International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Standardizing the Evaluation of Usability Test Results: Criteria Development and Human-AI Collaborative Performance

Emily Kuang, Luyao Shen, Ehsan Jahangirzadeh Soure , Mingming Fan & Kristen Shinohara

To cite this article: Emily Kuang, Luyao Shen, Ehsan Jahangirzadeh Soure , Mingming Fan & Kristen Shinohara (15 Apr 2026): Standardizing the Evaluation of Usability Test Results: Criteria Development and Human-AI Collaborative Performance, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2026.2638554

To link to this article: https://doi.org/10.1080/10447318.2026.2638554

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2026.2638554

Standardizing the Evaluation of Usability Test Results: Criteria Development and Human-AI Collaborative Performance

Emily Kuang[a*], Luyao Shen[b*], Ehsan Jahangirzadeh Soure[c,d], Mingming Fan[b] and Kristen Shinohara[e]

[a]Electrical Engineering and Computer Science, York University, Toronto, Ontario, Canada; [b]Computational Media and Arts Thrust, The Hong Kong University of Science and Technology, Guangzhou, China; [c]School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada; [d]Snowflake, Toronto, Ontario, Canada; [e]School of Information, Rochester Institute of Technology, Rochester, NY, USA

ABSTRACT

The growing use of artificial intelligence (AI) across industries has spurred interest in its application to usability analysis. Yet, standardized criteria for evaluating AI-generated usability test results are lacking, and the impact of strategies such as role prompting and human review on the quality of results remains underexplored. To address these gaps, we first developed seven evaluation criteria grounded in user experience (UX) literature and a survey of 40 UX professionals. Second, we recruited UX experts to apply these criteria to usability findings generated under five conditions: human-only, baseline AI-only, tailored AI-only, baseline AI with human review, and tailored AI with human review. Expert evaluations show that human-AI collaboration, especially tailored AI combined with human review, produced significantly higher-quality results than either humans or AI alone. These findings provide practical methodologies and empirical evidence to support human-AI collaboration in usability analysis, informing system design for complex human–computer interaction tasks.

KEYWORDS

Artificial intelligence applications and expert systems; interface design and evaluation methodologies; user-centered design; empirical studies of user behavior

1. Introduction

As artificial intelligence (AI) technologies become increasingly integrated into user experience (UX) workflows, they present new opportunities for supporting usability evaluation, which is a cognitively demanding process that traditionally depends on human expertise and contextual reasoning (Dumas & Redish, 1999; Sauro, 2010). Traditional usability evaluation methods require UX evaluators to synthesize diverse data sources into coherent descriptions of usability problems, including their causes, effects, and potential solutions (Følstad et al., 2010). However, in practice, such analyses are often unstructured, incomplete, or lack rigor (Fan, Shi, et al., 2020; Nørgaard & Hornbæk, 2006). Practitioners must attend simultaneously to multiple information streams (eg, visual and audio cues, user actions, and note-taking) and often under significant time constraints (Chilana et al., 2010; Følstad et al., 2012; Kuang et al., 2022). As a result, many usability findings remain incomplete or underspecified.

Recent advances in generative AI (GenAI) offer promising avenues to address these limitations. AI systems can automatically identify usability problems, summarize user behaviors, and propose redesign ideas (Cheng & Zhang, 2025; Kuang et al., 2024; Z. Liu et al., 2024). Yet, simply automating these tasks is insufficient. AI systems often fail to capture the nuanced, context-dependent aspects of user experience, such as interpreting user goals, emotional responses, or situational factors, that human evaluators naturally consider (Fan, Li, et al., 2020; Grigera et al., 2017; Jeong et al., 2020). To realize the full potential of AI in usability analysis, it is essential to move beyond viewing AI as a stand-alone evaluator and instead explore how it can effectively collaborate with humans.

CONTACT Mingming Fan mingmingfan@ust.hk Computational Media and Arts Thrust, The Hong Kong University of Science and Technology, Guangzhou, China; Kristen Shinohara kristen.shinohara@rit.edu School of Information, Rochester Institute of Technology, Rochester, NY, USA

*Emily Kuang and Luyao Shen contributed equally to this research.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10447318.2026.2638554.

Over the past few years, researchers have increasingly explored how AI can augment the efficiency and completeness of usability analysis (Duan, Warner, et al., 2024; Fan et al., 2022; Kuang et al., 2023, 2024; Soure et al., 2022). For example, Duan et al. utilized GPT − 4 to conduct automatic heuristic evaluations by examining whether user interface designs violated the usability heuristics, such as Nielsen's 10 Usability Heuristics. Their findings showed that GPT − 4 achieved a precision of 60.3% and a recall of 38%. Moreover, UX evaluators rated the AI-generated suggestions as both accurate and helpful (Duan, Warner, et al., 2024). Similarly, Kuang et al. used GPT-3.5 to identify usability problems by analyzing transcripts of usability test videos. This approach achieved a precision of 86% and a recall of 71%, and UX evaluators agreed with 78.6% of the AI-generated usability problems in the user study (Kuang et al., 2024). Despite these advancements, two key research gaps remain.

First, there is a **lack of standardized criteria for evaluating AI-generated usability test results**. Prior work has largely emphasized quantitative metrics, such as the number of problems identified or the overlap between AI and human analyses (Duan, Warner, et al., 2024; Hu et al., 2025; Kuang et al., 2024; Z. Liu et al., 2024), while paying less attention to the content quality and practical relevance of AI outputs. Moreover, although human evaluators often rely on established heuristics and guidelines, AI-generated findings are rarely assessed using consistent or validated standards. Informal best practices suggest that high-quality usability findings should be accurate, actionable, specific, and clearly articulated (Hotjar, 2023; Interaction Design Foundation [IxDF], 2018), yet these dimensions are unevenly emphasized across sources. Without standardized evaluation criteria, it is difficult to meaningfully compare AI-generated findings to human analyses or to assess their utility in professional UX practice. To address this gap, we *synthesize and validate a structured set of evaluation criteria* grounded in UX literature and refined through feedback from UX professionals.

Second, there is **limited understanding of how different moderation strategies, aimed at enhancing the accuracy and relevance of AI-generated findings, affect the quality of usability test results**. Few studies have examined how AI systems can be tailored to better support human evaluators or how human input can improve AI-generated usability test results through review and refinement. Much of the existing work has relied on Wizard-of-Oz techniques (Fan et al., 2022; Kuang et al., 2023) or default, unmodified versions of ChatGPT (Kuang et al., 2024), offering limited insights into how AI capabilities might be adapted for usability tasks. Recent advances in prompt engineering suggest that large language models (LLMs) can be guided to adopt specific personas or expert roles, such as a supportive peer or critical tutor, to influence the tone and depth of their outputs (Krapp et al., 2024).

To fill these gaps, this study investigates two moderation strategies for improving AI-generated usability test results: (1) **Role prompting**, where an AI model is explicitly prompted to act as an expert UX evaluator (Shin et al., 2025), compared against a baseline AI with no persona; and (2) **Human review**, where professional UX evaluators review and refine AI-generated findings. By combining these two strategies, we derived five evaluation conditions: three *single-evaluator* settings ("baseline AI," "tailored AI," and "human only") and *two human-AI collaborative* settings ("baseline AI with human review," and "tailored AI with human review"). These conditions allow us to isolate and compare the effects of AI prompting and human moderation, both individually and in combination. We examine how each condition influences the content quality of usability test results using the validated evaluation criteria. Specifically, this study addresses the following research questions (RQs):

- **RQ1**: What criteria do UX professionals apply to evaluate the quality of usability test results, and how well do these criteria capture key dimensions of usability findings?

- **RQ2**: How do usability test results compare across five collaboration conditions—human-only, baseline AI-only, tailored AI-only, baseline AI with human review, and tailored AI with human review—when evaluated using the criteria identified in RQ1?

In summary, this work makes the following contributions:

- A validated set of criteria for evaluating the quality of usability test results, grounded in UX literature and UX professionals' input.

- Empirical evidence on the effectiveness of role prompting and human review as strategies for improving AI-assisted usability evaluation; and
- Comparative analysis demonstrating that human–AI collaboration outperforms human-only or AI-only approaches across multiple criteria for assessing usability test results.

2. Related work

2.1. From traditional usability evaluation to AI-generated results

Usability evaluation is an essential part of user-centered product development, aiming at assessing how easily and pleasantly users can achieve their goals when interacting with a product. Evaluations can be conducted without users through heuristic evaluations, or with users through user testing methods (Riihiaho, 2018). Heuristic evaluation is a method for finding usability problems in a user interface design by having a small set of evaluators examine the interface and judge its compliance with recognized usability principles, such as Nielsen's 10 Usability Heuristics (Nielsen, 1992, 2024). While heuristic evaluation remains useful, an international survey found that the most frequently employed method for identifying usability problems is through usability testing, in which one or more representative users perform tasks and verbalize their cognitive processes under observation (Fan, Shi, et al., 2020). Usability testing usually produces a report consisting of identified problems, underlying causes, and redesign recommendations (Hotjar, 2023; IxDF, 2018). As the "gold standard" in usability evaluation, the problem sets identified through usability testing are often used to benchmark other evaluation methods, such as heuristic evaluation (Landauer, 1996), highlighting the importance of generating high-quality usability results. Usability test results can be evaluated in two ways. The first involves measuring quantitative objects after implementing redesigns, such as task success rate, which requires additional development effort before assessment (Dzida & Freitag, 2001). The second focuses on assessing the quantity and quality of identified usability problems.

As GenAI becomes increasingly integrated into UX workflows, it presents new opportunities for supporting usability evaluation, which is a cognitively complex task traditionally dependent on human expertise and contextual reasoning (Dumas & Redish, 1999; Sauro, 2010). Recent work has explored three main approaches: automatically generating heuristic evaluation suggestions (Duan, Cheng, et al., 2024; Duan, Warner, et al., 2024), simulating users to provide usability feedback (Hu et al., 2025; Z. Liu et al., 2024; Xiang et al., 2024), and analyzing usability test videos to identify problems (Kuang, 2025; Kuang et al., 2023, 2024, 2026; Shen et al., 2026; Soure et al., 2022). These studies have primarily evaluated AI's problem detection capabilities through metrics such as precision and recall of GPT − 4 suggestions, human evaluator agreement rates on GPT-3.5 outputs, and thematic overlap between human and AI-identified problems. However, this emphasis on the quantity of detected problems overlooks the description of AI-generated usability results. While some prior research has gathered qualitative feedback (eg, that AI suggestions are "too vague"), these efforts have lacked systematic quality assessment frameworks. Our work addresses this gap by developing a set of standardized criteria for evaluating the quality of both human- and AI-generated usability test results, shifting the focus from detection performance to practical value and actionability in real-world UX contexts.

2.2. AI-generated content quality assessment in usability evaluation

The widespread adoption of AIGC across various domains has heightened concerns regarding quality assurance, encompassing issues that range from factual inaccuracies and hallucinations to contextual inappropriateness and lack of domain-specific relevance. Hallucinations are defined as generated content that is nonsensical, unfaithful, and undesirable (Ji et al., 2023; Leiser et al., 2024). Recent research categorizes hallucinations into three types: input-conflicting hallucination, context-conflicting hallucination, and fact-conflicting hallucination (Joshi et al., 2025; Zhang et al., 2025). Specifically, Large Vision Language Models encounter challenges with semantic ambiguity, where synthetic images lack consistency and fail to respond appropriately to text prompts (Gao et al., 2024). Large Language Models (LLMs) may exhibit self-contradictions when generating lengthy or multi-turn responses due to their limitations in maintaining long-term memory and identifying relevant context (N. F. Liu et al., 2024;

Shi et al., 2023). Moreover, ChatGPT has been observed to generate factually incorrect responses, such as incorrectly assessing that 1000 is larger than 1062 (Borji, 2023). Another study used the LIX score, a readability metric that reflects the difficulty of a given text (Björnsson, 1983), to compare legal advice generated by LLMs and human lawyers, finding that LLM outputs were harder to read. Beyond these factual inaccuracies, AI-generated responses in domain-specific applications are sometimes not incorrect, but rather inappropriate or irrelevant to the specific domain context.

In the usability evaluation domain, prior studies employing LLMs have identified similar quality issues. For example, Duan et al. used GPT − 4 to generate usability feedback and found that some outputs lacked consideration of contextual factors such as comprehensive user interfaces and common design conventions. They also noted that some redesign suggestions were vague, lacking specificity and actionability (Duan, Warner, et al., 2024). Similarly, Kuang et al. used GPT-3.5 to detect usability problems and observed that problem descriptions remained surface-level, without elaborating on underlying causes (Kuang et al., 2024). Considering that usability test results are used by designers, technical colleagues, and managers to guide design and business decisions (Scholtz, 2000), such quality issues can impede their usefulness by influencing the interpretation, prioritization, and resolution of identified problems.

Various research efforts have evaluated AIGC quality by constructing unified benchmark datasets for different scenarios and tasks (Ott et al., 2022). While this evaluation approach provides relatively fair and consistent standards for AI models, it often overlooks model performance in real-world applications, particularly from the perspective of actual user experience and perception (Peng et al., 2025; Subramonyam et al., 2025). Consequently, some studies have incorporated human evaluations to assess AI outputs. Given the complexity and specificity of professional domains, each field requires customized assessment dimensions that reflect its practical requirements. For example, in the writing domain, users across different writing scenarios have defined varying criteria for quality evaluation. In story writing, quality can be assessed based on three aspects: story arc, turning points, and affective dimensions such as arousal and valence (Tian et al., 2024), while in academic writing, users prioritized relevance, professionalism, and readability (Peng et al., 2025). When examining AIGC quality assessment in the usability evaluation domain, we found that evaluations primarily relied on participants' subjective perceptions (Duan, Warner, et al., 2024; Kuang et al., 2024; Lu et al., 2025), lacking standardized dimensions to structure the assessment process. Moreover, most studies stopped at identifying these usability issues without proposing or validating methods to mitigate them.

2.3. Moderation strategies for improving AI-generated content quality

This study focuses on improving AIGC quality through two broad categories of strategies. The first is prompt engineering, where humans refine prompts *before generation* to guide the model toward more optimized outputs. The second is human-AI collaboration, where humans edit model outputs *after generation* to improve their usefulness (Reinhard et al., 2025). These two strategies are not mutually exclusive and can be used together, but each has different strengths and limitations.

Prompts serve as inputs to LLMs, and their syntax and semantics play an important role in determining the model's output (Naveed et al., 2025). Prompt engineering involves the systematic design and optimization of prompts to guide AI systems toward producing more accurate, relevant, and contextually appropriate outputs (Ahmed et al., 2024). A recent study compared five techniques, including zero-shot prompting, role prompting, chain of thought prompting, self-refine prompting, and least-to-most prompting, for identifying usability problems in digital interfaces. The findings revealed that role prompting was highly effective in identifying problems and providing contextual analysis and actionable insights from a user perspective (Shin et al., 2025). Other domains, such as healthcare and education, have also demonstrated the effectiveness of role prompting in eliciting domain-specific capabilities (Pope et al., 2025; Sun et al., 2024). Despite its effectiveness and adaptability across various applications, prompt engineering cannot fully address challenges such as factual inaccuracies and interpretability gaps (Sahoo et al., 2025). Moreover, even with prompting engineering, LLM outputs can still be irrelevant to professional contexts, such as generating usability feedback that contradicts popular design conventions (Duan, Warner, et al., 2024). Additionally, prompt engineering poses new challenges in creating effective prompts, sometimes leading to what has been termed the "prompt loop" (Jung et al., 2026; S. Kim et al., 2025).

While prompt engineering happens before content generation, another strategy to improve AIGC quality happens *after content generation*, which is human-AI collaboration, where AI supplements human decision-making (Lai et al., 2023). AI outcomes can inform individuals through an "algorithm-in-the-loop" process, recognizing that AI systems should support rather than replace domain workers' decisions and tasks (Green & Chen, 2019; Zhang et al., 2020). Human-AI collaboration approaches, particularly those involving human editing and iterative refinement, represent another critical strategy for enhancing AIGC quality through expert knowledge integration. The human-AI collaboration approach has dominated AI-powered applications across various domains recently, including writing, education, programming, design, and accessibility (Pang et al., 2025). In recent studies within the usability analysis domain (Duan, Cheng, et al., 2024; Duan, Warner, et al., 2024; Kuang et al., 2024, 2026; Xiang et al., 2024), researchers have identified inaccurate AI outputs by asking design professionals to label the AI outputs and have proposed human editing as a moderation strategy. Human editing can improve AI outputs by integrating human judgment. It can adapt to complex tasks where AI tends to generate inaccurate or inappropriate outputs. However, this process heavily relies on user expertise. For example, due to the lack of essential knowledge for task decomposition and AI output evaluation, novices often lack the judgment needed to overcome AI hallucinations (Chen et al., 2024).

Shin et al.'s work evaluated the quantity of usability problems generated through different prompting techniques (Shin et al., 2025). However, there is a lack of evaluation of result descriptions under different conditions, which may influence how results are interpreted and applied in practice. Moreover, these conditions should be evaluated reliably employing standard criteria for comparison. In our work, we first propose standard criteria to facilitate the evaluation of usability test result descriptions, supporting systematic cross-condition comparison and a nuanced understanding of where each approach excels or underperforms. We then moderate quality issues in AI outputs using two methods: role prompting and human-AI collaboration. We refer to the original AI outputs as "baseline AI" and AI outputs after role prompting as "tailored AI." We systematically evaluate the content quality across five conditions: baseline AI, tailored AI, human, baseline AI & human, tailored AI & human. By applying comprehensive evaluation criteria to assess these moderation strategies in the usability analysis context, this study aims to provide empirical evidence for their relative effectiveness and practical utility.

3. Design of the evaluation criteria for usability test results

3.1. Method

Since there are no standardized criteria for evaluating the quality of AI-generated usability results, we developed a set of evaluation criteria by synthesizing insights from two sources: a literature review and a survey of UX professionals. Our analysis focused on identifying quality indicators related to three core components of usability testing[1]: **usability problem** description, problem **cause** identification, and **redesign recommendations** formulation.

3.1.1. Literature review

We followed the PRISMA framework (Moher et al., 2009), which outlines a four-phase procedure for conducting systematic reviews. Our goal was to identify existing criteria for evaluating the quality of usability results, particularly in the context of AI-generated content. We applied predefined inclusion criteria to guide our selection of relevant literature.

Phase 1: Identification. We identified both academic and industry sources that provided guidance on writing or structuring usability reports. Industry sources included Interaction Design, Maze, Nielsen Norman Group, and UXBoost, while academic sources included venues such as the *Journal of Usability Studies*, *Communication Design Quarterly*, and *CHI*. To ensure relevance to AI-generated content, we also included papers proposing evaluation metrics for AI-generated text in adjacent domains, such as education (eg Huang et al., 2025; Tan et al., 2025). Our inclusion criteria were as follows:

- **Topic**: Focused on the quality of usability results or AI-generated content.
- **Publication type**: Scholarly articles, industry reports, or books.

- **Date range**: Published between 2005 and 2025.
- **Language**: English.

An example search query used in the ACM Digital Library, which yielded 317 results, is shown below:

> Title:("usability analysis" OR "usability evaluation" OR "usability report" OR "usability quality" OR "usability criteria" OR "GenAI quality" OR "AI-generated content quality" OR "LLM quality" OR "evaluating AI-generated content" OR "evaluating LLM output") AND Abstract:("usability analysis" OR "usability evaluation" OR "usability report" OR "usability quality" OR "usability criteria" OR "GenAI quality" OR "AI-generated content quality" OR "LLM quality" OR "evaluating AI-generated content" OR "evaluating LLM output")
>
> E-Publication Date: (01/01/2005 TO 2/28/2025)

After supplementing this set with relevant industry articles, we identified a total of 329 sources in this phase.

Phase 2: Screening. We screened the titles and abstracts of academic papers and reviewed the full text of industry sources, applying the inclusion criteria outlined above. Papers were excluded if they did not explicitly discuss criteria for evaluating usability results or the quality of AI-generated content. For instance, most scholarly articles that contained "usability evaluations" in their title or abstract turned out to be application-specific case studies that lacked discussions of structured reporting practices or evaluation guidelines. After this screening, 304 sources were excluded, leaving 25 for further eligibility review.

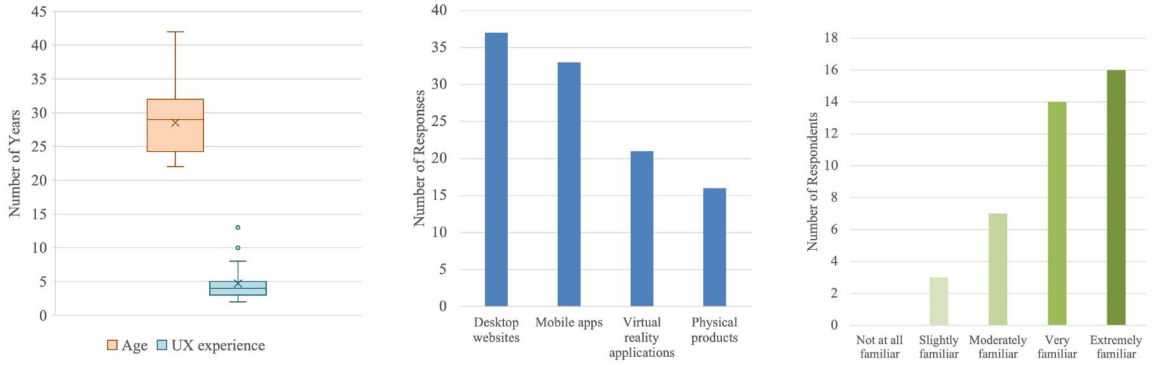
Phase 3: Eligibility. We conducted a full-text review of the remaining 25 sources. Some papers were excluded at this stage due to not being full articles or focusing on quantitative evaluations of AI-generated content (eg, performance scores based on large datasets). These papers did not include qualitative criteria for assessing content quality. As a result, we excluded 8 additional sources, leaving 17 sources for final analysis.

Phase 4: Data Set and Coding Process. The first two authors independently reviewed and open-coded the evaluation criteria described in the remaining sources. Emergent codes included concepts such as "clear problem statements," "actionable recommendations," and "justified reasoning." Through iterative discussion and consolidation, we synthesized a final set of evaluation criteria that would later serve as the basis for analyzing AI- and human-generated usability results in our study.

3.1.2. Survey of UX professionals

We conducted a survey to gather input from UX professionals on the criteria they use in their daily practice for assessing the quality of usability problems, their underlying causes, and redesign recommendations. The survey also collected demographic information, including years of UX experience, familiarity with usability analysis, and the types of products evaluated. The detailed list of survey questions is in Supplementary Appendix A.1. We distributed the survey via professional platforms such as LinkedIn and through snowball sampling within UX professionals' networks. Additionally, the survey served as a recruitment tool for participants in our subsequent user study, which aimed to evaluate the synthesized list of criteria.

In total, 48 UX professionals participated in the survey, with each participant providing one response. After conducting a quality check, we discarded several responses due to: selecting "not familiar at all" with usability analysis (N = 2), leaving criteria questions blank (N = 1), and evidence of AI-generated content (N = 5). To identify AI-generated responses, we applied a consistent screening protocol: (1) we examined linguistic patterns for excessive verbosity in open-ended responses (eg, excluded responses averaged 67.2 words, compared to 13.7 words for valid responses); (2) we flagged formatting artifacts inconsistent with the survey input fields (eg, Markdown syntax such as **bold**); and (3) we cross-checked survey completion times for irregularities (eg, excluded responses averaged 1.9 min, while valid responses averaged 6.7 min). Responses meeting any of these criteria were reviewed by two authors and excluded upon agreement. This process resulted in 40 valid responses for analysis. Figure 1 shows the demographic information of the survey respondents. Respondents were 27.5 years old on average (SD = 4.8), and had 4.7 years of UX experience (SD = 2.4). The majority had experience analyzing desktop websites (N = 37),



(a) Box plot showing the distribution of the age and years of UX experience of survey respondents.

(b) Bar chart showing responses to "What types of products have you evaluated for usability? (Select all that apply)?"

(c) Bar chart showing responses to "How familiar are you with identifying usability problems from usability test videos?"

Figure 1. Demographic information of survey respondents.

Table 1. Guidelines and assessment criteria for usability results.

Component	Findings from literature review	Findings from screening survey
Usability problem	- Accurate and consistent with user feedback (Friess, 2011, 2012; Xtensio, 2023); - Specific and clearly describes user actions (Huang et al., 2025; IxDF, 2018)	- Accurate and objective, presenting factual evidence ($N = 25$) - Specific, providing sufficient details ($N = 26$) - Relevant impact on UX ($N = 11$) - Includes established heuristics ($N = 2$)
Cause of the problem	- Accurate and specific, identifying the exact design element, flow, or interaction that caused the problem (Schade, 2013; Tan et al., 2025); - Provides justification and explanations, supports observations with data (Friess, 2011, 2012; Isherwood, 2018; Sauro, 2010)	- Justifies with evidence, providing support for the identified cause ($N = 23$) - Identifies the accurate root cause, going beyond surface-level observations ($N = 17$) - Clarity, describing the causes clearly and with sufficient details ($N = 19$) - Relevant, directly addressing the usability problem ($N = 17$)
Redesign recommend-ations	- Technically feasible, functionally correct, and actionable (eg, provides enough detail for implementation, broken down into smaller steps) (Dumas & Redish, 1999; Jury et al., 2024; Maze, 2024; Schade, 2013; UXBoost, 2021; Xtensio, 2023); - Provides reasoning for the recommendation and explains its projected impact (Woodmass, 2020); - Diverse range of ideas (Huang et al., 2025; Padmakumar & He, 2024)	- Actionable and feasible, considering technical constraints, time, and cost ($N = 22$) - Effective, directly addressing the usability problem ($N = 26$) - Positive impact on overall UX ($N = 21$) - Creative ($N = 3$)
Overall language	- Clear and easy to understand (Huang et al., 2025; Jury et al., 2024; Maze, 2024; Tan et al., 2025); - Logical and follows a structured format (Dumas & Redish, 1999; Huang et al., 2025; IxDF, 2018); - Professional and objective tone, avoiding excessive personal interpretations (Celikyilmaz et al., 2021; Friess, 2012; Xtensio, 2023)	- Logical and clear, presenting coherent information ($N = 31$)

mobile applications (N = 33), VR applications (N = 21), and physical products (N = 16). Thirty respondents reported being extremely or very familiar with usability analysis. Similar to our approach in the literature review, two authors independently coded the evaluation criteria from each respondent, and these findings were consolidated through iterative discussions.

3.2. Criteria synthesized from literature review and survey

To synthesize findings from the literature review and survey, the first two authors, each with professional UX experience as researchers and designers, coded all survey responses and compared them with the criteria identified in the literature. We counted the frequency of each criterion mentioned in survey responses (summarized in Table 1) and cross-checked whether these criteria also appeared in prior work. The authors then engaged in iterative discussions to refine and consolidate overlapping concepts. To ensure interpretability and avoid complexity, we selected two representative criteria for each

component of a usability result (problem, cause, and redesign recommendation), along with one overarching criterion for the overall language clarity of the result. We limited each component to two criteria to maintain a concise yet conceptually clear framework and prioritized distinct and meaningful criteria rather than creating a long or redundant list. For instance, among redesign recommendations, 21 responses referred to a "positive impact on overall UX" and 26 mentioned that the recommendation is "effective." Since both expressions reflected the same underlying concept of the redesign leading to meaningful improvement, we merged them into a single criterion of effectiveness. In contrast, we retained "effectiveness" and "feasibility" as separate criteria, as they represent different considerations that independently influence the quality and practicality of redesign decisions.

For the usability problem, the most important criterion was **accuracy**, meaning that the problem description should reflect what actually occurred in the usability video and align with user feedback and evidence. The second key criterion was **specificity**, meaning the problem should be described in sufficient detail to avoid ambiguity.

For the underlying cause, **accuracy** was again the most emphasized criterion. This includes correctly identifying the relevant design element or interaction flow. The second criterion was **justification**, requiring a clear explanation of why that element led to the observed problem.

For the redesign recommendation, the most important criterion was **feasibility**, where the solution should be realistic to implement given current technical constraints or design timelines. The second was **effectiveness**, meaning the recommendation should meaningfully address and resolve the identified problem.

In addition to these criteria, we included one overarching criterion: **language clarity**, which applies across all components. This ensures that the usability findings are clearly and professionally communicated.

Each of these criteria was translated into a statement that can be rated on a 5-point Likert scale (from *strongly disagree* to *strongly agree*). The final list of evaluation criteria is as follows:

1. The usability problem description is **accurate**, correctly reflecting the video content.
2. The usability problem description is **specific**, providing sufficient details for understanding.
3. The cause of the problem is **accurate**, identifying the factors that led to the issue.
4. The cause of the problem is **justified**, presenting evidence to support the explanation.
5. The redesign recommendations are **feasible**, considering technical constraints, time, and cost.
6. The redesign recommendations are **effective**, directly resolving the usability problem.
7. The overall language is **clear** and **easy to understand**.

4. User study

After developing the evaluation criteria, we needed a dataset of usability results to apply them to. This section outlines our method for curating that dataset and conducting a user study in which UX experts rated the usability results using the criteria. All experiments in this work were reviewed and approved by the Institutional Review Board at Rochester Institute of Technology, with protocol approval #01102723.

4.1. Curating a dataset of usability results

We curated a dataset of 15 usability test videos featuring five different users, each interacting with three distinct products: a desktop website, a smartphone app, and a VR headset. The tasks were designed to reflect the core functionalities of each product. For example, checking the forecast for a specific date in a weather app, creating a poster using a GenAI design website, or playing games while wearing a VR headset. The users were seniors and students recruited from the first author's institution. Because the users were unfamiliar with the products, they experienced numerous usability problems, making these videos well-suited for analysis.

4.1.1. Role prompting strategy

After collecting the videos, we used ChatGPT (powered by GPT − 4 at the time of the study) to generate usability results. For the **baseline AI** condition, we first uploaded screenshots of the main interface pages for each product directly into the ChatGPT interface. We then provided the corresponding user task descriptions and video transcripts. Finally, we prompted ChatGPT to generate usability problem descriptions, their underlying causes, and redesign recommendations using a structured format. The prompt was as follows:

> This is the transcript of a usability test where a participant used the think-aloud protocol to complete the following tasks: [insert video tasks]
>
> **Transcript**: [insert transcript from video]
>
> **Task**: Based on your knowledge of the screenshots 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.

For the tailored AI condition, we created a custom GPT, also powered by GPT − 4 at the time. Following OpenAI's Creating a GPT guide,[2] we configured the system instructions using the GPT Builder interface in the ChatGPT web platform. The custom GPT was assigned the following persona:

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 (Rynes et al., 1997). You frequently customize usability problem descriptions and prioritize ensuring reliability in your analyses (Følstad et al., 2010; Kuang et al., 2022). Additionally, you actively question inconsistencies and seek critical information to deepen your understanding of design goals (Foong et al., 2017).

We then followed the same procedure as with the baseline AI: first uploading product screenshots, and then using the prompt to instruct the custom GPT to generate usability problem descriptions, causes, and redesign recommendations. Prior research has shown that generating multiple responses can improve the consistency and breadth of ChatGPT's output (Giannakopoulos et al., 2023; Ronanki et al., 2024). To ensure a more comprehensive set of usability problems, we prompted both the baseline ChatGPT and the custom GPT three times per video. We then aggregated these responses by taking the union of all unique usability problems identified across runs. When redundant or overlapping problems were detected, we retained the first instance. In cases where the same problem had contradictory elements (eg, differing identified causes or redesign recommendations), we selected the version that appeared with higher frequency across the three runs. A manual review confirmed that there were no cases in which all three runs produced entirely different causes or redesign recommendations for the same problem.

4.1.2. Human review strategy

For the human review strategy, we recruited 12 UX professionals (10 females, 2 males) with an average of 4.6 years of UX experience (SD = 2.1).[3] Each participant analyzed all 15 usability videos under three different conditions: *no AI support* (manual detection of usability problems), *baseline AI support*, and *tailored AI support*. The analysis was conducted over five 1-hr sessions within a two-week period, where participants analyzed three videos per session. This provided sufficient time for iterative analysis and revision in the human-only, baseline AI & human, and tailored AI & human conditions, comparable to the multi-round generation used in the AI-only conditions. The order of conditions across participants was counterbalanced using a balanced Latin square design to ensure that each condition appeared before and after every other condition an equal number of times (Schwind et al., 2023).

In both AI-supported conditions, participants were shown AI-generated usability problem suggestions during their analysis, following a method similar to Kuang et al. (2024). We did not disclose the differences between the baseline and tailored AI to the participants. When an AI-generated suggestion appeared, participants had three options: they could *accept* the suggestion as is, *edit* it based on their own judgment, or *ignore* it entirely. This approach ensured that participants critically reviewed each suggestion, while ultimately retaining full responsibility for the final analysis.

4.1.3. Dataset sampling

Following the prompt customization and human review of the 15 usability videos, we collected over 1,000 usability problem descriptions, along with their causes and redesign recommendations. Given the size of this dataset, it was not feasible to have experts evaluate the full corpus without making the study overly long. To address this, we applied a two-step subsampling approach for each video. First, we identified the set of unique usability problems that appeared across all three human-involved conditions: *human-only* (no AI), *human & baseline AI*, and *human & tailored AI*. Since the *baseline AI* and *tailored AI* problems were first generated then subsequently reviewed by humans, the *human & baseline AI* and *human & tailored AI* conditions already covered problems in the AI-only conditions. Second, from this list of overlapping problems, we randomly selected one problem per video. While a single problem may not capture the range of severity or difficulty within a video, the study included 15 videos spanning three distinct product types (i.e., a website, a mobile app, and a VR game), thereby prioritizing diversity in interaction contexts and ensuring coverage of usability problems across domains.

The two-step process ensured that the selected problem appeared in all five experimental conditions, allowing for a consistent basis of comparison across AI and non-AI-generated results. In total, we curated 75 usability problem descriptions: five versions (one per condition) for each of the 15 videos. We acknowledge that focusing on overlapping problems across all conditions may bias the selection toward more apparent usability problems. However, this approach was necessary to enable a fair comparison of the *quality* and *content* of problem descriptions across conditions, rather than the quantity or variety of issues identified.

Examples of these usability problem descriptions are in Supplementary Appendix A.2 (Supplementary Table 3 for the app video, Supplementary Table 4 for the website video, and Supplementary Table 5 for the VR video). We report key characteristics of the dataset to provide context on the usability test results, including the total word count and the LIX score,[4] a readability metric that indicates the difficulty of a given text (Björnsson, 1983). As mentioned in Section 2.2, the LIX score has been employed to compare human-written and LLM-generated text (Schneiders et al., 2025). Table 2 presents these metrics for each condition.

We conducted one-way between-subjects ANOVA tests, which revealed significant differences among the five conditions in the word count for the cause ($F_{4,70} = 3.0, p < 0.05, \eta_p^2 = 0.15$) and redesign recommendations ($F_{4,70} = 14.1, p < 0.0001, \eta_p^2 = 0.45$). We also observed significant differences in the LIX scores for the problem ($F_{4,70} = 7.9, p < 0.0001, \eta_p^2 = 0.31$), cause ($F_{4,70} = 3.0, p < 0.05, \eta_p^2 = 0.15$), and redesign recommendations ($F_{4,70} = 9.7, p < 0.0001, \eta_p^2 = 0.36$). Post-hoc comparisons with Bonferroni correction indicated that the word count for the cause in the human-only condition was significantly lower than in the baseline AI, baseline AI & human, and tailored AI & human conditions. Similarly, the word count for redesign recommendations in the human-only condition was significantly lower than in all other conditions, while the tailored AI & human condition had a significantly higher word count than all others.

The LIX scores for the human-only condition ranged from 34.1 to 36.2, corresponding to readability levels between "easy" and "average."[5] For problem descriptions and redesign recommendations, the human-only condition had significantly lower LIX scores than all other conditions. Additionally, the LIX score for the cause in the human-only condition was significantly lower than in the tailored AI-only condition. The tailored AI condition produced the highest LIX score (64.2), categorized as

Table 2. Word count and LIX scores for the usability result dataset across conditions, presented as "mean (standard deviation)."

Variable	Human only	Baseline AI only	Tailored AI only	Baseline AI & human	Tailored AI & human
Word count (problem)	9.5 (3.6)	8.6 (4.4)	7.1 (2.7)	8.5 (4.5)	7.7 (2.3)
Word count (cause)*	10.5 (6.2)	17.2 (4.7)	15.0 (5.9)	17.8 (7.3)	17.7 (9.3)
Word count (redesign recommendation)***	12.1 (5.0)	25.1 (5.1)	24.9 (6.7)	23.5 (6.4)	33.2 (12.9)
LIX score (problem)***	34.1 (17.1)	54.9 (13.4)	64.2 (17.9)	53.8 (14.3)	56.4 (14.0)
LIX score (cause)*	36.2 (22.6)	46.2 (10.3)	53.2 (7.0)	45.3 (9.8)	42.1 (7.6)
LIX score (redesign recommendation)***	35.6 (10.3)	52.6 (9.6)	55.5 (9.8)	48.0 (11.9)	47.1 (11.8)

Note: Asterisks denote a significant difference between conditions (*$p < 0.05$ and ***$p < 0.001$).

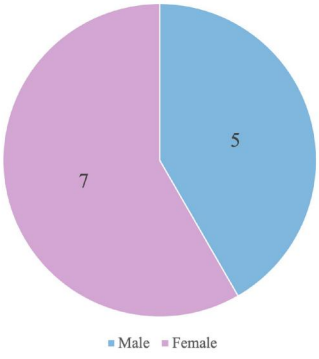
"very difficult," while the tailored AI & human condition ranged from 42.1 to 56.4, falling between "average" and "difficult." This pattern suggests that AI-generated text tends to be more complex and that human involvement can reduce, but not eliminate, this complexity.

However, while LIX scores highlight differences in the readability of usability test results across conditions, they do not directly reflect quality, as readability alone does not capture whether a problem is accurately identified or if a recommendation is actionable. Therefore, to evaluate the practical utility of the findings, we engaged expert UX evaluators to assess the dataset using structured criteria grounded in UX practice.

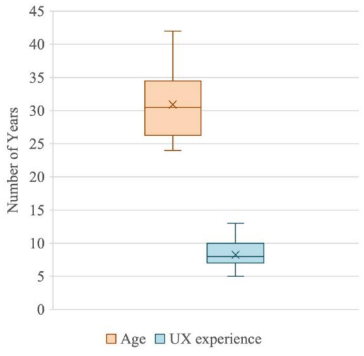
4.2. Participants

Based on the screening survey results (Section 3.1.2), we determined participants' eligibility for the expert evaluation. The inclusion criteria required at least five years of UX experience, strong familiarity with usability analysis, prior experience evaluating usability reports, and broad coverage across product domains. To develop the evaluation criteria (Section 3.1.2), we first surveyed a broader group of UX professionals (average 4.7 years of experience) to capture diverse usability practices. For the subsequent criteria evaluation, we purposefully recruited more experienced experts to provide deeper and more qualified judgments. Thus, some participants took part in both phases, ensuring continuity between the development and evaluation of the criteria. This design allowed the criteria to be informed by general UX practices while their assessment drew on expert-level insight.

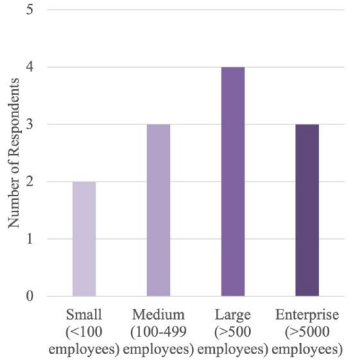
Figure 2 presents the demographic distribution of the participants. In total, twelve UX experts (7 female, 5 male) took part in the evaluation, with an average age of 30.9 years (SD = 5.5) and an average



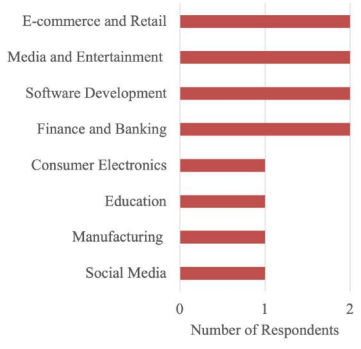
(a) Pie chart showing the gender of the participants in the expert evaluation.



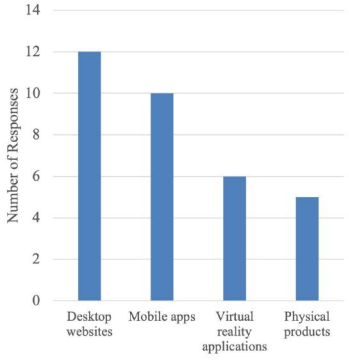
(b) Box plot showing the distribution of participants' age and years of UX experience.



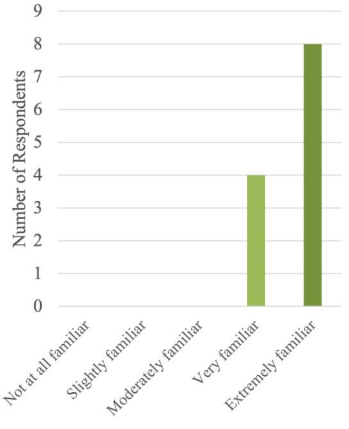
(c) Bar chart showing the company size of the participants.



(d) Bar chart showing the industry of the company that the participants currently work at.



(e) Bar chart showing responses to "What types of products have you evaluated for usability? (Select all that apply)?"



(f) Bar chart showing responses to "How familiar are you with identifying usability problems from usability test videos?"

Figure 2. Demographic information of participants in the expert evaluation.

of 8.2 years of UX experience (SD = 2.4). The participants represented a diverse professional background, working in organizations of varying sizes, from small companies with fewer than 100 employees to multinational enterprises with over 5,000 employees. They also came from a broad range of industries, such as e-commerce and retail, media and entertainment, and software development, reflecting a wide cross-section of UX practice contexts. All twelve had experience evaluating the usability of desktop websites, 10 had evaluated mobile applications, six had evaluated virtual reality applications, and five had worked with physical products. Eight experts reported being "extremely familiar" with usability analysis, while four indicated they were "very familiar." Their frequency of evaluations ranged from weekly to quarterly.

4.3. Procedure

The expert evaluation consisted of a 90-min synchronous remote session followed by an asynchronous survey. To facilitate the rating process, we created two Qualtrics surveys, each with 15 pages (one page per video). Each page presented a video snippet illustrating a user encountering a usability issue, followed by descriptions of the problem, its underlying cause, and redesign recommendations. The first survey included three conditions per page: (1) human-only, (2) human & baseline AI, and (3) human & tailored AI, while the second survey included two conditions: (1) baseline AI and (2) tailored AI. In both surveys, the order of conditions on each page was counterbalanced, and the sequence of the 15 pages was randomized to mitigate order effects. Although experts viewed each video once during the synchronous session and once again in the follow-up survey, potential learning effects between the two viewings cannot be fully ruled out. These effects may have increased familiarity with the videos, but the one to two week gap and randomized order likely reduced their impact.

During the synchronous session, the experts were first introduced to the evaluation goals and then provided with a survey link to complete their ratings. They were not informed of the differences between conditions while completing the survey. After submitting their ratings, they participated in a semi-structured interview. In the interview, experts first discussed their experiences using the evaluation criteria and whether they believed the criteria accurately reflected the actual quality of the reports. Subsequently, we presented the dataset of usability results that they just rated with the corresponding conditions, allowing the experts to assess the overall quality of the reports and discuss any observed differences between conditions. All sessions were video-recorded for later analysis. Within two weeks of the synchronous session, experts completed the follow-up survey asynchronously.

4.4. Data analysis

For the qualitative data, the synchronous session recordings were initially transcribed using automatic speech-to-text software. One researcher rewatched the recordings to manually correct any transcription errors. Two researchers independently analyzed the corrected transcripts using inductive coding. We then met to discuss the themes that emerged from the data, refining and grouping the codes into overarching themes through iterative discussions.

For the quantitative analysis of the ratings across various criteria, we performed Friedman's Test, a non-parametric test appropriate for within-subjects designs with three or more conditions, to assess whether there were significant differences among the five conditions. We then conducted post hoc pairwise comparisons using Conover's F test to identify which specific condition pairs showed significant differences. Additionally, we reported the effect size for Friedman's Test using Kendall's W coefficient (W) to show the strength of the observed differences.

5. Results

5.1. RQ1: Effectiveness of quality evaluation criteria

To assess the effectiveness of the criteria set, we analyzed interview responses from expert evaluators, using their feedback to validate the criteria. Overall, participants considered the seven criteria to be

highly effective and user-friendly for evaluating the quality of usability reports. They also suggested potential enhancements to the criteria. We characterized the effectiveness of our criteria (see list in Section 3.2) from two perspectives: (1) Functionality and (2) Usability.

5.1.1. Functionality

Functionality encompasses the **standardization** and **comprehensiveness** of the criteria. Nine participants believed that the established criteria offered a new method of standardizing quality assessments. Currently, participants relied on personally established (P7, P10, P11) or client-supplied (P5) criteria to evaluate quality. A key advantage of establishing the set of criteria was facilitating the standardization of all the information contained in the usability report. As P4 said, "Having this as a rubric can help frame the report in a consistent way, including structure and level of details." Additionally, participants mentioned that this set of criteria was beneficial for generalized application across different types of products.

Seven participants said that the set of criteria was sufficient to capture most of the important aspects of usability results. In addition, serving as a foundation, it was flexible to add less frequent criteria for specific cases, such as "adoption" (P5) and "sustainability" (P7) of the redesign recommendations.

5.1.2. Usability

Usability encompasses the **content interpretability** and **ease of use** of the criteria. Four participants noted that learning the set of criteria required some effort. For instance, for the problem specificity, P3 said, "This implies that the statement is indeed more precise. I was a bit confused initially, but eventually I understood it on my own." P8 recommended refining the criteria description to improve clarity and make it easier to interpret for first-time users.

Seven participants emphasized that the rating system was "straightforward" (P12) and "easy to use" (P2, P3, P5). P1 said, "If there were more like seven or nine points, it would take me longer to make a decision." In addition, P7 and P8 suggested changing the labels for the points from agree or disagree to a score from one to five to easily quantify the ratings.

5.1.3. Qualitative feedback on the refinements of the criteria

Participants primarily proposed two refinements to our evaluation criteria. First, four participants identified potential overlap between "The usability problem description is accurate" and "specific." P9 found these two criteria to be quite similar, while P2 and P6 explained that they assigned identical ratings across different conditions for both criteria. These findings suggest that these two criteria could either be merged or their distinctions more clearly articulated.

Beyond our proposed criteria, participants suggested additional evaluation criteria outside our current scope. For example, P3 and P10 emphasized the importance of "priority" and "severity" when organizing multiple usability problems, with P3 noting, "Another consideration is arranging the problems according to their importance." Similarly, P7 and P8 proposed "sustainability" and "diversity" as additional criteria for evaluating redesign recommendations. P7 highlighted, "I think the sustainability, the environmental side, is something we do not consider all the time."

However, our criteria were specifically designed to evaluate the quality of individual usability problems rather than the structural organization of usability reports, which explains our exclusion of "priority" and "severity." Regarding "diversity," while this criterion emerged from our literature review, no participants mentioned it in our survey, indicating that it may not be widely applied in current practice.

5.2. RQ2: Comparison of usability results between human-only, AI-only, and human-AI collaborative analysis

Figure 3 presents a boxplot illustrating the ratings for the seven criteria across the five conditions. Supplementary Table 6 in the Supplementary Appendix shows the p-value and effect sizes (r) for all

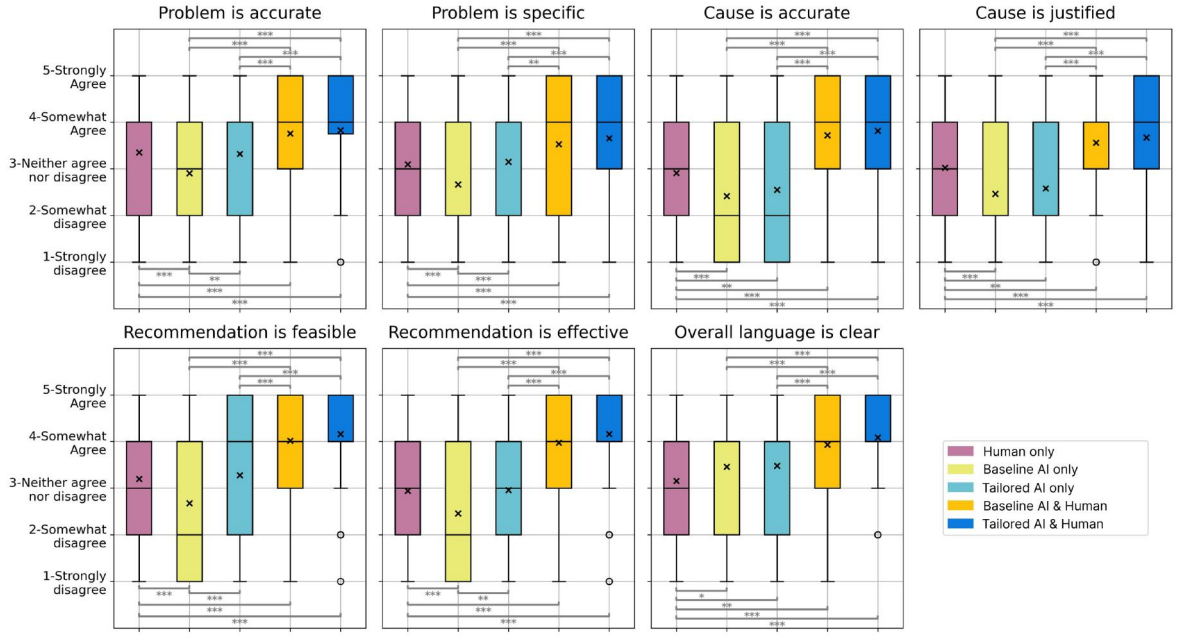


Figure 3. Boxplot showing the median, interquartile range, and overall distribution of the ratings for the seven criteria in each condition: human only, baseline AI only, tailored AI only, baseline AI & human, and tailored AI & human. The mean of each condition is marked with "x" and asterisks denote a significant difference between conditions ($^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$). These visualizations highlight differences in central tendency and variability among conditions. The boxes for the baseline AI & human and tailored AI & human conditions tended to rest above those for the other three.

pairwise comparisons. Our results revealed four main patterns in the quality of the usability results:

- The combination of human and AI was rated as significantly higher than either alone on all criteria.
- The tailored AI condition performed at a similar level as a human for problem accuracy, problem specificity, redesign recommendation feasibility, and redesign recommendation effectiveness, while both were significantly higher than the baseline AI.
- Both baseline and tailored AIs were rated as significantly worse at determining the accuracy and justification of the cause of usability problems than humans only.
- Usability findings generated by humans were rated significantly less clear in language than those produced under AI conditions.

We discuss the differences for each criterion below:

5.2.1. Criteria 1: Problem accuracy

Friedman's test revealed significant differences across the conditions ($\chi^2(4) = 67.38$, $p < 0.0001$, $W = 0.10$[6]). Post hoc pairwise comparisons with Conover's F indicated significant differences between all condition pairs, except for two: (1) the human-only condition compared to the tailored AI-only condition, and (2) the baseline AI & human combination compared to the tailored AI & human combination.

The baseline AI-only condition produced the least accurate problem descriptions, with pairwise effect sizes ranging from $r = 0.17$ (interpreted as small) to $r = 0.36$ (interpreted as medium).[7] In contrast, the tailored AI-only and human-only conditions resulted in more accurate descriptions, though no significant difference was found between these two conditions. For instance, P9 noted, "The tailored AI and human both identified the right issue, but the human used more casual language."

Furthermore, the AI and human combinations outperformed the conditions where only a AI or human was involved, with r ranging from 0.16 (small) to 0.36 (medium). However, no significant difference emerged between the two combined conditions. P5 observed, "I do like how the AI [and human] is more accurate," while P8 commented, "I think the human plus tailored AI definitely

produced the highest quality in defining the problem." This feedback suggests that human involvement enhanced the AI's accuracy, effectively compensating for the discrepancies in the AI's standalone performance.

5.2.2. Criteria 2: Problem specificity

Friedman's test revealed significant differences across the conditions ($\chi^2(4) = 69.96$, $p < 0.0001, W = 0.10$). Post hoc pairwise comparisons indicated significant differences between all pairs except two: (1) the human-only condition compared to the tailored AI-only condition, and (2) the baseline AI & human combination compared to the tailored AI & human combination.

The pattern of ratings for problem specificity mirrors that of problem accuracy (Criteria 1). The combination of AI and human consistently outperformed the conditions where either AI or human worked alone with r ranging from 0.14 (small) to 0.37 (medium), although no significant difference was found between the two combined conditions. As P12 noted, "The most detailed ones are from human and AI together." This suggests that human involvement improved the specificity of the AI's results, while the tailored AI performed comparably to human evaluators but significantly better than the baseline AI.

5.2.3. Criteria 3: Cause accuracy

Friedman's test revealed significant differences across the conditions ($\chi^2(4) = 164.45$, $p < 0.0001, W = 0.23$). Post hoc pairwise comparisons showed significant differences between all pairs except two: (1) baseline AI-only compared to tailored AI-only, and (2) baseline AI & human compared to tailored AI & human.

As shown in Figure 4, there were no significant differences between the baseline AI-only and tailored AI-only conditions, suggesting that role prompting did not improve the AI's ability to accurately identify the causes of usability problems. Human-identified causes were rated significantly more accurate than those generated by the baseline AI-only ($r = 0.20$) and tailored AI-only ($r = 0.14$) conditions. Although both effect sizes were small, several participants noticed differences. For instance, P1 mentioned, "The human identified the more correct root cause," while P7 noted, "I feel some of the AI descriptions didn't identify the root cause since it only reads or watches a specific part of the video." Moreover, the combination of AI and human consistently outperformed the conditions where either a AI or human worked alone, with r ranging from 0.32 (medium) to 0.51 (large). Again, no significant difference was found between the two combined conditions.

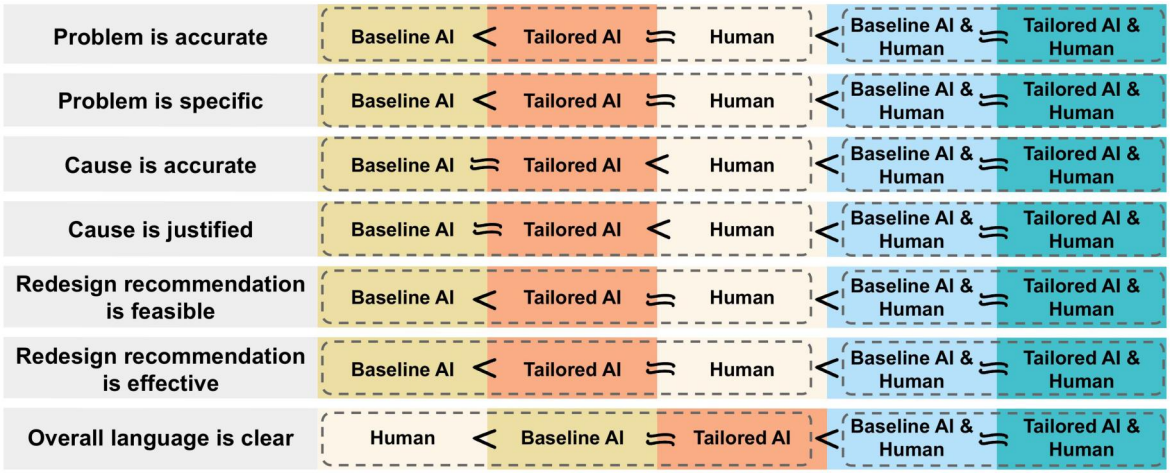


Figure 4. Table showing the patterns for the different conditions: human only, baseline AI only, tailored AI only, baseline AI & human, and tailored AI & human for each criteria. ("<" indicates that the condition on the left was rated significantly lower than the one on the right, and "≈" indicates no significant difference between the two.).

5.2.4. Criteria 4: Cause justification

Friedman's test revealed significant differences across the conditions ($\chi^2(4) = 125.31$, $p < 0.0001, W = 0.18$). Post hoc pairwise comparisons showed significant differences between all pairs, except for two: (1) baseline AI-only compared to tailored AI-only, and (2) baseline AI & human compared to tailored AI & human.

This pattern aligns with the results for cause accuracy (Criteria 3), indicating that human-only descriptions provided stronger justifications than baseline AI-only ($r = 0.22$) and tailored AI-only ($r = 0.17$), although the effect sizes were both small. As P2 observed, "I feel like humans interpret data better and explain it better to other humans." Furthermore, the combination of AI and human consistently produced the best results, with r ranging from 0.22 (small) to 0.46 (medium). P8 also noted, "The human and AI together provide more interpretations of what could be the reasons behind the problem."

5.2.5. Criteria 5: Redesign recommendation feasibility

A statistical analysis using Friedman's test revealed significant differences across the conditions ($\chi^2(4) = 125.60, p < 0.0001, W = 0.18$). Post hoc pairwise comparisons showed significant differences between all pairs, except for two: (1) human-only compared to tailored AI-only, and (2) baseline AI & human compared to tailored AI & human.

As illustrated in Figure 4, the ratings for the feasibility of redesign recommendations followed the same pattern as those for problem accuracy (Criteria 1) and problem specificity (Criteria 2). This suggests that human involvement enhanced the feasibility of the AI's recommendations, with r ranging from 0.27 (small) to 0.51 (large). The tailored AI performed at a similar level to human evaluators but significantly outperformed the baseline AI ($r = 0.20$). For instance, P8 commented, "The tailored AI with human gives a more diverse and feasible set of ideas compared to the others." Moreover, the combination of AI and human consistently produced the most feasible redesign recommendations, as AI alone sometimes provides "a very mainstream solution" (P2).

5.2.6. Criteria 6: Redesign recommendation effectiveness

A statistical analysis using Friedman's test revealed significant differences across the conditions ($\chi^2(4) = 174.58, p < 0.0001, W = 0.24$). Post hoc pairwise comparisons showed significant differences between all pairs, except for two: (1) human-only compared to tailored AI-only, and (2) baseline AI & human compared to tailored AI & human.

The ratings for the effectiveness of redesign recommendations followed a similar pattern to those for problem accuracy (Criteria 1), problem specificity (Criteria 2), and feasibility of redesign recommendations (Criteria 5). Human involvement improved the effectiveness of the AI's recommendations, with the tailored AI performing comparably to human evaluators but significantly better than the baseline AI ($r = 0.17$). The combination of AI and human consistently led to the most effective redesign recommendations, with r ranging from 0.36 (medium) to 0.58 (large). Participants also appeared more comfortable using AI for redesign recommendations compared to identifying problems and root causes. As P5 noted, "I definitely like the AI for the redesign recommendations, those are pretty good. So I would always use AI for redesign recommendations, while for the others, I rely more on humans."

5.2.7. Criteria 7: Clarity of language

Friedman's test revealed significant differences across the conditions ($\chi^2(4) = 101.65$, $p < 0.0001, W = 0.14$). Post hoc pairwise comparisons showed significant differences between all pairs, except for two: (1) baseline AI-only compared to tailored AI-only, and (2) baseline AI & human compared to tailored AI & human.

Unlike the other criteria, the human-only condition was rated significantly lower than all other conditions, with r ranging from 0.13 (small) to 0.42 (medium). No significant differences were found between the baseline AI-only and tailored AI-only conditions, nor between the two combined human-AI conditions. Participants noted key differences in language clarity based on the length of descriptions, level of detail, and formality. The human-only condition tended to have "shorter descriptions" (P1-3, P6-9), "grammar mistakes" (P1, P10), and was often written in a "very free-form manner, like spoken

language" (P6). In contrast, the AI-involved conditions were perceived to have "better wording" (P9) and were written in a more formal style, with "academic definitions, like in a human-computer interaction textbook" (P10). However, participants also identified instances where AIs used "weird terms" (P8), were "difficult to understand" (P3), or sounded "just like AI" (P4). For example, the baseline AI described a problem as "difficulty in achieving desired design elements," which participants found overly verbose. Our findings suggest that incorporating a baseline or tailored AI improved the formality of the human evaluators' language. The combination of AI and human achieved the highest language clarity, with effect sizes ranging from $r = 0.21$ (small) to $r = 0.42$ (medium). This improvement may stem from human input mitigating the AI's tendency to use jargon or complex phrasing. As shown in the LIX metrics (Table 2), the baseline AI & human and tailored AI & human conditions produced text that was easier to read than either AI-only condition.

5.2.8. Qualitative feedback on the role of AI

While completing the ratings, seven out of twelve participants did not suspect that any of the usability results were AI-generated. Some assumed they were written by different UX researchers with varying levels of experience. However, four participants suspected AI involvement due to the language style and the use of awkward or sophisticated terminology. For instance, P4 remarked, "The AI said 'manipulation of something,' whereas a human would say something more conversational, like 'it's hard to manipulate something'." One participant noted that they did not mind whether the results were AI-generated, as AI tools are already integrated into their workplace, and it is expected that these tools will be in use.

Overall, participants believed that AI can help overcome language barriers and improve the polish of human-written content. However, they suggested that AI language should become more conversational and easier to understand. A key strength of AI was seen in its ability to offer a broader range of redesign suggestions, complementing those proposed by humans. Additionally, participants felt that AI provides an objective perspective when analyzing usability problems, offering quick starting points that improve efficiency. While AI can accelerate the initial analysis, participants emphasized that humans should remain responsible for reviewing AI suggestions and identifying deeper root causes. AI can also contribute insights from domains outside of design, further enhancing the analysis. Given the increasing integration of AI into human workflows, participants anticipated that AI involvement would become a standard part of future work processes.

6. Discussion

6.1. Impact of AI on the perceived quality of usability results

6.1.1. Usability problem identification results

Our results indicate that, in terms of problem accuracy and specificity, the baseline AI received the lowest ratings, the tailored AI performed at an equivalent level to human-only condition, and the two AI-human conditions received the highest ratings. Given that AI identified usability problems based on transcripts of usability test videos in our study, the identification performance is influenced by semantic understanding capabilities.

Prior studies have primarily compared the accuracy and recall of different LLMs in identifying usability problems (Duan, Warner, et al., 2024). Our study extends this line of research by examining how LLM prompt engineering influences the quality of usability result descriptions. The results demonstrate that role prompting can improve problem accuracy and specificity to a level comparable to human-only evaluators. This finding aligns with previous work showing that role prompting is an effective strategy for accurately identifying usability problems and generating clear responses (Shin et al., 2025). A possible explanation is that prompt engineering enhances the semantic understanding capabilities of LLMs. Evidence from the healthcare domain supports this explanation, where the initial semantic understanding accuracy of ChatGPT 4 (78.73%) increased to 79.54% after refining prompts by incorporating core responsibilities and duties into the instructions (Pope et al., 2025).

Previous studies have also found that LLMs occasionally generate inaccurate descriptions of usability problems. For example, Kuang et al. reported that GPT-3.5 inaccurately described users' difficulty finding Coke in the drinks section as difficulty locating the "drinks" category (Kuang et al., 2024). Similarly, as shown in Supplementary Appendix A.2, LLMs sometimes generated general descriptions, such as "Unclear control scheme of gameplay," that obscure the specific design elements requiring redesign. In such cases, human evaluators needed to specify these elements to facilitate future redesign. This may explain why human involvement improved accuracy and specificity. Moreover, prior research has identified issues in human-written usability test reports, such as emotional wording, usability jargon, and vague comments (Dumas et al., 2004). LLMs, in contrast, showed potential to generate more neutral, accessible, and specific descriptions. This may further explain why human-AI collaboration outperforms the human-only condition in describing usability problems.

6.1.2. Cause identification results

For cause identification, the human-only condition was rated higher than both the baseline and tailored AI conditions. This suggests that AI on its own still lacks the reasoning and interpretive abilities needed to accurately identify root causes of usability problems and provide sufficient justification. This finding complements prior work on LLM-supported usability evaluation, which usually overlooked the understanding of causes. Previous studies primarily focused on generating the problem descriptions and improvement suggestions (Duan, Warner, et al., 2024; Kuang et al., 2024; Shin et al., 2025).

Root cause analysis in usability evaluation requires more than surface-level observations. It involves understanding the user's intent and expectations, analyzing the specific product features involved, and identifying the actions leading up to the occurrence of the problem. This process typically draws on user interaction data, qualitative feedback, and system logs to determine why a problem emerged (Ghulam Jillani, 2025). This type of reasoning often draws on "Theory of Mind" (ToM): the ability to infer others' beliefs, desires, and intentions based on their behavior (Leslie et al., 2004). While GPT − 4 has demonstrated some ToM-like capabilities resembling human inference, these remain unreliable, and other LLMs perform even more poorly (Gandhi et al., 2023). In addition, LLMs show limitations in processing visual-modal information. Although they can recognize objects, actions, and events, they require large amounts of labeled data and may not capture contextual and fine-grained details of the content (Zhao et al., 2023; Zhou et al., 2024). Both ToM reasoning and visual understanding are critical in usability evaluation, especially when evaluators must analyze recorded usability sessions to identify the root causes of problems. Even in text-based evaluations, limitations persist. For example, a recent study applied LLMs to conduct heuristic evaluations of the codebases for thirty open-source websites. While the models could identify some usability problems, their reasoning and severity assessments were inconsistent and often unreliable (Platt et al., 2025).

Moreover, such reasoning requires substantial domain knowledge and contextual understanding, which general-purpose LLMs may lack. In our study, we provided contextual information as comprehensively as possible, including usability test tasks, think-aloud transcripts, and product interfaces. However, the problem causes may extend beyond these interface- or interaction-based cues to user-specific characteristics (Mendoza & Novick, 2005). Therefore, LLMs cannot access the same depth of contextual understanding as human evaluators, which may contribute to their poorer performance in identifying underlying causes. As previous research indicates, improving the ToM reasoning capabilities of LLMs requires structured frameworks or fine-tuning with domain-specific datasets (Galitsky, 2025). This finding may help explain our results. Role prompting alone cannot improve performance in identifying underlying causes, since it neither strengthens ToM reasoning nor supplements contextual understanding. Together, these findings underscore the importance of human expertise in root cause identification. While LLMs can assist with surface-level problem detection, they are currently insufficient for the deeper reasoning required to explain **why** usability problems occur.

6.1.3. Redesign recommendation ideation results

Generating redesign recommendations is an ideation activity under constraints, focusing on resolving identified usability problems. Our results indicate that, in terms of recommendation feasibility and

effectiveness, the baseline AI received the lowest ratings, the tailored AI performed at an equivalent level to the human-only condition, and the two AI-human conditions received the highest ratings. Although previous studies have also leveraged LLMs to generate redesign recommendation, they either overlooked evaluation or assessed them using general metrics such as "helpfulness" (Duan, Warner, et al., 2024; Shin et al., 2025). Our findings complement this literature by providing a more nuanced understanding of why recommendations are helpful or not under different conditions. It should be noted that, compared to initial design ideation that emphasizes creativity, redesign focuses on feasibility and effectiveness, facilitating quick problem fixes (Hornbæk & Frøkjær, 2005). Therefore, our findings may be limited to product redesign contexts.

Previous studies have showed that LLMs can effectively generate design suggestions under various constraints, ranging from broad ones such as product industry to specific ones like color schemes (Shokrizadeh et al., 2025). Our findings extend this point by demonstrating that role prompting is an effective method to further improve the quality of design suggestions under constraints. A possible explanation is that generating high-quality redesign recommendations requires certain domain knowledge, such as formal rules (eg, Nielsen's 10 Usability Heuristic) and best practices (eg, existing features of similar applications) (Adinda & Suzianti, 2018). Role prompting can help elicit the LLMs' domain-specific capabilities (Pope et al., 2025; Sun et al., 2024), thereby raising the feasibility and effectiveness of redesign recommendations to a level comparable to human evaluators. However, Duan et al. also identified that even with prompting engineering, LLMs may still generate some vague suggestions, and users called for more concrete guidance such as "change this color to a specific value" (Duan, Warner, et al., 2024). This highlights the necessity of human oversight.

Previous studies have also pointed out that a common problem with human-written redesign recommendations is vagueness, such as "change the visual representation to discourage … behavior" (Dumas et al., 2004; Molich et al., 2007). Our findings align with these observations, revealing that the word count of redesign recommendations proposed by human evaluators is significantly lower than that of AI-generated ones, suggesting insufficient elaboration to clearly articulate the details. AI can mitigate this challenge by providing more specific drafts for human evaluators. This may explain why the two human-AI collaboration conditions outperformed the human-only condition in ideating redesign recommendations.

6.1.4. Language for usability results

Previous studies on LLM-supported usability evaluation have largely overlooked the linguistic dimension of generated results (Duan, Warner, et al., 2024; Kuang et al., 2024; Shin et al., 2025). Given that little domain-specific evidence exists, our discussion of this section draws on cross-domain findings from studies of AI-generated text quality to interpret the linguistic patterns observed in reporting the usability test results. Our results indicate that, in terms of language clarity, the human-only condition received the lowest ratings, followed by the two AI-only conditions, while the two AI-human conditions received the highest ratings. Although the AI-only conditions were rated higher than the human-only condition, prior research has identified several issues in AI-generated text. For instance, LLMs have been found to fall short as effective creative support tools for comedy writing, often generating bland and biased tropes (Mirowski et al., 2024). They also struggled with specificity and interpreting subtext in story summarization (Subbiah et al., 2024) and tended to produce overly positive narratives that lacked tension (Tian et al., 2024). Additionally, AI-generated text exhibits undesirable idiosyncrasies, which can be categorized into a seven-category taxonomy, including clichés, unnecessary exposition, purple prose, poor sentence structure, lack of specificity and detail, awkward word choice and phrasing, and tense inconsistency (Chakrabarty et al., 2025). These cross-domain observations, although drawn from creative and narrative writing, reveal similar tendencies that can undermine the clarity and perceived usefulness of usability test results. Consistent with this findings, participants in our study also noted many of these issues, suspecting AI involvement due to the language style and the use of awkward or overly sophisticated terminology.

Language Length and Complexity: Interestingly, a study in the legal domain found that participants were significantly more willing to act on LLM-generated advice than lawyer-generated advice when

unaware of the source (Schneiders et al., 2025). The researchers suggested that this could be due to participants conflating complexity with quality, as the LLM-generated advice contained more words on average and had a higher LIX score than the lawyer-generated advice. A similar pattern emerged in our study: the four conditions incorporating AI-generated suggestions had significantly higher word counts for cause and redesign recommendations, as well as significantly higher LIX scores for all three elements, compared to the human-only condition. However, we found that human involvement reduced text complexity. For example, the average LIX score for problem descriptions decreased from 64.2 in the tailored AI condition to 56.4 in the tailored AI & human condition (Table 2). Although this reduction was not statistically significant, it suggests that human involvement helped refine AI-generated usability results, potentially by eliminating awkward phrasing and overly sophisticated terminology. In reporting usability test results, linguistic moderation is essential for reducing unnecessary complexity and improving communicative quality, thereby enhancing both their comprehensibility and actionability (Dumas et al., 2004).

Comparability to Real-World Usability Results: LLMs often exhibit overconfidence in their descriptions, using complex vocabulary, structured sentence patterns, and formal syntax (Chhikara, 2025). While this can enhance the perceived fluency and authority, it also increases the risk of exposing users to decisively worded but hallucinated facts. LLMs are trained to follow grammatical rules and generate convincing, well-structured responses, making them particularly effective for formal writing tasks such as reports (Schneiders et al., 2025). However, their strengths in structured language contrast with the more free-form and rough-note style commonly used by human evaluators during video analysis.

In real-world usability evaluations, human-generated notes often require post-processing to refine their structure and clarity. For instance, humans may need to edit their own descriptions or have them reviewed by colleagues to ensure readability and professionalism (Woodmass, 2020; Xtensio, 2023). Moreover, when presenting findings to stakeholders, UX evaluators typically invest significant time in creating slides and refining language for clarity and impact—an effort that was not replicated in the controlled context of our studies (IxDF, 2018; Schade, 2013). Thus, a key implication of our study is that AI can serve as a valuable tool for refining usability reports, improving clarity, structure, and coherence for professional communication. Rather than replacing human judgment, AI could be leveraged as an assistant to enhance the presentation of usability results, helping UX evaluators more efficiently and effectively convey insights to stakeholders. Therefore, while parallels with other domains help us interpret the linguistic behavior of LLMs, our study focuses on revealing how different conditions shape distinct linguistic tendencies, which influence the communicative quality of usability test results.

6.2. Implications for human-AI synergy in usability analysis

When comparing the three conditions involving human participation—human-only, baseline AI & human, and tailored AI & human—we found that the latter two were consistently rated significantly higher than the human-only condition. This suggests that joint performance results in the highest quality of usability problem descriptions, root causes, and redesign recommendations. The benefits of AI-assisted collaboration have been widely demonstrated in prior research. By combining human creativity and contextual understanding with AI's scalability and analytical capabilities, human–AI teams can produce more innovative solutions and improve decision-making (Nicolescu & Tudorache, 2022; Vaccaro et al., 2024; Zhang et al., 2020). This aligns with the concept of "human augmentation," in which human–AI collaboration enhances human abilities and outperforms humans working alone (Vaccaro et al., 2024). A related but distinct concept is "human–AI synergy," where the collaboration between humans and AI not only surpasses human-only performance but also exceeds the capabilities of AI alone (Bansal et al., 2021; Vaccaro et al., 2024). Our findings contribute to this growing body of evidence, demonstrating that AI can play a valuable role in usability analysis by refining human-generated insights, improving clarity, and aiding in structured decision-making.

Interestingly, a meta-analysis of 106 experiments comparing human-only, AI-only, and human–AI systems found strong evidence for human augmentation but negative human–AI synergy. In other words, human–AI collaboration often underperformed compared to either humans or AI alone in certain cases (Vaccaro et al., 2024). However, a closer analysis revealed that the effect size for human–AI

synergy varied by task type: it was positive for content creation tasks (eg, generating open-ended responses) but negative for decision-making tasks (eg, selecting from a predefined set of options) (Vaccaro et al., 2024). Since our study involved generating open-ended usability descriptions based on video observations, our findings align with the positive human–AI synergy observed in content creation tasks. This suggests that AI can effectively support usability analysis by enhancing human-generated descriptions, reinforcing the idea that human–AI collaboration is particularly beneficial in tasks requiring creativity and contextual interpretation.

Another important question raised by our study is whether AI could eventually replace humans in usability analysis. Our results show that with carefully customized prompts, ChatGPT—when prompted to act as an experienced UX evaluator—performed at a comparable level to human evaluators in terms of problem description accuracy and specificity, as well as the feasibility and effectiveness of redesign recommendations. However, AI still lacked the reasoning and interpretive capabilities needed to accurately identify root causes and provide sufficient justification. Despite AI's ability to generate structured and well-articulated insights, human involvement remains crucial for interpreting those insights with empathy, which involves understanding users' contexts, emotions, and needs. Each evaluator brings a unique perspective shaped by personal beliefs, experiences, and cultural background, which AI cannot fully replicate. This need for human insight is supported by studies showing that while ChatGPT helped users generate a greater number of ideas, it also introduced homogenization effects, which reduced semantic diversity among different users (Anderson et al., 2024) and decreased overall lexical and content variation in augmentative essay writing (Moon et al., 2025). Similarly, when a Western-centric AI model provided writing suggestions to users from different cultural backgrounds, participants adapted to Western writing norms, diminishing nuances essential for cultural expression (Agarwal et al., 2025). In usability analysis, designing for diverse audiences requires a deep understanding of cultural context, lived experience, and historical background, which are factors that AI alone cannot fully capture. This diversity in perspective is vital for identifying problems and solutions that are meaningful and inclusive across user groups. While AI can help structure and surface insights, human evaluators are essential for interpreting findings in ways that honor cultural nuance and ensure that design recommendations are contextually appropriate and equitable.

Overall, our study reinforces the continued need for human involvement in the usability analysis process, highlighting the potential for human–AI synergy. Rather than replacing human evaluators, AI can serve as a valuable tool for augmenting usability analysis by enhancing clarity, structure, and idea generation. Future research could further optimize this synergy by exploring methods to calibrate users' trust in AI, such as integrating explanations and confidence scores (Zhang et al., 2020), providing source references for AI-generated insights (S. S. Y. Kim et al., 2025), and personalizing AI assistance based on individual tendencies toward over-reliance or under-reliance on AI (Swaroop et al., 2025). By refining AI's role in usability analysis, we can ensure that AI serves as a collaborative partner that enhances, rather than diminishes, the depth and reliability of human-driven insights.

6.3. Effectiveness and generalizability of the proposed evaluation criteria

Our work introduces a foundational criteria set for evaluating the quality of usability test results (see Section 3.2) and demonstrates its practical effectiveness through expert evaluation. In our study, UX experts assessed AI-, human-, and hybrid-generated usability test results across all seven criteria. The findings in Section 5.2 indicate that the proposed criteria set is effective in distinguishing the strengths and limitations of different conditions. Moreover, because our study included usability test results spanning a range of product types (eg, mobile apps, AI tools, and VR environments) as well as tasks like daily routines, creative work, and games, the criteria have demonstrated applicability across diverse contexts. This suggests that our framework can serve as a reliable foundation for assessing usability insight quality in practical UX workflows.

However, much like Nielsen's original 10 usability heuristics, which were first proposed in 1990 and subsequently adapted to other product types such as VR products,[8] video games,[9] and cartoons,[10] our criteria set is not static and can be adapted. When applied to emerging AI products, Nielsen's heuristics were expanded to include additional heuristics such as "transparency," "explainability," and "bias and fairness."[11] Similarly, when applying our criteria beyond the scope of product types in our dataset, such

as human-machine interfaces in vehicles and smart home systems, additional criteria closely related to specific product characteristics may be required.

Expert feedback in our study also pointed to additional criteria beyond our initial seven, such as "sustainability" and "diversity" (see Section 5.1). These suggestions reflect how companies may tailor usability evaluation to align with strategic priorities. For instance, firms emphasizing environmental impact may incorporate "sustainability" as a key usability concern (Santolaria et al., 2011), while start-ups prioritizing innovation might value criteria related to idea diversity or differentiation from competitors (Aminova & Marchi, 2021).

In summary, our criteria set has proven effective as a core framework for evaluating usability insight quality and can be adapted or extended to meet the evolving demands of emerging technologies and organizational contexts. It offers a structured, expert-validated foundation that can guide both research and practice in UX evaluation, while remaining open to domain-specific enhancements.

6.4. Limitations and future work

While our study provides valuable insights into human–AI collaboration in usability analysis, several limitations highlight opportunities for future research. A growing concern in AI-assisted design is the tendency of AI models to prioritize Western values and norms. For instance, studies have shown that AI-generated writing suggestions can homogenize text toward Western linguistic and rhetorical styles, diminishing cultural nuances (Agarwal et al., 2025). Since our study focused solely on Western-designed applications and was conducted in English, it remains unclear how AI impacts usability analysis in non-Western contexts or multilingual settings. Future research should explore the role of AI in usability evaluations across diverse cultural perspectives, languages, and user expectations. Investigating how AI-generated usability test results align with or diverge from region-specific design principles and user behaviors would be particularly valuable.

Our study was based on a relatively small dataset of 15 usability testing videos across three product categories: websites, mobile applications, and VR applications. While these categories represent a broad range of digital experiences, they do not capture the full diversity of usability challenges across different domains, such as smart home devices, automotive interfaces, or enterprise software. Future studies should expand the scope to include a wider variety of products and interaction modalities, ensuring that findings generalize across different types of usability evaluations. Furthermore, the dataset of usability test results showed significant differences in word count and LIX, which may bias perceived "accuracy" (eg, longer or more technical text may be perceived as more accurate). While we did not control for these factors across conditions to preserve ecological validity, future work could systematically control or normalize them to better isolate their effects on perceived accuracy.

We developed a set of evaluation criteria for usability results based on a literature review and a survey of UX evaluators. Our expert evaluation with twelve participants confirmed that these criteria were useful and accurately reflected usability result quality. However, given the small sample size, a larger-scale study is needed to validate whether these criteria generalize across a broader population of UX professionals. Although we reported participants' UX-related demographics (eg, years of experience, company size, industry), we did not collect information on seniority levels, which could further validate their expertise. Furthermore, we did not provide explicit rating guidelines (eg, examples of "high" vs. "low" feasibility) for the evaluation criteria. Instead, we relied on the experts' professional judgment to interpret and apply each criterion. While this approach reflects real-world evaluation practices, it may have introduced some variability in interpretation. Future work should develop and validate clearer operational definitions and rating examples to improve consistency and replicability across evaluators.

7. Conclusion

This study addressed two gaps in the field: the lack of standardized criteria for evaluating AI-generated usability test results, and the limited understanding of how moderation strategies, such as role prompting and human review, impact the quality of those insights. We developed a validated set of seven evaluation criteria grounded in UX literature and professional feedback, offering a practical framework

for assessing the quality of usability findings. Using these criteria, we evaluated outputs generated under five conditions, revealing that human-AI collaboration—particularly with a tailored AI and human moderation—consistently produced higher-quality results than either humans or AI alone. While AI contributed strengths in language clarity and structure, human evaluators were critical for interpreting root causes and ensuring the relevance and feasibility of redesign recommendations. Rather than humans or AI working in isolation, we advocate for collaborative workflows where each complements the other's strengths. These insights contribute to the broader human–computer interaction discourse on designing intelligent systems that collaborate with human experts in cognitively complex tasks.

Notes

1. See https://measuringu.com/three-goals/.
2. See https://help.openai.com/en/articles/8554397-creating-a-gpt.
3. This manuscript draws on usability results collected from a prior user study that examined the longitudinal impact of AI-assisted tools on UX evaluators' ability to identify usability problems. That study is currently under review. All data were collected in accordance with IRB-approved protocols, and informed consent was obtained from all participants.
4. $LIX = \frac{A}{B} + \frac{C \times 100}{A}$, where A = number of words in the text, B = number of sentences, and C = number of words with more than six letters
5. LIX ranking: Very Easy: 20; Easy: 30; Average: 40; Difficult: 50; Very Difficult: 60
6. Kendall's W uses Cohen's interpretation guidelines of $0.1 < 0.3$ (small effect), $0.3 < 0.5$ (moderate effect) and $>= 0.5$ (large effect), as summarized in a commonly used online reference (https://rpkgs.datanovia.com/rstatix/reference/friedman_effsize.html).
7. *r* also uses Cohen's interpretation guidelines of $0.1 < 0.3$ (small effect), $0.3 < 0.5$ (moderate effect) and $>= 0.5$ (large effect) (Brydges, 2019).
8. See https://www.nngroup.com/articles/usability-heuristics-virtual-reality/.
9. See https://www.nngroup.com/articles/usability-heuristics-applied-video-games/.
10. See https://www.uxtigers.com/post/heuristics-cartoons.
11. See https://www.uxstudioteam.com/ux-blog/10-usability-principles-for-ai.

Author contributions

CRediT: **Emily Kuang**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing; **Luyao Shen**: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing; **Ehsan Jahangirzadeh Soure**: Methodology, Software; **Mingming Fan**: Conceptualization, Supervision; **Kristen Shinohara**: Conceptualization, Supervision, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by Meta.

ORCID

Emily Kuang http://orcid.org/0000-0003-4635-0703
Luyao Shen http://orcid.org/0009-0006-8403-6271
Mingming Fan http://orcid.org/0000-0002-0356-4712
Kristen Shinohara http://orcid.org/0000-0002-1157-6369

References

Adinda, P. P., & Suzianti, A. (2018). Redesign of user interface for e-government application using usability testing method. In H. Unger & M. Arai (Eds.), *Proceedings of the 4th International Conference on Communication and*

Information Processing (Qingdao, China) (ICCIP '18) (pp. 145–149). Association for Computing Machinery. https://doi.org/10.1145/3290420.3290433

Agarwal, D., Naaman, M., & Vashistha, A. (2025). AI suggestions homogenize writing toward western styles and diminish cultural nuances. In N. Yamashita, V. Evers, K. Yatani, X. (S.) Ding, B. Lee, M. Chetty, & P. Toups-Dugas (Eds.), *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery. https://doi.org/10.1145/3706598.3713564

Ahmed, A., Hou, M., Xi, R., Zeng, X., & Shah, S. A. (2024). Prompt-eng: Healthcare prompt engineering: Revolutionizing healthcare applications with precision prompts. In R. Kumar, H. W. Lauw, & R. Ka-Wei Lee (Eds.), *Companion Proceedings of the ACM Web Conference 2024 (Singapore, Singapore) (WWW '24)* (pp. 1329–1337). Association for Computing Machinery. https://doi.org/10.1145/3589335.3651904

Aminova, M., & Marchi, E. (2021). The role of innovation on start-up failure vs. its success. *EuroMid Journal of Business and Tech-Innovation*, *4*(1), 41–72.

Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition (C&C '24)* (pp. 413–425). Association for Computing Machinery. https://doi.org/10.1145/3635636.3656204

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In P. Bjørn & S. Drucker (Eds.), *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21) (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445717

Björnsson, C. H. (1983). Readability of newspapers in 11 languages. *Reading Research Quarterly*, *18*(4), 480–497. https://doi.org/10.2307/747382

Borji, A. (2023). *A categorical archive of ChatGPT failures*. https://arxiv.org/abs/2302.03494

Brydges, C. R. (2019). Effect size guidelines, sample size calculations, and statistical power in gerontology. *Innovation in Aging*, *3*(4), igz036. https://doi.org/10.1093/geroni/igz036

Celikyilmaz, A., Clark, E., & Gao, J. (2021). Evaluation of text generation: A survey. https://doi.org/10.48550/arXiv.2006.14799

Chakrabarty, T., Laban, P., & Wu, C.-S. (2025). Can AI writing be salvaged? Mitigating idiosyncrasies and improving human-AI alignment in the writing process through edits. https://doi.org/10.48550/arXiv.2409.14509

Chen, J., Lu, X., Du, Y., Rejtig, M., Bagley, R., Horn, M., & Wilensky, U. (2024). Learning agent-based modeling with LLM companions: Experiences of novices and experts using ChatGPT & NetLogo chat. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24)*. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642377

Cheng, X., & Zhang, L. (2025). Inspiration booster or creative fixation? The dual mechanisms of LLMs in shaping individual creativity in tasks of different complexity. *Humanities and Social Sciences Communications*, *12*(1), 1–10. https://doi.org/10.1057/s41599-025-05867-9

Chhikara, P. (2025). Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=lyaHnHDdZl

Chilana, P. K., Wobbrock, J. O., & Ko, A. J. (2010). Understanding usability practices in complex domains. In G. Fitzpatrick, S. Hudson, K. Edwards, & T. Rodden (Eds.), *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10* (pp. 2337–2346). ACM Press. https://doi.org/10.1145/1753326.1753678

Duan, P., Cheng, C.-Y., Li, G., Hartmann, B., & Li, Y. (2024). UICrit: Enhancing automated design evaluation with a UI critique dataset. In L. Yao, M. Goel, A. Ion, & P. Lopes (Eds.), *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (Pittsburgh, PA, USA) (UIST '24)*. Association for Computing Machinery. https://doi.org/10.1145/3654777.3676381

Duan, P., Warner, J., Li, Y., & Hartmann, B. (2024). Generating automatic feedback on UI mockups with large language models. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24)*. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642782

Dumas, J. S., Molich, R., & Jeffries, R. (2004). Describing usability problems: Are we sending the right message? *Interactions*, *11*(4), 24–29. https://doi.org/10.1145/1005261.1005274

Dumas, J. S., & Redish, J. C. (1999). *A practical guide to usability testing* (Revised, subsequent ed.). Intellect Ltd.

Dzida, W., & Freitag, R. (2001). Usability testing—The Datech standard. In M. Wieczorek & D. Meyerhoff (Eds.), *Software quality: State of the art in management, testing, and tools* (pp. 160–177). Springer.

Fan, M., Li, Y., & Truong, K. N. (2020). Automatic detection of usability problem encounters in think-aloud sessions. *ACM Transactions on Interactive Intelligent Systems*, *10*(2), 1–24. https://doi.org/10.1145/3385732

Fan, M., Shi, S., & Truong, K. N. (2020). Practices and challenges of using think-aloud protocols in industry: An international survey. *Journal of Usability Studies*, *15*(2), 85–102. https://dl.acm.org/doi/10.5555/3532708.3532711

Fan, M., Yang, X., Yu, T., Liao, Q. V., & Zhao, J. (2022). Human-AI collaboration for UX evaluation: Effects of explanation and synchronization. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–32. https://doi.org/10.1145/3512943

Følstad, A., Lai-Chong Law, E., & Hornbæk, K. (2010). 6th Analysis in usability evaluations: An exploratory study. In A. Blandford & J. Gulliksen (Eds.), *Proceedings of the Nordic Conference on Human-Computer Interaction: Extending Boundaries (NordiCHI '10)* (pp. 647–650). Association for Computing Machinery. https://doi.org/10.1145/1868914.1868995

Følstad, A., Lai-Chong Law, E., & Hornbæk, K. (2012). Analysis in practical usability evaluation: A survey study. In E. H. Chi & K. Höök (Eds.), *Proceedings of the 30th SIGCHI Conference on Human Factors in Computing Systems - CHI '12* (pp. 2127–2136). ACM Press. https://doi.org/10.1145/2207676.2208365

Foong, E., Gergle, D., & Gerber, E. M. (2017). Novice and expert sensemaking of crowdsourced design feedback. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–18. https://doi.org/10.1145/3134680

Friess, E. (2011). Discourse variations between usability tests and usability reports. *Journal of Usability Studies*, *6*(3), 102–116. https://dl.acm.org/doi/10.5555/2007456.2007458

Friess, E. (2012). Do usability evaluators do what we think usability evaluators do? *Communication Design Quarterly Review*, *13*(1), 9–13. https://doi.org/10.1145/2424837.2424838

Galitsky, B. A. (2025). Improving ToM capabilities of LLMs in applied domains. https://www.preprints.org/manuscript/202502.1456

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. D. (2023). Understanding social reasoning in language models with language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '23)* (Article 595; pp. 13518–13529). Curran Associates.

Gao, Y., Wang, J., Lin, Z., & Sang, J. (2024). AIGCs confuse AI too: Investigating and explaining synthetic image-induced hallucinations in large vision-language models. In R. Subramanian, L. Zheng, V. K. Singh, P. Cesar, L. Xie, & D. Xu (Eds.), *Proceedings of the 32nd ACM International Conference on Multimedia (Melbourne VIC, Australia) (MM '24)* (pp. 9010–9018). Association for Computing Machinery. https://doi.org/10.1145/3664647.3681467

Ghulam Jillani, M. (2025). *How can you predict usability problems with root cause analysis before product launch?* https://www.linkedin.com/advice/3/how-can-you-predict-usability-problems-root-p6x8c

Giannakopoulos, K., Kavadella, A., Aaqel Salim, A., Stamatopoulos, V., & Kaklamanos, E. G. (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), e51580. https://doi.org/10.2196/51580

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24. https://doi.org/10.1145/3359152

Grigera, J., Garrido, A., Rivero, J. M., & Rossi, G. (2017). Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies*, *97*(2017), 129–148. https://doi.org/10.1016/j.ijhcs.2016.09.009

Hornbæk, K., & Frøkjær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. In C. Gale (Ed.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Portland, Oregon, USA) (CHI '05)* (pp. 391–400). Association for Computing Machinery. https://doi.org/10.1145/1054972.1055027

Hotjar. (2023). *How to analyze and evaluate usability tests in 5 steps*. https://www.hotjar.com/usability-testing/evaluation-analysis/

Hu, X., Mo, X., Jin, X., Chai, Y., Hu, Y., Fan, M., & Braud, T. (2025). Toward AI-driven UI transition intuitiveness inspection for smartphone apps. *International Journal of Human-Computer Studies*, *206*(2025), 103661. https://doi.org/10.1016/j.ijhcs.2025.103661

Huang, Q., Lv, C., Lu, L., & Tu, S. (2025). Evaluating the quality of AI-generated digital educational resources for university teaching and learning. *Systems*, *13*(3), 174. https://doi.org/10.3390/systems13030174

Interaction Design Foundation. (2018). *What are usability reports?* https://www.interaction-design.org/literature/topics/usability-reports

Isherwood, M. (2018). *How to write a user testing report that people will actually read*. https://uxdesign.cc/how-to-write-a-user-testing-report-that-people-will-actually-read-652d15d2f92e

Jeong, J., Kim, N., & In, H. (2020). Detecting usability problems in mobile applications on the basis of dissimilarity in user behavior. *International Journal of Human-Computer Studies*, *139*, Article 102364. https://doi.org/10.1016/j.ijhcs.2019.10.001

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Joshi, I., Shahid, S., Venneti, S. M., Vasu, M., Zheng, Y., Li, Y., Krishnamurthy, B., & Chan, G. Y.-Y. (2025). CoPrompter: User-centric evaluation of LLM instruction alignment for improved prompt engineering. In T. Li, F. Paternò, K. Väänänen, L. Leiva, D. Spano, & K. Verbert (Eds.), *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)* (pp. 341–365). Association for Computing Machinery. https://doi.org/10.1145/3708359.3712102

Jung, H., Jeong, Y., & Lee, J. (2026). Trapped in the prompt loop: Reprompt behavior in writing with ChatGPT. *International Journal of Human–Computer Interaction*, *42*(4), 2831–2859. https://doi.org/10.1080/10447318.2025.2530100

Jury, B., Lorusso, A., Leinonen, J., Denny, P., & Luxton-Reilly, A. (2024). Evaluating LLM-generated worked examples in an introductory programming course. In N. Herbert & C. Seton (Eds.), *Proceedings of the 26th Australasian Computing Education Conference (ACE '24)* (pp. 77–86). Association for Computing Machinery. https://doi.org/10.1145/3636243.3636252

Kim, S., Eun, J., Park, Y. E., Lee, K., Lee, G., & Lee, J. (2025). PromptPilot: Exploring user experience of prompting with AI-enhanced initiative in LLMs. *International Journal of Human–Computer Interaction*, *41*(23), 14779–14801. https://doi.org/10.1080/10447318.2025.2489030

Kim, S. S. Y., Vaughan, J. W., Liao, Q. V., Lombrozo, T., & Russakovsky, O. (2025). Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In N. Yamashita, V. Evers, K. Yatani, X. (S.) Ding, B. Lee, M. Chetty, & P. Toups-Dugas (Eds.), *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery. https://doi.org/10.1145/3706598.3714020

Krapp, E., Neuhaus, R., Hassenzahl, M., & Laschke, M. (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)* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3679318.3685501

Kuang, E. (2025). Evaluating usability challenges in VR games for older adults: A comparison with and without AI assistance. In A. Bianchi, E. Glassman, W. E. Mackay, S. Zhao, J. Kim, & I. Oakley (Eds.), *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '25)* (pp. 1–3). Association for Computing Machinery. https://doi.org/10.1145/3746058.3758343

Kuang, E., Jahangirzadeh Soure, E., Fan, M., Zhao, J., & Shinohara, K. (2023). Collaboration with conversational AI assistants for UX evaluation: Questions and how to ask them (voice vs. text). In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, & M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)* (pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3544548.3581247

Kuang, E., Jin, X., & Fan, M. (2022). "Merging results is no easy task": An international survey study of collaborative data analysis practices among UX practitioners. In S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3491102.3517647

Kuang, E., Li, M., Fan, M., & Shinohara, K. (2024). Enhancing UX evaluation through collaboration with conversational AI assistants: Effects of proactive dialogue and timing. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642168

Kuang, E., Soure, E. J., Shen, L., Goyal, N., Fan, M., & Shinohara, K. (2026). "It became my buddy, but I'm not afraid to disagree": A multi-session study of UX evaluators collaborating with conversational AI assistants. In A. Bozzon, T. Kosch, V. Liao, & X. Ma (Eds.), *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)* (pp. 1–26). Association for Computing Machinery. https://doi.org/10.1145/3772318.3790536

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Chicago, IL, USA) (FAccT '23)* (pp. 1369–1385). Association for Computing Machinery. https://doi.org/10.1145/3593013.3594087

Landauer, T. K. (1996). *The trouble with computers: Usefulness, usability, and productivity*. The MIT Press. https://doi.org/10.7551/mitpress/6918.001.0001

Leiser, F., Eckhardt, S., Leuthe, V., Knaeble, M., Mädche, A., Schwabe, G., & Sunyaev, A. (2024). HILL: A hallucination identifier for large language models. In Florian F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24)* (pp. 482–513). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642428

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences*, *8*(12), 528–533. https://doi.org/10.1016/j.tics.2004.10.001

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638

Liu, Z., Chen, C., Wang, J., Chen, M., Wu, B., Che, X., Wang, D., & Wang, Q. (2024). Make LLM a testing expert: Bringing human-like interaction to mobile GUI testing via functionality-aware decisions. In A. Roychoudhury & M. Storey (Eds.), *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (ICSE '24)* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3597503.3639180

Lu, Y., Yao, B., Gu, H., Huang, J., Wang, Z. J., Li, Y., Gesi, J., He, Q., Li, T. J.-J., & Wang, D. (2025). UXAgent: An LLM agent-based usability testing framework for web design. In N. Yamashita, V. Evers, K. Yatani, & X. (S.) Ding (Eds.), *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery. https://doi.org/10.1145/3706599.3719729

Maze. (2024). *How to report usability test results for maximum impact*. https://maze.co/guides/usability-testing/results/

Mendoza, V., & Novick, D. G. (2005). Usability over time. In R. M. Newman (Ed.), *Proceedings of the 23rd Annual International Conference on Design of Communication: Documenting & Designing for Pervasive Information (Coventry, United Kingdom) (SIGDOC '05)* (pp. 151–158). Association for Computing Machinery. In. https://doi.org/10.1145/1085313.1085348

Mirowski, P., Love, J., Mathewson, K., & Mohamed, S. (2024). A robot walks into a bar: Can language models serve as creativity supporttools for comedy? An evaluation of LLMs' humour alignment with comedians. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)* (pp. 1622–1636). Association for Computing Machinery. https://doi.org/10.1145/3630106.3658993

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. (2009). *Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement*. https://pubmed.ncbi.nlm.nih.gov/19621072/

Molich, R., Jeffries, R., & Dumas, J. S. (2007). Making usability recommendations useful and usable. *Journal of Usability Studies*, *2*(4), 162–179. https://dl.acm.org/doi/abs/10.5555/2835552.2835554

Moon, K., Adam, E. G., & Kushlev, K. (2025). Homogenizing effect of large language models (LLMs) on creative diversity: An empirical comparison of human and ChatGPT writing. *Computers in Human Behavior: Artificial Humans*, *6*(2025), 100207. https://doi.org/10.1016/j.chbah.2025.100207

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, *16*(5), 1–72. https://doi.org/10.1145/3744746

Nicolescu, L., & Tudorache, M. T. (2022). Human-computer interaction in customer service: The experience with AI chatbots—A systematic literature review. *Electronics*, *11*(10), 1579. https://doi.org/10.3390/electronics11101579

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In P. Bauersfeld, J. Bennett, & G. Lynch (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Monterey, California, USA) (CHI '92)* (pp. 373–380). Association for Computing Machinery. https://doi.org/10.1145/142750.142834

Nielsen, J. (2024). *10 usability heuristics for user interface design*. https://www.nngroup.com/articles/ten-usability-heuristics/

Nørgaard, M., & Hornbæk, K. (2006). What do usability evaluators do in practice? An explorative study of think-aloud testing. In S. Bødker & J. Coughlin (Eds.), *Proceedings of the 6th Conference on Designing Interactive Systems (DIS '06)* (pp. 209–218). Association for Computing Machinery. https://doi.org/10.1145/1142405.1142439

Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J., & Samwald, M. (2022). Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, *13*(1), 6793. https://doi.org/10.1038/s41467-022-34591-0

Padmakumar, V., & He, H. (2024). Does writing with language models reduce content diversity? https://doi.org/10.48550/arXiv.2309.05196

Pang, R. Y., Schroeder, H., Smith, K. S., Barocas, S., Xiao, Z., Tseng, E., & Bragg, D. (2025). Understanding the LLM-ification of CHI: Unpacking the impact of LLMs at CHI through a systematic literature review. In N. Yamashita, V. Evers, K. Yatani, X. (S.) Ding, B. Lee, M. Chetty, & P. Toups-Dugas (Eds.), *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)* (pp. 1–20). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713726

Peng, H., Liu, S., & Li, L. (2025). Evaluation of the quality of AI-generated scientific text under different types of cognitive complexity tasks. In G. Oliver, V. Frings-Hessami, J. T. Du, & T. Tezuka (Eds.), *Sustainability and empowerment in the context of digital libraries* (pp. 212–221). Springer Nature. https://doi.org/10.1007/978-981-96-0865-2_17

Platt, N., Luchs, E., & Nizamani, S. B. (2025). Data analysis for catching UX flaws in code: Leveraging LLMs to identify usability flaws at the development stage. In *2025 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, Raleigh, NC, USA (pp. 152–158). https://doi.org/10.1109/VL-HCC65237.2025.00024

Pope, T., Gilbertson-White, S., & Patooghy, A. (2025). Evaluating GPT-4's semantic understanding of obstetric-based healthcare text through Nurse Ruth. *ACM Transactions on Intelligent Systems and Technology*. https://doi.org/10.1145/3735647

Reinhard, P., Li, M. M., Fina, M., & Leimeister, J. M. (2025). Fact or fiction? Exploring explanations to identify factual confabulations in RAG-based LLM systems. In N. Yamashita, V. Evers, K. Yatani, & X. (S.) Ding (Eds.), *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)* (pp. 274–313). Association for Computing Machinery. https://doi.org/10.1145/3706599.3720249

Riihiaho, S. (2018). *Usability testing*. John Wiley. https://doi.org/10.1002/9781118976005.ch14
Ronanki, K., Cabrero-Daniel, B., & Berger, C. (2024). ChatGPT as a tool for user story quality evaluation: Trustworthy out of the box?. In P. Kruchten & Peggy Gregory (Eds.), *Agile processes in software engineering and extreme programming – Workshops (lecture notes in business information processing)* (pp. 173–181). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-48550-3_17
Rynes, S. L., Orlitzky, M. O., & Bretz, R. D., Jr. (1997). Experienced hiring versus college recruiting: Practices and emerging trends. *Personnel Psychology*, *50*(2), 309–339. https://doi.org/10.1111/j.1744-6570.1997.tb00910.x
Sahoo, P., Singh, A. K., Saha, S., Jain, V., Samrat, M., & Aman, C. (2025). *A systematic survey of prompt engineering in large language models: Techniques and applications*. https://arxiv.org/abs/2402.07927
Santolaria, M., Oliver-Solà, J., Gasol, C. M., Morales-Pinzón, T., & Rieradevall, J. (2011). Eco-design in innovation driven companies: Perception, predictions and the main drivers of integration. The Spanish example. *Journal of Cleaner Production*, *19*(12), 1315–1323. https://doi.org/10.1016/j.jclepro.2011.03.009
Sauro, J. (2010). *A practical guide to measuring usability: 72 Answers to the most common questions about quantifying the usability of websites and software*. Measuring Usability LLC.
Schade, A. (2013). *Making usability findings actionable*. https://www.nngroup.com/articles/actionable-usability-findings/
Schneiders, E., Seabrooke, T., Krook, J., Hyde, R., Leesakul, N., Clos, J., & Fischer, J. E. (2025). Objection overruled! Lay people can distinguish large language models from lawyers, but still favour advice from an LLM. In N. Yamashita, V. Evers, K. Yatani, X. (S.) Ding, B. Lee, M. Chetty, & P. Toups-Dugas (Eds.), *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery. https://doi.org/10.1145/3706598.3713470
Scholtz, J. (2000). Common industry format for usability test reports. In M. Tremaine (Ed.), *CHI '00 extended abstracts on human factors in computing systems (The Hague, The Netherlands) (CHI EA '00)* (p. 301). Association for Computing Machinery. https://doi.org/10.1145/633292.633470
Schwind, V., Resch, S., & Sehrt, J. (2023). The HCI user studies toolkit: Supporting study designing and planning for undergraduates and novice researchers in human-computer interaction. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, & A. Peters (Eds.), *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)* (pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3544549.3585890
Shen, L., Shi, Q., Kuang, E., Qiu, L., Zhou, S., Hui, P., & Fan, M. (2026). MultiUX: A human-AI collaborative tool to facilitate multiple usability test video analyses. *International Journal of Human–Computer Interaction*, *0*(0), 1–26. https://doi.org/10.1080/10447318.2025.2605526
Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., Schärli, N., & Zhou, D. (2023). Large language models can be easily distracted by irrelevant context. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning (Honolulu, Hawaii, USA) (ICML'23)* (pp. 31210–31227). JMLR.org
Shin, S., Oh, J., & Lee, S. (2025). Can LLMs see what I see? A study on five prompt engineering techniques for evaluating UX on a shopping site. In N. Yamashita, V. Evers, K. Yatani, & X. (S.) Ding (Eds.), *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)* (pp. 125–127). Association for Computing Machinery. https://doi.org/10.1145/3706599.3720079
Shokrizadeh, A., Bahati Tadjuidje, B., Kumar, S., Kamble, S., & Cheng, J. (2025). Dancing with chains: Ideating under constraints with UIDEC in UI/UX design. In N. Yamashita, V. Evers, K. Yatani, X. (S.) Ding, B. Lee, M. Chetty, & P. Toups-Dugas (Eds.), *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)* (pp. 1106–1123). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713785
Soure, E. J., Kuang, E., Fan, M., & Zhao, J. (2022). CoUX: Collaborative visual analysis of think-aloud usability test videos for digital interfaces. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 643–653. https://doi.org/10.1109/TVCG.2021.3114822
Subbiah, M., Zhang, S., Chilton, L. B., & McKeown, K. (2024). Reading subtext: Evaluating large language models on short story summarization with writers. *Transactions of the Association for Computational Linguistics*, *12*, 1290–1310. https://doi.org/10.1162/tacl_a_00702
Subramonyam, H., Thakkar, D., Ku, A., Dieber, J., & Sinha, A. K. (2025). Prototyping with prompts: Emerging approaches and challenges in generative AI design for collaborative software teams. In N. Yamashita, V. Evers, K. Yatani, X. (S.) Ding, B. Lee, M. Chetty, & P. Toups-Dugas (Eds.), *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery. https://doi.org/10.1145/3706598.3713166
Sun, G., Zhan, X., & Such, J. (2024). Building better AI Agents: A provocation on the utilisation of persona in LLM-based conversational agents. In M. Dubiel, L. A. Leiva, J. Trippas, J. Fischer, & I. Torre (Eds.), *Proceedings of the 6th ACM Conference on Conversational User Interfaces (Luxembourg, Luxembourg) (CUI '24)* (pp. 35–36). Association for Computing Machinery. https://doi.org/10.1145/3640794.3665887
Swaroop, S., Buçinca, Z., Gajos, K. Z., & Doshi-Velez, F. (2025, March 2025). Personalising AI assistance based on overreliance rate in AI-assisted decision making. In T. Li, F. Paternò, K. Väänänen, L. Leiva, D. Spano, & K.

Verbert (Eds.), *Proceedings of the 30th International Conference on Intelligent User Interfaces* (pp. 1107–1122). Association for Computing Machinery. https://doi.org/10.1145/3708359.3712128

Tan, K., Yao, J., Pang, T., Fan, C., & Song, Y. (2025). ELF: Educational LLM framework of improving and evaluating AI generated content for classroom teaching. *Journal of Data and Information Quality*, *17*(3), 1–23. https://doi.org/10.1145/3712065

Tian, Y., Huang, T., Liu, M., Jiang, D., Spangher, A., Chen, M., May, J., & Peng, N. (2024). Are large language models capable of generating human-level narratives?. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17659–17681). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.978

UXBoost. (2021). *How to write a usability testing report*. https://www.uxboost.com/post/how-to-write-a-usability-testing-report

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Woodmass, R. (2020). *Creating usability reports from usability test findings*. https://medium.com/thinking-design/creating-usability-reports-from-usability-test-findings-93fceceac571

Xiang, W., Zhu, H., Lou, S., Chen, X., Pan, Z., Jin, Y., Chen, S., & Sun, L. (2024). SimUser: Generating usability feedback by simulating various users interacting with mobile applications. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24)*. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642481

Xtensio. (2023). *How to write a usability testing report (with templates and examples)*. https://xtensio.com/how-to-write-a-usability-testing-report/Section:Product Management

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, *51*(4), 1373–1418. https://doi.org/10.1162/COLI.a.16

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In E. Celis, S. Ruggieri, L. Taylor, & G. Zanfir-Fortuna (Eds.), *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20)* (pp. 295–305). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372852

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M., & Lin, M. (2023). On evaluating adversarial robustness of large vision-language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '23)*. Curran Associates.

Zhou, P., Wang, L., Liu, Z., Hao, Y., Hui, P., Sasu, T., & Jussi, K. (2024). *A survey on generative AI and LLM for video generation, understanding, and streaming*. https://arxiv.org/abs/2404.16038

About the authors

Emily Kuang is an Assistant Professor at York University. She leads the COCOA (CO-designing COllaborations with AI) Lab, which explores how to design human-AI collaborative tools to enhance productivity, creativity, and accessibility. She completed a PhD in Computing and Information Sciences from Rochester Institute of Technology in 2025.

Luyao Shen received an MA in Artistic Design from the University of Science and Technology Beijing in 2020. She is conducting her PhD in Computational Media and Arts (CMA) at HKUST (GZ). Her research interests include human–computer interaction and user experience design.

Ehsan Jahangirzadeh Soure received an MMath degree in Computer Science from the University of Waterloo in 2023. He is currently a Software Engineer at Snowflake. His research interests include human–computer interaction, information visualization, and artificial intelligence.

Mingming Fan is an Associate Professor at Hong Kong University of Science and Technology (Guangzhou) and Hong Kong University of Science and Technology. He directs the Augmenting People *via* Empowering X (APEX) Group, which explores the design, interaction, and applications of Human-AI Interaction, Accessible Computing, and VR/AR/MR.

Kristen Shinohara is an Associate Professor in the School of Information at the Rochester Institute of Technology and the Director of the Center for Accessibility and Inclusion Research (CAIR) Lab. Her research focuses on the design of technologies usable by people with disabilities and teaching accessibility in computing.