and caught more of the problem than the other version"*. While most participants echoed this sentiment, some still found the novice CA valuable for specific use cases. P10 mentioned that *"it would be good for training new researchers and interns since they can learn effectively with someone*
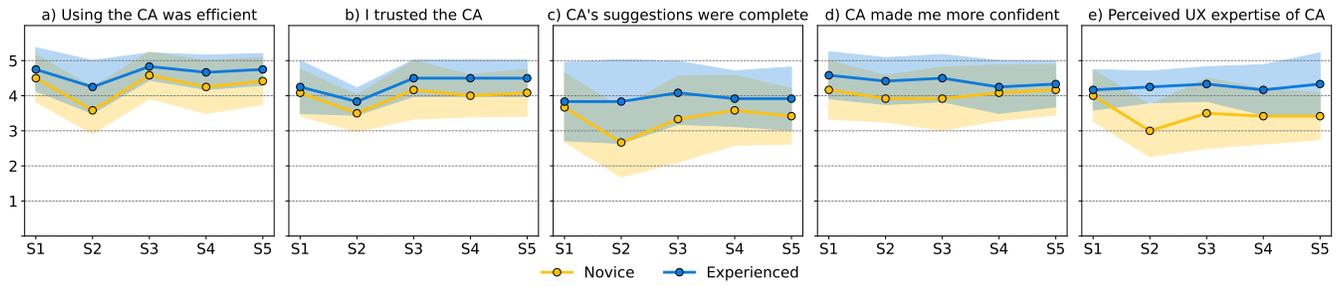
Figure 11: Line charts displaying the mean (circular markers) and standard deviation (shaded area) of Likert ratings for a) perceived efficiency, b) trust, c) suggestion completeness, d) analysis confidence, and e) UX expertise over five sessions. Statements a-d were rated from 1 (strongly disagree) to 5 (strongly agree), while e was rated from 1 (none) to 5 (expert). The novice CA saw drops in efficiency, trust, and completeness in session 2, while the experienced CA showed declines in efficiency and trust. The perceived UX expertise of the novice CA declined starting in session 2.

at their own level. This way they can apply their skills and judge how well the AI is doing without overdepending on the AI since it's not that good." Overall, our study revealed key differences in analytic behaviors and perceptions between novice and experienced CAs, which we elaborate on in the following section.

7 Discussion

This section expands on the impact of multi-session analysis, discussing the implications for longitudinal studies in human-AI collaboration. Additionally, it discusses practical use cases for both novice and experienced CAs.

7.1 Impact of Multi-Session Analysis (RQ1)

7.1.1 Evolution of Video Analysis Behaviors. We identified five main video analysis strategies, consistent with Fan et al.'s findings on using visualizations to flag usability problems [25]. In Fan et al.'s study, participants analyzed three videos in one session, with two-pass strategies common early on. Similarly, our participants initially used time-intensive two-pass strategies. By session 3, however, the one-pass, no-pause-write strategy emerged, and by session 4, two-pass approaches disappeared. This progression reflects participants' growing expertise: as they became more familiar with recurring usability issues and more comfortable working with the CA, their behavior shifted toward the linear, streamlined strategies typical of expert designers [12].

7.1.2 Changes in Attitudes Toward AI. Participants' perceived efficiency and trust in the CAs were high in Session 1 but declined in Session 2, suggesting a *novelty effect*. Similar declines have been observed in human-robot interaction, where enthusiasm often fades as users gain more experience—for example, children's interest in a dancing robot dropped over three months [117], families' excitement about a Roomba decreased over six months [116], and the perceived reliability of a pharmacy robot declined after fifteen months [41]. Another possible explanation is that prior evidence shows *trust often declines over time* as users apply greater scrutiny to AI systems [20, 40, 44]. This may also explain the sharper drop in perceived expertise for the novice CA, who made more early mistakes (e.g., omitting a usability problem) than the experienced

CA, which is an important factor given that first impressions are crucial for trust formation [111] and that failures exert stronger negative effects on trust than successes [137].

After Session 2, trust and efficiency ratings increased again, indicating a familiarization effect, in which users adapt to and internalize new technologies as they gain experience [96]. Prior work shows that as users settle into stable mental models, trust judgments become less volatile over time [137], which may explain why the ratings became stable between sessions 3 to 5. Our results mirror the trajectory reported in gamification systems: an initial dip followed by recovery [96], suggesting this pattern may generalize across technology domains. A recent longitudinal study of generative AI for science communication similarly found that familiarization unfolded over several sessions [66], though differences in session length (20 min vs. 90 min here) limit direct comparison. Consistent with that work, participants in our study used the familiarization period to probe the CAs' capabilities, which aligned with the emergence of message categories such as "AI capabilities" and "problem generation" and ultimately improved collaboration.

Caveats on Interpreting Trust Ratings. Some participants recognized differences between the CA conditions by the second or third session, meaning changes in trust may reflect both adaptation and perceived system differences. In addition, trust was assessed primarily through self-reports, which capture deliberate attitudes but overlook implicit behaviors and remain vulnerable to social desirability bias [34, 40]. For example, users may report trusting an AI while behaviorally avoiding its suggestions [8, 34]. We included acceptance rates as a behavioral proxy, but the highest rates occurred in Session 5—possibly due to increased trust, accumulated familiarity with the task, task saturation, or participant fatigue—and differences were not significant. Future work should incorporate richer behavioral indicators (e.g., intervention frequency, response times, eye-tracking) and physiological measures (e.g., heart rate variability, skin conductance, EEG), which have been shown to approximate trust [1, 49, 133]. Despite having limitations, self-report scales, especially validated instruments such as the Trust in Automation Scale (TIAS), remain practical and widely used for measuring trust in human-AI studies [72].

7.1.3   Implications for Future Longitudinal Studies on Human-AI Collaboration. We found that perceived UX expertise ratings significantly differed between the novice and experienced CAs overall, but these differences were also shaped by *session*. In the first session, participants rated both CAs similarly, showing no clear distinction in perceived expertise. However, across subsequent sessions, ratings for the novice CA declined, while the experienced CA maintained consistently higher scores. This pattern highlights that users' initial perceptions of AI systems can be overly optimistic, shaped by novelty, rather than by accurate assessments of capability. If our study relied on a single-session evaluation, which is common in HCI work, we might have drawn the misleading conclusion that CA expertise does not impact user perception. Instead, our longitudinal approach revealed how these perceptions evolve with continued interaction, as users recalibrate their expectations of the AI over time. These findings underscore the necessity of longitudinal methods in human-AI collaboration research.

Reflections and Lessons Learned. Conducting a multi-session, longitudinal study also offered several methodological lessons. Sustaining participant engagement over time required balancing task familiarity with novelty. Introducing new videos in each session, while keeping the CAs the same, helped maintain interest without altering the study design. Because the participants were professionals with full-time jobs, flexibility in scheduling was essential. Using an online scheduling tool (e.g., Calendly) allowed participants to easily reschedule sessions as needed, reducing attrition and accommodating diverse work hours. Logistically, maintaining consistent procedures and clear communication across sessions was critical for ensuring data comparability while allowing participants' understanding of the system to evolve. We also observed that participants' reflections became richer as familiarity grew, suggesting that later sessions are particularly valuable for capturing deeper insights into human-AI collaboration. Future researchers conducting longitudinal studies may benefit from protocols that explicitly account for engagement, scheduling flexibility, and reflection across time.

7.2   Impact of Perceived UX Expertise (RQ2)

Our study showed that having any CA is better than none, as the no-CA condition resulted in significantly fewer identified problems, fewer unique problems, and lower inter-rater reliability, as indicated by Fleiss' kappa. Moreover, not only is any CA better than none, but the *expertise of the CA also matters*: we found that the experienced CA led to significantly higher trust and efficiency than the novice CA, particularly after participants gained experience with these tools. While this partially aligns with prior research showing that high AI accuracy leads to greater trust in domains like legal decision-making [44], our study extends these findings by investigating a different domain (usability analysis) over a longer time horizon, and by examining how perceived AI expertise influences evaluators' analytic behaviors beyond trust. We further explore the implications of our findings concerning the differences between novice and experienced CAs.

7.2.1   Implications for Real-World Collaboration with the Experienced CA. The overarching goal of usability testing is to identify usability problems and uncover opportunities for improvement [74]. However, the reliability of analysis results is hindered by the

"evaluator effect," where different evaluators tend to uncover different usability problems [37, 38]. Prior work has demonstrated weak inter-rater reliability in heuristic evaluations [114, 131] and high variability between evaluators in usability tests [97]. While the evaluator effect suggests that some core UX methods may be unreliable, agreement is not the only goal of usability testing [101]. Completeness is also crucial, as identifying a diverse set of problems can lead to overall usability improvements.

Prior research recommends collaboration among multiple evaluators to leverage diverse perspectives and identify a broader range of usability problems [38, 101]. Yet surveys show that only 23% [29] and 37% [56] of UX evaluators in industry actually collaborate on analyzing usability videos. In practice, this data means a single evaluator often analyzes multiple videos. Our findings suggest that **collaborating with an experienced CA can effectively provide diverse perspectives without compromising reliability**: the experienced CA achieved the highest problem coverage and inter-rater reliability, making it a valuable aid for evaluators. To further enhance evaluator-CA collaboration, future tools could include configurable prompts that specify the AI's level of expertise. They could also adopt distributed-agent approaches, such as DesignGPT, where a human inputs requirements and collaborates with multiple AI "employees" (e.g., product managers, design directors) [21]. Such capabilities would allow UX professionals to efficiently leverage CAs with greater expertise, improving both the completeness and reliability of usability analyses.

7.2.2   Implications for Real-World Collaboration with the Novice CA. While the experienced CA led to better analytic performance and was generally preferred, the novice CA was still perceived as valuable, particularly for *training purposes*. The UX field is rapidly evolving with the introduction of new tools, design philosophies, and most recently, the integration of AI and machine learning into UX practices [42]. Research has shown that CAs can boost productivity, especially for novice users, by helping them bridge skill gaps and accelerate learning [6]. Building on these insights, we propose leveraging the **novice CA as a training partner to support the development of emerging UX evaluators**, providing a low-stakes environment to practice, reflect, and refine their analytic skills.

This perspective aligns with prior research on the *protégé effect*, which shows that adopting a mentoring role, such as guiding a novice, can enhance one's own learning and critical thinking [11]. Similarly, previous studies have found that UX evaluators perceived AI-generated usability problems as incomplete and likened the AI to a "junior colleague" or "new intern who needs to be monitored" [26]. Our findings echo this dynamic: participants described the novice CA in similar terms and engaged more critically with its suggestions. Recent work also highlights the potential benefits of prosocial attitudes toward AI agents, suggesting that helping AI can enhance human well-being [139]. Future research could build on this by refining the novice-AI framing and exploring how designs that invite humans to "support" AI might better meet human needs and foster well-being.

A common fear surrounding AI is the belief that it will eventually match human capabilities and replace jobs, which has been dubbed the "fear of obsolescence" (FOBO) [132]. In contrast, our

study demonstrated that over time, participants recognized both the capabilities and limitations of the CA, leading to a productive division of labor rather than displacement. Recent work comparing a Nielsen-heuristics-trained GPT with human UX experts also highlighted its difficulty handling ambiguous scenarios, underscoring the need for human oversight in complex cases [120]. Likewise, we argue that human expertise remains essential and that novice CAs may serve as effective training partners, helping evaluators sharpen their skills. For instance, P4, who had two years of UX experience, shared that participating in this multi-session study felt like a training program, improving her attention to detail and articulation of usability problems. Our data also showed higher ignore rates for novice CAs, suggesting that participants actively evaluated and filtered AI suggestions rather than accepting them by default. This process of identifying and correcting AI mistakes encouraged evaluators to reflect on their own reasoning, reinforcing critical thinking and domain expertise. Extending this idea, future work could explore *deliberative AI* for usability analysis, in which humans and AI jointly deliberate to resolve conflicting perspectives in decision-making tasks [70].

Furthermore, the CA's expertise should be considered relative to that of the human evaluator, meaning that, to remain an effective training partner, the CA's expertise may need to be updated over time as the human evaluator's skills improve. Adaptation, where the AI learns the human decision-making process and updates its behavior accordingly, has shown benefits in training contexts such as UX and product design [7, 119], flight simulation [19], higher education [100], and navigation tasks [138]. In usability analysis, AI could similarly adapt to evaluators' evolving skills by adjusting its recommendations based on real-time signals like user hesitation, time spent on tasks, or patterns in the feedback process. This adaptation could enhance the AI's ability to provide more tailored support as evaluators' expertise develops.

7.3 Limitations and Future Work

In this study, we simulated novice and experienced CAs to explore how perceived UX expertise influences evaluators' analysis behaviors. While our prompts were grounded in general UX knowledge and practices, they may not fully reflect the diversity of individual evaluator styles and capabilities. Although the novice CA showed potential as a training aid, further research is needed to determine whether such interactions lead to measurable improvements in usability analysis skills. While we aimed to reduce confounds related to tone or semantics by focusing on task-related expertise and enforcing a strict response format, prior work shows that a CA's personality traits, such as sociability, enthusiasm, or critical tone, can influence user perceptions and behavior [61, 89, 95]. Future studies could investigate how personality design choices impact evaluator engagement and collaboration, further refining the role of CAs in usability analysis.

Participants also interacted with a single CA per video, which differs from real-world settings where multiple evaluators with diverse expertise could collaborate on the same project. Expanding on previous research of chatbots facilitating consensus-building in co-design [108], future studies could investigate environments where multiple CAs, each with different areas of expertise, collaborate on the same video. This would provide insights into how mixed-expertise teams interact, negotiate findings, and influence each other's analysis strategies.

Lastly, this study involved a limited number of participants. The gender distribution (10 females and 2 males) is roughly consistent with the gender breakdown in crowdsourced data on UX researchers (76.3% female, 20.3% male) [91], though potential gender bias remains a consideration. Individual differences in participants' prior AI experience may also have shaped their behaviors during the study. Future research with larger, more diverse samples could help validate these findings and determine whether similar longitudinal patterns emerge.

8 Conclusion

This study explored how evaluator behaviors evolve when collaborating with AI in usability analysis. Through a multi-session, within-subjects study, we found that participants adapted their analysis strategies over time and leveraged novice and experienced CAs in distinct ways, leading to effective analyses and a clear division of work between humans and AI. The presence of CA assistance significantly outperformed no assistance, boosting both analysis efficiency and evaluator confidence. The experienced CA was preferred, increasing the reliability of results without sacrificing diversity. Ultimately, our study highlights the potential of AI in improving usability evaluations and fostering more efficient and effective human-AI partnerships over time.

References

[1] Ighoyota Ben. Ajenaghughrure, Sonia C. Sousa, Ilkka Johannes Kosunen, and David Lamas. 2019. Predictive model to assess user trust: a psycho-physiological approach. In *Proceedings of the 10th Indian Conference on Human-Computer Interaction (IndiaHCI '19)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3364183.3364195

[2] AltexSoft. 2019. UX Researcher: Methods, Skills and Process. https://www.altexsoft.com/blog/ux-researcher-methods-skills-process/

[3] Theo Araujo and Nadine Bol. 2024. From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. *Computers in Human Behavior: Artificial Humans* 2, 1 (Jan. 2024), 100030. https://doi.org/10.1016/j.chbah.2023.100030

[4] Andrea Batch, Yipeng Ji, Mingming Fan, Jian Zhao, and Niklas Elmqvist. 2023. uxSense: Supporting User Experience Analysis with Visualization and Computer Vision. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–15. https://doi.org/10.1109/TVCG.2023.3241581 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[5] Patricia Benner. 1982. From Novice To Expert. *AJN The American Journal of Nursing* 82, 3 (1982), 402–407. https://journals.lww.com/ajnonline/citation/1982/82030/from_novice_to_expert.4.a

[6] 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

[7] Félix Buendía-García and Javier Piris-Ruano. 2025. Using Generative AI to Support UX Design Students in Web Development Courses. *Applied Sciences* 15, 13 (June 2025), 7389. https://doi.org/10.3390/app15137389 Publisher: Multidisciplinary Digital Publishing Institute.

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[9] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 188:1–188:21. https://doi.org/10.1145/3449287

[10] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. SAGE. Google-Books-ID: v1qP1KbXz1AC.

[11] Catherine C. Chase, Doris B. Chin, Marily A. Oppezzo, and Daniel L. Schwartz. 2009. Teachable Agents and the Protégé Effect: Increasing the Effort Towards Learning. *Journal of Science Education and Technology* 18, 4 (Aug. 2009), 334–352. https://doi.org/10.1007/s10956-009-9180-4

[12] Hsi-Jen Chen, Yan-Ting Chen, and Chia-Han Yang. 2022. Behaviors of Novice and Expert Designers in the Design Process: From Discovery to Design. *International Journal of Design* 16, 3 (2022), 59–76. https://doi.org/10.57698/v16i3.04

[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] Dorothee Clasen and Marc Hassenzahl. 2024. Fostering people's autonomy by foregrounding and questioning daily choices. In *Adjunct Proceedings of the 2024 Nordic Conference on Human-Computer Interaction (NordiCHI '24 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3677045.3685442

[15] Andy Cockburn, Carl Gutwin, Joey Scarr, and Sylvain Malacria. 2014. Supporting Novice to Expert Transitions in User Interfaces. *ACM Comput. Surv.* 47, 2 (Nov. 2014), 31:1–31:36. https://doi.org/10.1145/2659796

[16] Tyler Corwin, Mehmet Kosa, Mahsa Nasri, Christoffer Holmgård, and Casper Harteveld. 2023. The Teaching Efficacy of the Protégé Effect in Gamified Education. In *2023 IEEE Conference on Games (CoG)*. 1–8. https://doi.org/10.1109/CoG57401.2023.10333166 ISSN: 2325-4289.

[17] Emmelyn A. J. Croes and Marjolijn L. Antheunis. 2021. Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot - Emmelyn A. J. Croes, Marjolijn L. Antheunis, 2021. *Journal of Social and Personal Relationships* 38, 1 (2021), 279–300. https://doi.org/10.1177/0265407520959463

[18] Fleur Deken, Maaike Kleinsmann, Marco Aurisicchio, Kristina Lauche, and Rob Bracewell. 2011. Tapping into past design experiences: knowledge sharing and creation during novice–expert design consultations. *Research in Engineering Design* 23, 3 (Oct. 2011), 203–218. https://doi.org/10.1007/s00163-011-0123-8 Number: 3 Publisher: Springer.

[19] Jean-François Delisle. 2022. *Intelligent Adaptive Flight Training System- A Human Performance in the Loop for Real-Time Decision Making*. phd. Polytechnique Montréal. https://publications.polymtl.ca/10551/

[20] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126. https://doi.org/10.1037/xge0000033 Place: US Publisher: American Psychological Association.

[21] Shiying Ding, Xinyi Chen, Yan Fang, Wenrui Liu, Yiwu Qiu, and Chunlei Chai. 2023. DesignGPT: Multi-Agent Collaboration in Design. In *2023 16th International Symposium on Computational Intelligence and Design (ISCID)*. 204–208. https://doi.org/10.1109/ISCID59865.2023.00056 ISSN: 2473-3547.

[22] Peitong Duan, Jeremy Warner, Yang Li, and Bjoern Hartmann. 2024. Generating Automatic Feedback on UI Mockups with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3613904.3642782

[23] 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

[24] 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.

[25] 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

[26] 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

[27] James Fisher. 1991. Defining the novice user. *Behaviour & Information Technology* 10, 5 (1991), 437–441. https://doi.org/10.1080/01449299108924301 Place: United Kingdom Publisher: Taylor & Francis.

[28] 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

[29] 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

[30] Eureka Foong, Darren Gergle, and Elizabeth M. Gerber. 2017. Novice and Expert Sensemaking of Crowdsourced Design Feedback. *Proc. ACM Hum.-Comput.*

[31] Jens Gerken. 2010. *Longitudinal Research in Human-Computer Interaction*. Ph. D. Dissertation. Universität Konstanz, Konstanz, Germany. https://kops.uni-konstanz.de/entities/publication/ef10c26c-3f4c-41a9-8045-da53586fc360

[32] Kostis Giannakopoulos, Argyro Kavadella, Anas Aaqel Salim, Vassilis Stamatopoulos, and Eleftherios G. Kaklamanos. 2023. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. *Journal of Medical Internet Research* 25, 1 (Dec. 2023), e51580. https://doi.org/10.2196/51580 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[33] 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.

[34] Siddharth Gulati, Joe McDonagh, Sonia Sousa, and David Lamas. 2024. Trust models and theories in human–computer interaction: A systematic literature review. *Computers in Human Behavior Reports* 16 (Dec. 2024), 100495. https://doi.org/10.1016/j.chbr.2024.100495

[35] Kasper Hald, Katharina Weitz, Elisabeth André, and Matthias Rehm. 2021. "An Error Occurred!" - Trust Repair With Virtual Robot Using Levels of Mistake Explanation. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21)*. Association for Computing Machinery, New York, NY, USA, 218–226. https://doi.org/10.1145/3472307.3484170

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

[37] 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

[38] 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

[39] Annabell Ho, Jeff Hancock, and Adam S Miner. 2018. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication* 68, 4 (Aug. 2018), 712–733. https://doi.org/10.1093/joc/jqy026

[40] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. 2013. Trust in Automation. *IEEE Intelligent Systems* 28, 1 (Jan. 2013), 84–88. https://doi.org/10.1109/MIS.2013.24

[41] Jane Hogan, Gary Grant, Fiona Kelly, and Jennie O'Hare. 2020. Factors influencing acceptance of robotics in hospital pharmacy: a longitudinal study using the Extended Technology Acceptance Model. *International Journal of Pharmacy Practice* 28, 5 (Oct. 2020), 483–490. https://doi.org/10.1111/ijpp.12637

[42] Ironhack. 2023. The Evolving Role of a UX/UI Designer: Trends and Skills for Success. https://www.ironhack.com/gb/blog/the-evolving-role-of-a-ux-ui-designer-trends-and-skills-for-success

[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] Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C. P. Snijders. 2024. Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions. *ACM Trans. Interact. Intell. Syst.* (Aug. 2024). https://doi.org/10.1145/3686164 Just Accepted.

[45] Mahmut Kalman. 2019. "It Requires Interest, Time, Patience and Struggle": Novice Researchers' Perspectives on and Experiences of the Qualitative Research Journey. *Qualitative Research in Education* 8, 3 (Oct. 2019), 341–377. https://doi.org/10.17583/qre.2019.4483 Number: 3.

[46] Takayuki Kanda, Rumi Sato, Naoki Saiwaki, and Hiroshi Ishiguro. 2007. A Two-Month Field Trial in an Elementary School for Long-Term Human–Robot Interaction. *IEEE Transactions on Robotics* 23, 5 (Oct. 2007), 962–971. https://doi.org/10.1109/TRO.2007.904904 Conference Name: IEEE Transactions on Robotics.

[47] Evangelos Karapanos, Jhilmil Jain, and Marc Hassenzahl. 2012. Theories, methods and case studies of longitudinal HCI research. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems (CHI EA '12)*. Association for Computing Machinery, New York, NY, USA, 2727–2730. https://doi.org/10.1145/2212776.2212706

[48] Abdolvahab Khademi. 2023. Can ChatGPT and Bard Generate Aligned Assessment Items? A Reliability Analysis against Human Performance. *Journal of Applied Learning & Teaching* 6, 1 (May 2023). https://doi.org/10.37074/jalt.2023.6.1.28 arXiv:2304.05372 [cs].

[49] Ahmad Khawaji, Jianlong Zhou, Fang Chen, and Nadine Marcus. 2015. Using Galvanic Skin Response (GSR) to Measure Trust and Cognitive Load in the Text-Chat Environment. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. Association for Computing Machinery, New York, NY, USA, 1989–1994. https://doi.org/10.1145/2702613.2732766

[50] Jay Kim. 2020. Rubber Ducking- What It Is and Why It Works. https://medium.com/@jkarma0920/rubber-ducking-what-it-is-and-why-it-works-5d026fd9ae58

[51] Sunyoung Kim and Abhishek Choudhury. 2021. Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study. *Computers in Human Behavior* 124 (Nov. 2021), 106914. https://doi.org/10.1016/j.chb.2021.106914

[52] Maria Kjærup, Mikael B. Skov, Peter Axel Nielsen, Jesper Kjeldskov, Jens Gerken, and Harald Reiterer. 2021. Longitudinal Studies in HCI Research: A Review of CHI Publications From 1982–2019. In *Advances in Longitudinal HCI Research*, Evangelos Karapanos, Jens Gerken, Jesper Kjeldskov, and Mikael B. Skov (Eds.). Springer International Publishing, Cham, 11–39. https://doi.org/10.1007/978-3-030-67322-2_2

[53] Eva Krapp, Robin Neuhaus, Marc Hassenzahl, and Matthias Laschke. 2024. In a Quasi-Social Relationship With ChatGPT. An Autoethnography on Engaging With Prompt-Engineered LLM Personas. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction (NordiCHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3679318.3685501

[54] Emily Kuang. 2025. Evaluating Usability Challenges in VR Games for Older Adults: A Comparison With and Without AI Assistance. In *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '25)*. Association for Computing Machinery, New York, NY, USA, Article 92, 3 pages. https://doi.org/10.1145/3746058.3758343

[55] 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

[56] 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

[57] 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)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3613904.3642168

[58] 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. https://doi.org/10.48550/arXiv.2112.11471 arXiv:2112.11471 [cs].

[59] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (March 1977), 159. https://doi.org/10.2307/2529310

[60] Claire Lauer, Danielle Storey, and Romit Soley. 2024. Vector Personas: How UX Researchers Can Use AI to Bring New Dimension to Traditional Persona Development. In *Proceedings of the 42nd ACM International Conference on Design of Communication (SIGDOC '24)*. Association for Computing Machinery, New York, NY, USA, 151–157. https://doi.org/10.1145/3641237.3691663

[61] Nadine Lessio and Alexis Morris. 2020. Toward Design Archetypes for Conversational Agent Personality. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 3221–3228. https://doi.org/10.1109/SMC42975.2020.9283254 ISSN: 2577-1655.

[62] Tobias Lieberei, Virginia Deborah Elaine Welter, Leroy Großmann, and Moritz Krell. 2023. Findings from the expert-novice paradigm on differential response behavior among multiple-choice items of a pedagogical content knowledge test – implications for test development. 14 (2023). https://doi.org/10.3389/fpsyg.2023.1240120

[63] Yiren Liu, Pranav Sharma, Mehul Oswal, Haijun Xia, and Yun Huang. 2025. PersonaFlow: Designing LLM-Simulated Expert Perspectives for Enhanced Research Ideation. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 506–534. https://doi.org/10.1145/3715336.3735789

[64] Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Xing Che, Dandan Wang, and Qing Wang. 2023. Chatting with GPT-3 for Zero-Shot Human-Like Mobile Automated GUI Testing. https://doi.org/10.48550/arXiv.2305.09434 arXiv:2305.09434 [cs].

[65] Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Xing Che, Dandan Wang, and Qing Wang. 2024. Make LLM a Testing Expert: Bringing Human-like Interaction to Mobile GUI Testing via Functionality-aware Decisions. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (ICSE '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3597503.3639180

[66] Tao Long, Katy Ilonka Gero, and Lydia B Chilton. 2024. Not Just Novelty: A Longitudinal Study on Utility and Customization of an AI Workflow. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. Association for Computing Machinery, New York, NY, USA, 782–803. https://doi.org/10.1145/3643834.3661587

[67] Yuwen Lu, Yuewen Yang, Qinyi Zhao, Chengzhi Zhang, and Toby Jia-Jun Li. 2024. AI Assistance for UX: A Literature Review Through Human-Centered AI. https://doi.org/10.48550/arXiv.2402.06089 arXiv:2402.06089 [cs].

[68] 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

[69] Kim Man Lui and Keith C. C. Chan. 2006. Pair programming productivity: Novice–novice vs. expert–expert. *International Journal of Human-Computer Studies* 64, 9 (Sept. 2006), 915–925. https://doi.org/10.1016/j.ijhcs.2006.04.010

[70] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3706598.3713423

[71] Ellie Martin. 2016. Why 5 is the magic number for UX usability testing | Inside Design Blog. https://www.invisionapp.com/inside-design/ux-usability-research-testing/

[72] Melanie J. McGrath, Oliver Lack, James Tisch, and Andreas Duenser. 2025. Measuring trust in artificial intelligence: validation of an established scale and its short form. *Frontiers in Artificial Intelligence* 8 (May 2025). https://doi.org/10.3389/frai.2025.1582880 Publisher: Frontiers.

[73] Valerie Mendoza and David G. Novick. 2005. Usability over time. In *Proceedings of the 23rd annual international conference on Design of communication: documenting & designing for pervasive information (SIGDOC '05)*. Association for Computing Machinery, New York, NY, USA, 151–158. https://doi.org/10.1145/1085313.1085348

[74] Kate Moran. 2019. Usability Testing 101. https://www.nngroup.com/articles/usability-testing-101/

[75] Adity U. Mutsuddi and Kay Connelly. 2012. Text messages for encouraging physical activity Are they effective after the novelty effect wears off?. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. 33–40. https://doi.org/10.4108/icst.pervasivehealth.2012.248715 ISSN: 2153-1641.

[76] Harold G. Nelson and Erik Stolterman. 2012. The Ultimate Particular. In *The Design Way: Intentional Change in an Unpredictable World*. MIT Press, 27–40. https://ieeexplore.ieee.org/document/6354147 Conference Name: The Design Way: Intentional Change in an Unpredictable World.

[77] Jakob Nielsen. 1994. Heuristic Evaluation. In *Usability Inspection Methods*. John Wiley & Sons, Ltd, New York, NY.

[78] Jakob Nielsen. 2000. Why You Only Need to Test with 5 Users. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

[79] Jakob Nielsen. 2012. Usability 101: Introduction to Usability. https://www.nngroup.com/articles/usability-101-introduction-to-usability/.

[80] Jakob Nielsen. 2024. What AI Can and Cannot Do for UX. https://www.uxtigers.com/post/ai-can-cannot-do-ux

[81] Jakob Nielsen and Robert L. Mack (Eds.). 1994. *Usability inspection methods*. John Wiley & Sons, Inc.

[82] 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

[83] Don Norman and Jakob Neilsen. 1998. The Definition of User Experience (UX). https://www.nngroup.com/articles/definition-user-experience/

[84] OpenAI. 2023. Introducing GPTs. https://openai.com/blog/introducing-gpts

[85] OpenAI. 2024. Creating a GPT | OpenAI Help Center. https://help.openai.com/en/articles/8554397-creating-a-gpt

[86] OpenAI. 2024. OpenAI Platform. https://platform.openai.com

[87] OpenAI. 2024. Prompt engineering. https://platform.openai.com/docs/guides/prompt-engineering

[88] 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

[89] Debajyoti Pal, Vajirasak Vanijja, Himanshu Thapliyal, and Xiangmin Zhang. 2023. What affects the usage of artificial conversational agents? An agent personality and love theory perspective. *Computers in Human Behavior* 145 (Aug. 2023), 107788. https://doi.org/10.1016/j.chb.2023.107788

[90] 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.

[91] Payscale. 2024. UX Researcher Salary in 2024 | PayScale. https://www.payscale.com/research/US/Job=UX_Researcher/Salary

[92] Iryna Pentina, Tianling Xie, Tyler Hancock, and Ainsworth Bailey. 2023. Consumer–machine relationships in the age of artificial intelligence: Systematic literature review and research directions. *Psychology & Marketing* 40, 8 (2023), 1593–1614. https://doi.org/10.1002/mar.21853 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.21853.

[93] Adam M. Persky and Jennifer D. Robinson. 2017. Moving from Novice to Expertise and Its Implications for Instruction. *American Journal of Pharmaceutical Education* 81, 9 (Nov. 2017), 6065. https://doi.org/10.5688/ajpe6065

[94] Robert E. Ployhart and Robert J. Vandenberg. 2010. Longitudinal research: The theory, design, and analysis of change. *Journal of Management* 36, 1 (2010), 94–120. https://doi.org/10.1177/0149206309352110 Place: US Publisher: Sage Publications.

[95] Alisha Pradhan and Amanda Lazar. 2021. Hey Google, Do You Have a Personality? Designing Personality and Personas for Conversational Agents. In *Proceedings of the 3rd Conference on Conversational User Interfaces (CUI '21)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3469595.3469607

[96] Luiz Rodrigues, Filipe D. Pereira, Armando M. Toda, Paula T. Palomino, Marcela Pessoa, Leandro Silva Galvão Carvalho, David Fernandes, Elaine H. T. Oliveira, Alexandra I. Cristea, and Seiji Isotani. 2022. Gamification suffers from the novelty effect but benefits from the familiarization effect: Findings from a longitudinal study. *International Journal of Educational Technology in Higher Education* 19, 1 (Feb. 2022), 13. https://doi.org/10.1186/s41239-021-00314-6

[97] Christian P. Rohrer, James Wendt, Jeff Sauro, Frederick Boyle, and Sara Cole. 2016. Practical Usability Rating by Experts (PURE): A Pragmatic Approach for Scoring Product Usability. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 786–795. https://doi.org/10.1145/2851581.2851607

[98] Krishna Ronanki, Beatriz Cabrero-Daniel, and Christian Berger. 2024. ChatGPT as a Tool for User Story Quality Evaluation: Trustworthy Out of the Box?. In *Agile Processes in Software Engineering and Extreme Programming – Workshops (Lecture Notes in Business Information Processing)*, Philippe Kruchten and Peggy Gregory (Eds.). Springer Nature Switzerland, Cham, 173–181. https://doi.org/10.1007/978-3-031-48550-3_17

[99] Sara L. Rynes, Marc O. Orlitzky, and Robert D. Bretz Jr. 1997. Experienced Hiring Versus College Recruiting: Practices and Emerging Trends. *Personnel Psychology* 50, 2 (1997), 309–339. https://doi.org/10.1111/j.1744-6570.1997.tb00909.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6570.1997.tb00910.x.

[100] Ramteja Sajja, Yusuf Sermet, Muhammed Cikmaz, David Cwiertny, and Ibrahim Demir. 2024. Artificial Intelligence-Enabled Intelligent Assistant for Personalized and Adaptive Learning in Higher Education. *Information* 15, 10 (Sept. 2024), 596. https://doi.org/10.3390/info15100596 Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[101] Jeff Sauro. 2018. How Large Is the Evaluator Effect in Usability Testing? https://measuringu.com/evaluator-effect/

[102] M. Schmidt, A. A. Tawfik, I. Jahnke, and Y. Earnshaw. 2020. Learner and User Experience Research. *Learner and User Experience Research* (2020). https://doi.org/10.59668/36 Publisher: EdTech Books.

[103] Mark A. Schmuckler. 2001. What Is Ecological Validity? A Dimensional Analysis. *Infancy* 2, 4 (2001), 419–436. https://doi.org/10.1207/S15327078IN0204_02 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/S15327078IN0204_02.

[104] Valentin Schwind, Stefan Resch, and Jessica Sehrt. 2023. The HCI User Studies Toolkit: Supporting Study Designing and Planning for Undergraduates and Novice Researchers in Human-Computer Interaction. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3544549.3585890

[105] Hyojung Julia Seo. 2021. *An Exploration of Knowledge Practices Among User Experience (UX) Professionals*. Thesis. https://tspace.library.utoronto.ca/handle/1807/108705 Accepted: 2021-11-30T16:44:14Z.

[106] Luyao Shen, Qing Shi, Emily Kuang, Linjie Qiu, Shixu Zhou, Pan Hui, and Mingming Fan. 2026. MultiUX: A Human-AI Collaborative Tool to Facilitate Multiple Usability Test Video Analyses. *International Journal of Human–Computer Interaction* 0, 0 (2026), 1–26. https://doi.org/10.1080/10447318.2025.2605526

[107] Grace Shin, Yuanyuan Feng, Mohammad Hossein Jarrahi, and Nicci Gafinowitz. 2019. Beyond novelty effect: a mixed-methods exploration into the motivation for long-term activity tracker use. *JAMIA Open* 2, 1 (April 2019), 62–72. https://doi.org/10.1093/jamiaopen/ooy048

[108] Joongi Shin, Michael A. Hedderich, AndréS Lucero, and Antti Oulasvirta. 2022. Chatbots Facilitating Consensus-Building in Asynchronous Co-Design. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3526113.3545671

[109] Joongi Shin, Michael A. Hedderich, Bartłomiej Jakub Rey, Andrés Lucero, and Antti Oulasvirta. 2024. Understanding Human-AI Workflows for Generating Personas. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. Association for Computing Machinery, New York, NY, USA, 757–781. https://doi.org/10.1145/3643834.3660729

[110] Subin Shin, Jeesun Oh, and Sangwon Lee. 2025. Can LLMs See What I See? A Study on Five Prompt Engineering Techniques for Evaluating UX on a Shopping Site. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3706599.3720079

[111] Keng Siau and Weiyu Wang. 2018. Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53. https://ink.library.smu.edu.sg/sis_research/9371

[112] Auste Simkute, Ewa Luger, Michael Evans, and Rhianne Jones. 2024. "It is there, and you need it, so why do you not use it?" Achieving better adoption of AI systems by domain experts, in the case study of natural science research. https://doi.org/10.48550/arXiv.2403.16895 arXiv:2403.16895 [cs].

[113] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581075

[114] Thomas J. Smith and Cindy Kheng. 2021. Reliability of Heuristic Evaluation During Usability Analysis. https://doi.org/10.1007/978-3-030-74614-8_88

[115] 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

[116] JaYoung Sung, Henrik I. Christensen, and Rebecca E. Grinter. 2009. Robots in the wild: Understanding long-term use. In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 45–52. https://ieeexplore.ieee.org/document/6256092 ISSN: 2167-2148.

[117] Fumihide Tanaka, Javier R. Movellan, Bret Fortenberry, and Kazuki Aisaka. 2006. Daily HRI evaluation at a classroom environment: reports from dance interaction experiments. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction (HRI '06)*. Association for Computing Machinery, New York, NY, USA, 3–9. https://doi.org/10.1145/1121241.1121245

[118] Armin Tanovic. 2025. AI Thematic Analysis + the Best Tools for the Job. https://maze.co/collections/ai/thematic-analysis/

[119] Sirui Tao, Ivan Liang, Cindy Peng, Zhiqing Wang, Srishti Palani, and Steven P. Dow. 2025. DesignWeaver: Dimensional Scaffolding for Text-to-Image Product Design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–26. https://doi.org/10.1145/3706598.3714211

[120] Phuoc Thai, Fernando Montalvo, Promise Stephens, Jordan Sasser, and Sean Hinkle. 2025. Smarter UX Evaluations? Comparing AI and Human Experts in Usability Analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 69, 1 (Sept. 2025), 1219–1222. https://doi.org/10.1177/10711813251369790 Publisher: SAGE Publications Inc.

[121] Jonathan G. Tullis and Brennen Feder. 2022. The "curse of knowledge" when predicting others' knowledge. *Memory & Cognition* 51, 5 (Dec. 2022), 1214–1234. https://doi.org/10.3758/s13421-022-01382-3 Publisher: Springer.

[122] Janne Tyni, Aatu Turunen, Juho Kahila, Roman Bednarik, and Matti Tedre. 2024. Can ChatGPT Match the Experts? A Feedback Comparison for Serious Game Development. *International Journal of Serious Games* 11, 2 (June 2024), 87–106. https://doi.org/10.17083/ijsg.v11i2.744

[123] UserTesting. 2020. UserTesting: How to Analyze and Share Results. https://help.usertesting.com/hc/en-us/articles/115003371891-How-to-analyze-and-share-results.

[124] UserTesting. 2021. UserTesting: The Human Insight Platform. https://www.usertesting.com/.

[125] UXTesting. 2022. *UXTesting: The Best and Most Intelligent User Experience Solution for Enterprises*. https://www.uxtesting.io/user-experience-program

[126] Zombor Varnagy-Toth. 2022. 2021 — In-demand skills for UX Researchers. https://uxplanet.org/2021-in-demand-skills-for-ux-researchers-f997753a9f8

[127] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023), 129:1–129:38. https://doi.org/10.1145/3579605

[128] Dana Vertsberger, Navot Naor, and Mirène Winsberg. 2022. Adolescents' Well-being While Using a Mobile Artificial Intelligence–Powered Acceptance Commitment Therapy Tool: Evidence From a Longitudinal Study. *JMIR AI 2022;1(1):e38171 https://ai.jmir.org/2022/1/e38171* (Nov. 2022). https://doi.org/10.2196/38171 Publisher: JMIR AI.

[129] Karel Vredenburg, Ji-Ye Mao, Paul W. Smith, and Tom Carey. 2002. A Survey of User-Centered Design Practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*. Association for Computing Machinery, New York, NY, USA, 471–478. https://doi.org/10.1145/503376.503460

[130] Wen-Fan Wang, Chien-Ting Lu, Nil Ponsa i Campanyà, Bing-Yu Chen, and Mike Y. Chen. 2025. AIdeation: Designing a Human-AI Collaborative Ideation System for Concept Designers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–28. https://doi.org/10.1145/3706598.3714148

[131] Gareth R. White, Pejman Mirza-babaei, Graham McAllister, and Judith Good. 2011. Weak inter-rater reliability in heuristic evaluation of video games. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. Association for Computing Machinery, New York, NY, USA, 1441–1446. https://doi.org/10.1145/1979742.1979788

[132] Kate Whiting. 2023. Is AI making you suffer from FOBO? Here's what can help. https://www.weforum.org/stories/2023/12/ai-fobo-jobs-anxiety/

[133] Kelvin K. L. Wong, Yong Han, Yifeng Cai, Wumin Ouyang, Hemin Du, and Chao Liu. 2025. From Trust in Automation to Trust in AI in Healthcare: A 30-Year Longitudinal Review and an Interdisciplinary Framework. *Bioengineering* 12, 10 (Oct. 2025), 1070. https://doi.org/10.3390/bioengineering12101070 Publisher: Multidisciplinary Digital Publishing Institute.

[134] Wei Xiang, Hanfei Zhu, Suqi Lou, Xinli Chen, Zhenghua Pan, Yuping Jin, Shi Chen, and Lingyun Sun. 2024. SimUser: Generating Usability Feedback by Simulating Various Users Interacting with Mobile Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3613904.3642481

[135] Cindy Xiong, Lisanne Van Weelden, and Steven Franconeri. 2020. The Curse of Knowledge in Visual Data Communication. *IEEE Transactions on Visualization and Computer Graphics* 26, 10 (Oct. 2020), 3051–3062. https://doi.org/10.1109/TVCG.2019.2917689

[136] 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* (New York, NY, USA) *(CHI '20)*. Association for Computing Machinery, 1–13. https://doi.org/10.1145/3313831.3376301

[137] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3025171.3025219

[138] Michelle Zhao, Reid Simmons, and Henny Admoni. 2022. The Role of Adaptation in Collective Human–AI Teaming. *Topics in Cognitive Science* n/a, n/a (2022). https://doi.org/10.1111/tops.12633 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12633.

[139] Zicheng Zhu, Yugin Tan, Naomi Yamashita, Yi-Chieh Lee, and Renwen Zhang. 2025. The Benefits of Prosociality towards AI Agents: Examining the Effects of Helping AI Agents on Human Well-Being. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3706598.3713116

[140] 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: Prompts used to generate automatic suggestions and respond to questions

| Task	Prompt
Generate automatic suggestions	This is the transcript of a usability test where a participant used the think-aloud protocol to complete the following tasks: *[insert video tasks from Table 6]* Transcript: *[insert transcript from video]* Based on your role as a *[novice/experienced]* UX evaluator, as well as your knowledge of the user interfaces in the *[insert product]*, please identify all the usability problems that the participant may have encountered, the timestamps when the problem occurred, the potential causes of why the problem occurred, and redesign recommendations to address the problem. Provide your response in the format: Problem description, Start time, End time, Cause of Problem, and Redesign recommendation
Respond to questions	This is the transcript of a usability test where a participant used the think-aloud protocol to complete the following tasks: *[insert video tasks from Table 6]* Transcript: *[insert transcript from video]* This is the list of usability problems, causes, and redesign recommendations that you have previously identified in this video: *[insert previously identified problem descriptions, causes, and redesign recommendations]* Respond to the following questions based on your role as a *[novice/experienced]*, as well as your knowledge of the user interfaces in the *[insert product]*. Make your responses concise.
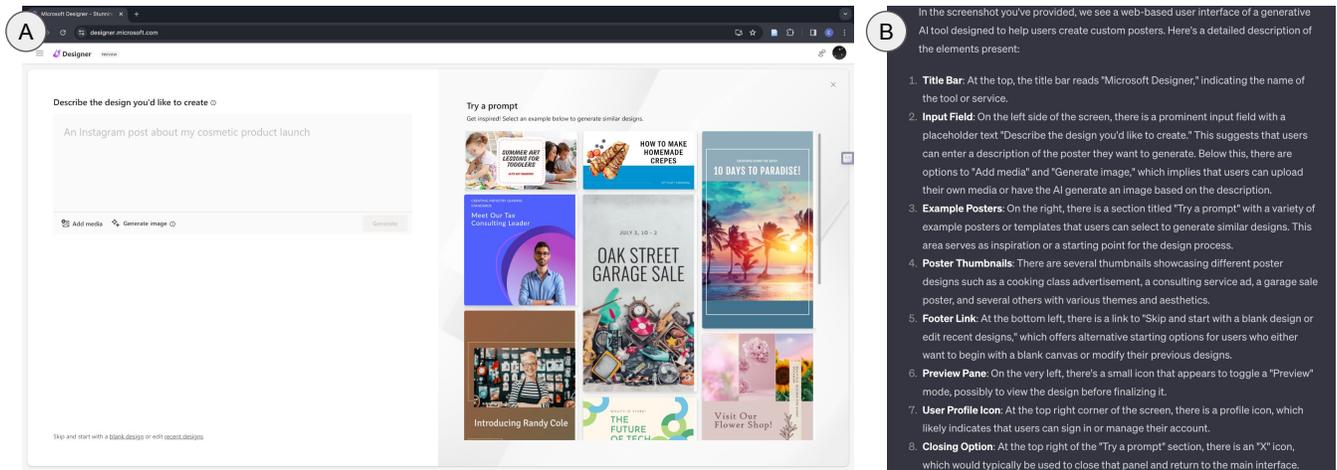
Figure 12: Screenshot of the homepage of a test website (A) and the corresponding response from GPT describing the UI elements shown in the screenshot (B).
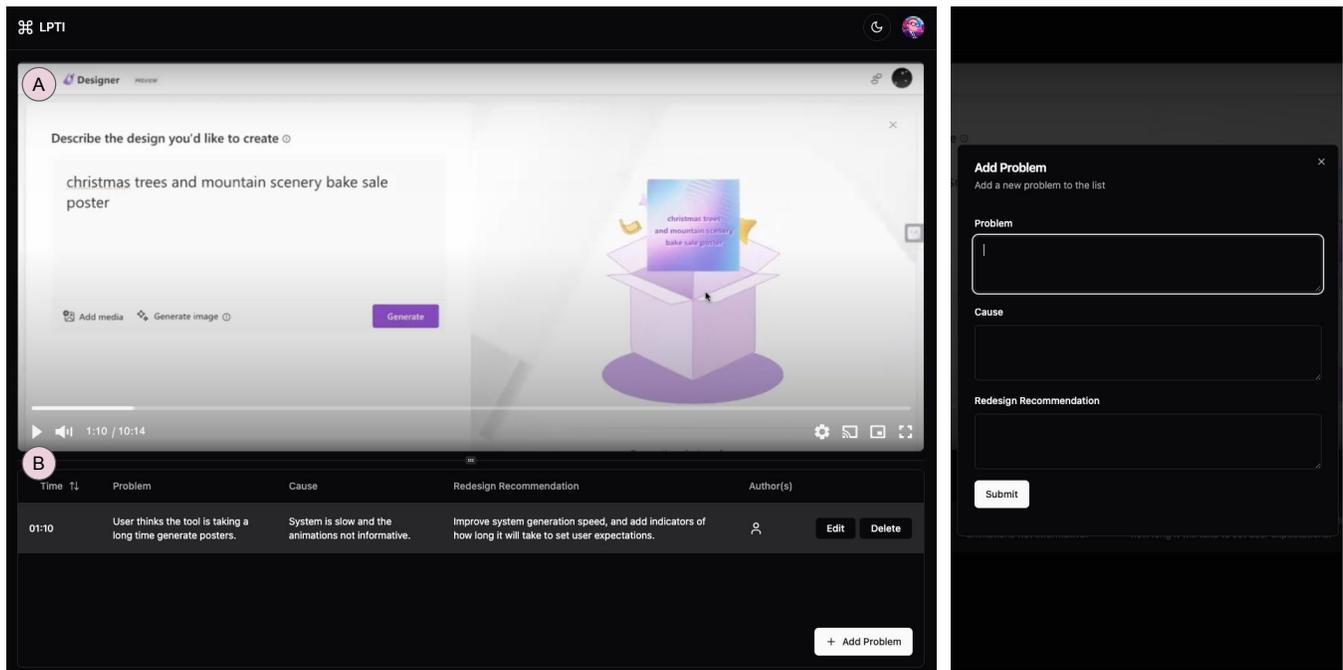

Figure 13: User interface of the UX analysis tool for the baseline condition containing: A) video player, and B) table of usability problem descriptions, causes, and redesign recommendations with the corresponding timestamp.
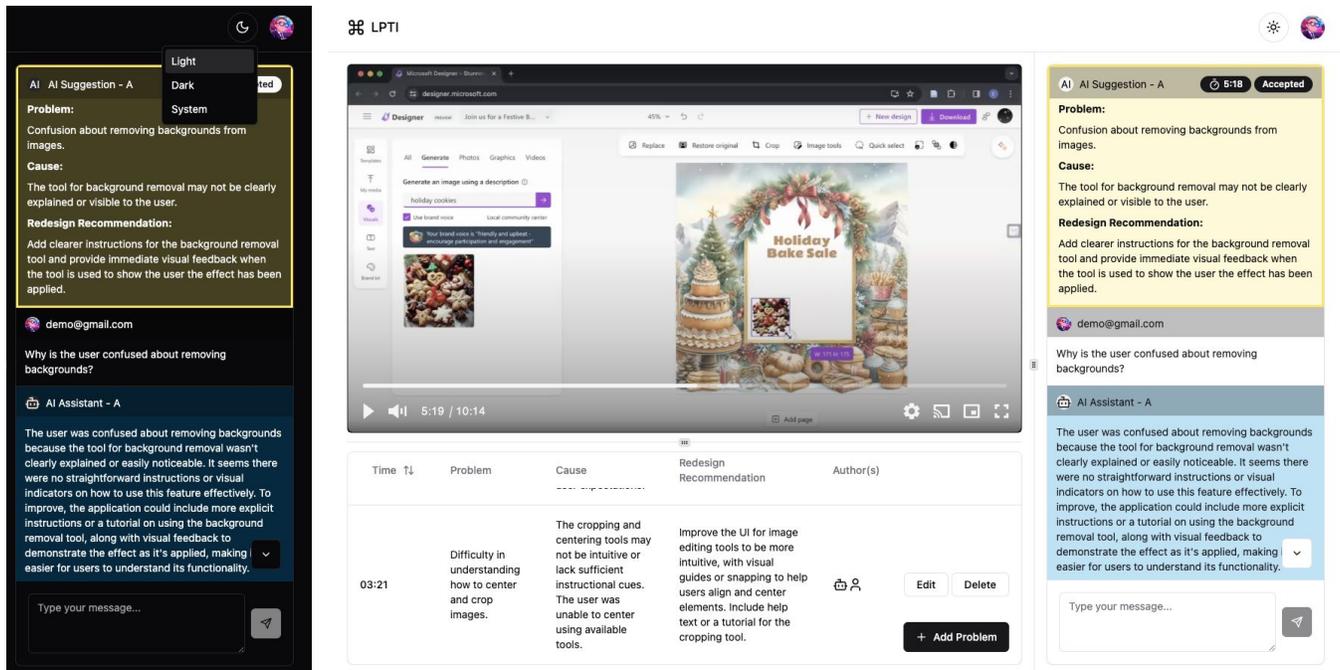
Figure 14: User interface of the UX analysis tool in light mode. Users can switch between light and dark modes by clicking on the "moon" or "sun" icon.

Table 6: Information on the 15 usability videos used in the study

Product	Tasks	Length (m:ss)	Number of Suggestions (Novice)	Number of Suggestions (Experienced)
Weather App	1. Customize the home screen so that UV is visible at the top.	7:27	4	6
	2. Add Copenhagen to the list of saved locations and change the nickname of this city to "Vacation."	14:15	6	6
		7:56	5	6
	3. Find the weather forecast for Copenhagen on Oct 16.	7:29	7	9
	4. Remove the nickname "Vacation" and reset it to the default name.	11:08	5	7
GenAI Website	1. Create a poster to advertise a fundraising bake sale in a local community center that includes the date, time, and location of the bake sale and at least 3 different visualizations.	18:08	6	8
		10:14	7	9
		13:49	6	8
	2. Export the poster, then generate a catchy caption and hashtags to post on Instagram.	17:07	7	9
		18:00	9	13
VR Game	1. Select the "Blocking" game from the menu and play it.	7:51	5	7
	2. Select the "Volleyball" game from the menu and play it.	9:42	5	8
	3. Select the "Skeet" game from the menu and play it.	10:56	7	7
		9:06	6	7
		11:27	6	6

Table 7: Categories of Messages sent by participants.

Session Appearance	Category	Frequency (%)	Example Question
All sessions	Problem identification	39 (22.9%)	"Do you think there is a usability issue here at this timestamp?"
Sessions 1 to 3	Instructions for AI	11 (6.5%)	"Just wanted to let you know you should report each and every minor and major usability issue to me."
	Task information	10 (5.9%)	"What tasks did they complete?"
	AI capabilities	7 (4.1%)	"Can you generate content for a usability problem, cause, and suggestion if I give you some basic information?"
	Product information	6 (3.5%)	"Where is the UV placed on the app?"
	Timestamp information	6 (3.5%)	"please list all the time stamps for usability problems."
	User background	2 (1.2%)	"information about the user"
Sessions 3 to 5	Problem generation	24 (14.1%)	"Write the problem, cause, redesign recommendation for users not knowing how to reload in the tutorial because they didn't see the demo first"
	Problem prioritization	9 (5.3%)	"Prioritize the changes that we need to make to the app design to fix all of the usability issues found in this video"
	Suggestion correction	8 (4.7%)	"It is also caused because important objects for the game are out of the user's point of view. A redesign option for this could also include having the user start facing the blocks."
Scattered throughout sessions 1 to 5	User actions	12 (7.1%)	"How long did it take the user to find the right page to rename the location?"
	Social expressions	8 (4.7%)	"cool thanks for helping"
	Suggestion clarification	7 (4.1%)	"Why did you suggest this organization is the best way?"
	User perceptions	6 (3.5%)	"why is the user struggling with the game?"
	Heuristics identification	5 (2.9%)	"Does this fall under the category of any heuristics?"
	Redesign recommendations	5 (2.9%)	"Do you think renaming the buttons is also a good redesign recommendation?"
	Product response clarification	5 (2.9%)	"How many times did the AI only give one image?"

Table 8: Results of ANOVA Tests for Quantitative Data (Significant effects are bolded)

| Variable	Session	Condition	Interaction Effect
Number of Pauses	$\mathbf{F_{4,44} = 2.7, p < .05, \eta_p^2 = 0.20}$	$F_{2,22} = 1.6, p > .05, \eta_p^2 = 0.13$	$F_{8,88} = 0.3, p > .05, \eta_p^2 = 0.02$
Acceptance Rate	$F_{4,44} = 1.3, p > .05, \eta_p^2 = 0.11$	$\mathbf{F_{1,11} = 9.9, p < .01, \eta_p^2 = 0.47}$	$F_{4,44} = 1.2, p > .05, \eta_p^2 = 0.1$
Edit Rate	$F_{4,44} = 0.7, p > .05, \eta_p^2 = 0.06$	$F_{1,11} = 1.0, p > .05, \eta_p^2 = 0.08$	$F_{4,44} = 2.5, p > .05, \eta_p^2 = 0.18$
Ignore Rate	$\mathbf{F_{4,44} = 3.1, p < .05, \eta_p^2 = 0.22}$	$\mathbf{F_{1,11} = 24.4, p < .001, \eta_p^2 = 0.69}$	$F_{4,44} = 0.8, p > .05, \eta_p^2 = 0.07$
Number of Total Problems	$\mathbf{F_{4,44} = 3.1, p < .05, \eta_p^2 = 0.22}$	$\mathbf{F_{2,22} = 13.3, p < .001, \eta_p^2 = 0.55}$	$F_{8,88} = 0.4, p > .05, \eta_p^2 = 0.04$
Number of Unique Problems	$F_{4,8} = 0.5, p > .05, \eta_p^2 = 0.21$	$\mathbf{F_{2,4} = 9.1, p < .05, \eta_p^2 = 0.82}$	$F_{8,16} = 0.8, p > .05, \eta_p^2 = 0.30$
Problem Coverage	$F_{4,8} = 0.6, p > .05, \eta_p^2 = 0.24$	$\mathbf{F_{2,4} = 9.1, p < .05, \eta_p^2 = 0.82}$	$F_{8,16} = 0.8, p > .05, \eta_p^2 = 0.28$
Fleiss' Kappa	$F_{4,8} = 0.7, p > .05, \eta_p^2 = 0.25$	$\mathbf{F_{2,4} = 9.1, p < .05, \eta_p^2 = 0.82}$	$F_{8,16} = 0.4, p > .05, \eta_p^2 = 0.18$
Perceived Efficiency	$\mathbf{F_{4,44} = 5.9, p < .01, \eta_p^2 = 0.35}$	$\mathbf{F_{1,11} = 12.8, p < .01, \eta_p^2 = 0.54}$	$F_{4,44} = 2.2, p > .05, \eta_p^2 = 0.17$
Trust	$\mathbf{F_{4,44} = 4.2, p < .01, \eta_p^2 = 0.27}$	$\mathbf{F_{1,11} = 16.7, p < .01, \eta_p^2 = 0.60}$	$F_{4,44} = 0.4, p > .05, \eta_p^2 = 0.04$
Perceived Completeness	$F_{4,44} = 1.2, p > .05, \eta_p^2 = 0.10$	$\mathbf{F_{1,11} = 14.2, p < .01, \eta_p^2 = 0.56}$	$F_{4,44} = 1.6, p > .05, \eta_p^2 = 0.12$
Confidence	$F_{4,44} = 0.48, p > .05, \eta_p^2 = 0.04$	$\mathbf{F_{1,11} = 6.0, p < .05, \eta_p^2 = 0.35}$	$F_{4,44} = 0.49, p > .05, \eta_p^2 = 0.04$
Perceived UX Experience	$\mathbf{F_{4,44} = 2.7, p < .05, \eta_p^2 = 0.20}$	$\mathbf{F_{1,11} = 24.4, p < .001, \eta_p^2 = 0.69}$	$\mathbf{F_{4,44} = 3.1, p < .05, \eta_p^2 = 0.22}$
Overall Satisfaction	N/A	$\mathbf{F_{1,11} = 12.3, p < .01, \eta_p^2 = 0.53}$	N/A
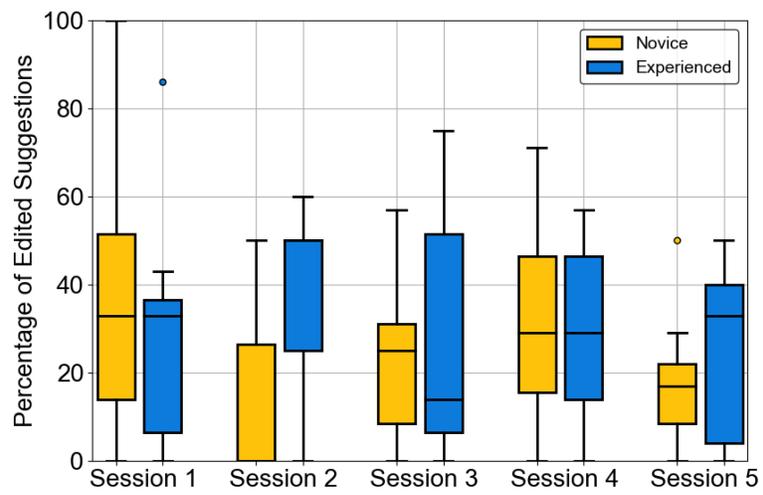
Figure 15: Box plot showing the percentage of edited suggestions separated by session and CA condition.