EvAlignUX: Advancing UX Evaluation through LLM-Supported Metrics Exploration

Qingxiao Zheng
Computer Science and Engineering
University at Buffalo
Buffalo, New York, USA
qingxiao@buffalo.edu

Minrui Chen
Informatics
University of Illinois Urbana-Champaign
Champaign, Illinois, USA
minruic2@illinois.edu

Pranav Sharma
Computer Science
University of Illinois Urbana-Champaign
Champaign, Illinois, USA
pranav24@illinois.edu

Yiliu Tang
Informatics
University of Illinois at Urbana-Champaign
Champaign, Illinois, USA
yiliut2@illinois.edu

Mehul Oswal
Computer Science
University of Illinois Urbana-Champaign
Champaign, Illinois, USA
mehuljo2@illinois.edu

Yiren Liu
Informatics
University of Illinois Urbana-Champaign
Champaign, Illinois, USA
yirenl2@illinois.edu

Yun Huang
Information Sciences
University of Illinois at Urbana-Champaign
Champaign, Illinois, USA
yunhuang@illinois.edu

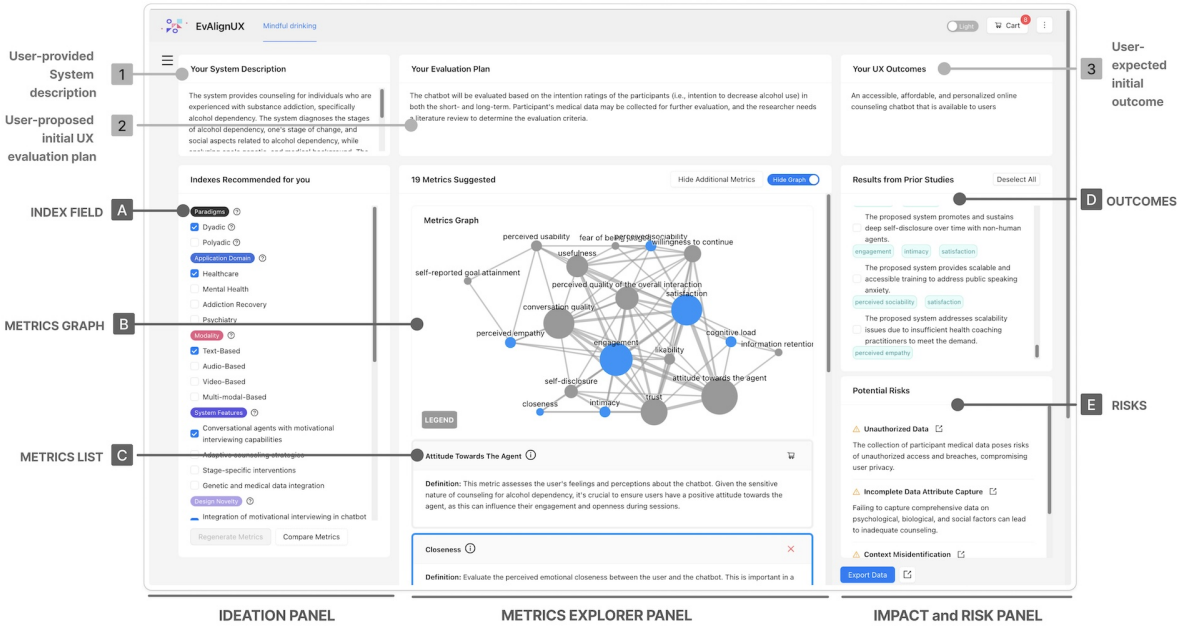


Figure 1: EvAlignUX comprises three main panels, each powered by large language models (LLMs): the project IDEATION Panel (left), the METRICS EXPLORER Panel (center), and the IMPACT AND RISK Panel (right). After the user provides 1) a project description, 2) an initial evaluation plan, and 3) expected outcomes, they can utilize five key features–`INDEX` [A], `METRICS GRAPH` [B], `METRICS LIST` [C], `OUTCOMES` [D], and `RISKS` [E]–to refine the UX evaluation plan and hypothesize potential results.

Abstract

Evaluating UX in the context of AI's complexity, unpredictability, and generative nature presents unique challenges. *How can we support HCI researchers to create comprehensive UX evaluation plans?* In this paper, we introduce EvAlignUX , a system powered by large language models and grounded in scientific literature, designed to help HCI researchers explore evaluation metrics and their relationship to research outcomes. A user study with 19 HCI scholars showed that EvAlignUX improved the perceived quality

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

CHI '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714045

and confidence in UX evaluation plans while prompting deeper consideration of research impact and risks. The system enhanced participants' thought processes, leading to the creation of a *"UX Question Bank"* to guide UX evaluation development. Findings also highlight how researchers' backgrounds influence their inspiration and concerns about AI over-reliance, pointing to future research on AI's role in fostering critical thinking. In a world where experience defines impact, we discuss the importance of shifting UX evaluation from a *"method-centric"* to a *"mindset-centric"* approach as the key to meaningful and lasting design evaluation.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; *Empirical studies in HCI*; **HCI design and evaluation methods**.

Keywords

User experience, Evaluation, Human-AI Interaction, Large Language Models, Usability

ACM Reference Format:

Qingxiao Zheng, Minrui Chen, Pranav Sharma, Yiliu Tang, Mehul Oswal, Yiren Liu, and Yun Huang. 2025. EvAlignUX: Advancing UX Evaluation through LLM-Supported Metrics Exploration. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3706598.3714045

1 Introduction

The rapid integration of AI into interactive systems presents a growing challenge for user experience (UX) evaluation [12, 88]. Traditional usability evaluation guided by standard frameworks such as the System Usability Scale (SUS) [6], User Experience Questionnaire (UEQ) [82], NASA Task Load Index (NASA-TLX) [36], falls short in capturing the complex, dynamic interaction between humans and AI [48]. As AI systems become more context-aware and capable of generating responses in real-time, UX researchers are confronted with new questions about trust, emotional engagement, and ethical considerations [2, 55, 68, 85].

Addressing these new considerations is especially challenging for systems research, where a range of evaluation methods, often borrowing and combining techniques from software engineering, design, and usability evaluation, is triangulated to evaluate the UX of interactive computing systems [59]. Current UX evaluation methods, e.g., the Post-Study System Usability Questionnaire (PSSUQ) [54], the Subjective Mental Effort Questionnaire (SMEQ) [40], while robust for traditional systems, are inadequate for AI-powered interfaces. Researchers are striving to develop new frameworks that take into account ethical considerations, trust dynamics, intimacy, and multimodal behaviors, to adequately guide scholars and practitioners in developing their UX evaluation plans for human-AI interaction research, e.g., [57, 110]. Given the plethora of well-documented evaluation methods, heuristics, and metrics (e.g. [49, 63, 95]), much less is known about *How researchers select appropriate metrics to conduct a comprehensive evaluation of a system.*

Also, despite the availability of various UX evaluation methods [14, 45, 74, 77, 95], there is a lack of practical tools to guide researchers in developing comprehensive evaluation plans. Research has shown that incorporating evaluation metrics as part of the design materials in the early stages can improve the clarity in the design of human-AI interaction research [109]. UX is frequently regarded as a rigid set of methods, leading to calls for research and education that encourage framing UX as "*more of a mindset than a method*" within the community [34]. However, with the rapid advancement of AI models, the literature on AI-mediated multi-stakeholder interactions has expanded exponentially. This growth presents challenges for novice researchers in acquiring knowledge of relevant evaluation metrics, while seasoned researchers also face increasing pressure to stay current with the latest developments.

To bridge this gap, we designed, implemented, and evaluated EvAlignUX , an interactive system that supports novice and experienced HCI researchers in developing UX evaluation plans for their research projects. EvAlignUX is powered by the generative capabilities of Large Language Models (LLMs) to facilitate the exploration of UX evaluation ideas, as LLMs have demonstrated their potential for the ideation of scientific research by processing vast amount of information, generating context-aware insights, retrieving literature, and brainstorming [21, 29, 58]. Specifically, as shown in Fig. 1, EvAlignUX comprises three primary panels: the `Project Ideation` Panel, the `Metric Explorer` Panel, and the `Outcomes and Risks` Panel. These panels allow users to explore UX metrics, potential outcomes, and potential risks associated with human-AI interaction research. In an evaluation with 19 HCI scholars, EvAlignUX led to significant improvements in the clarity, concreteness, and confidence of their revised UX evaluation plans. Participants were able to better justify selected UX metrics, self-critique their plans, and raise awareness of potential risks using the system features.

This work makes several significant contributions to the HCI community. **First**, EvAlignUX is a novel system designed to support the ideation process of developing UX evaluation plans. Powered by LLMs, it enables users to explore UX metrics and their relationships with potential research outcomes. The unique visual interface of the metrics graph supports users in examining the connections between different UX metrics. To the best of our knowledge, EvAlignUX fills the gap in tool support available to researchers by providing insights into the research literature on human-AI evaluation metrics and facilitating better decision-making regarding UX evaluation. **Second**, our user studies demonstrate the effectiveness of EvAlignUX in helping users achieve significantly greater clarity and feasibility in their proposed UX evaluation plans. Our findings revealed that researchers' backgrounds impacted their perceived levels of inspiration and their concerns about over-reliance on AI. They also integrated the system features into their thought processes differently, focusing on varied utilities such as the comprehensiveness of self-critique and the efficiency of implementing the ideas. **Third**, our analysis of participants' think-aloud protocols and perceived value of different features produced a comprehensive *Question Bank for UX Evaluation Development*, highlighting the specific questions researchers asked at different steps of UX evaluation development and identifying how AI provides valuable support such as envisioning the research impact, assessing the risks, and

justifying metric selection. **Lastly**, our findings prompt a critical discussion on shifting the focus from method to mindset in future research on UX evaluation.

2 Related Work

In this section, we review the literature on (1) current UX evaluation approaches and (2) the potential of leveraging LLMs to provide tool support.

2.1 UX Evaluation of HCI System Research

Developing a UX evaluation plan can be challenging, as it is often unclear what constitutes "evaluation" and which metrics are appropriate [38, 50]. Hillman et al. [38] elaborated on several challenges involved in developing a solid evaluation plan, such as selecting metrics that accurately reflect UX and research goals, choosing data collection methods, establishing collection protocols, and ensuring consistency across products and organizations. Additionally, long-term effects and social or cultural barriers further complicate the process [14, 38, 95]. Several foundational conceptual frameworks serve as the backbone for UX evaluation, including Nielsen's usability engineering [65] and heuristic evaluation [64], Hassenzahl's UX framework [37], and Norman's emotional design theory [67]. These frameworks emphasize multi-dimensional assessments, moving beyond performance metrics to encompass user perceptions, attitudes, and context of use.

To measure users' perceptions quantitatively, researchers often use established surveys, such as the System Usability Scale (SUS), the User Experience Questionnaire (UEQ), the Post-Study System Usability Questionnaire (PSSUQ), the NASA Task Load Index (NASA-TLX), the Subjective Mental Effort Questionnaire (SMEQ), and the Customer Satisfaction (CSAT) [6, 36, 54, 75, 82]. PULSE metrics (e.g., page views, uptime, latency, etc.) are also commonly used to track user behavior [78]. However, these metrics are often "low-level or indirect measures of user experience" [78]. Several evaluation frameworks relating to user-centered metrics have been introduced to the HCI and UX domains, such as the Software Development Productivity Framework [81] and Google's HEART framework [78]. For instance, the HEART framework, which includes Happiness, Engagement, Adoption, Retention, and Task Success, was introduced to measure web-based applications at scale [78].

Beyond selecting appropriate metrics to evaluate systems, researchers have also developed methods and tools to enhance survey and interview question design. For example, the Question Appraisal System (QAS-99) [104] and Willis's cognitive interviewing methodology [105] provide structured approaches for detecting and refining data collection items, ensuring that questions are clear, unbiased, and aligned with research goals. Similarly, QUAID [33] offers diagnostic support for identifying comprehension issues in survey questions, while Survey Sidekick [39] assists in creating scientifically sound surveys. Semi-automated approaches [86] further aid in generating relevant and effective questions. Incorporating such methods into UX evaluation can yield instruments that better capture users' experiences, perceptions, and emotional responses.

As HCI rapidly shifts toward AI-driven systems, recent studies in UX research highlight several emerging trends and challenges. A group of HCI researchers aims to identify and propose new evaluation metrics for both usability and user-centered evaluation in the age of AI. For example, Zheng et al.'s literature review of human-AI interaction research in HCI, which included user evaluations, reported a suite of metrics that were collected using surveys, system logs, and interviews to collect [110] and argued the importance of including more user-centered metrics (e.g., sociability, socio-emotional, and AI anthropomorphism metrics) with AI's evolving role in human social interactions. Likewise, Faruk et al. proposed new subjective scales for measuring the UX of voice agents [24] and conversational agents [25]. In other domains, such as healthcare, Abbasian et al. summarized some foundation metrics for evaluating the effectiveness of healthcare conversations powered by AI [1].

However, there is a lack of tools to help researchers understand the evaluation metrics literature and design a comprehensive evaluation plan for human-AI interaction research. As Yang et al. [107] systematically reported that an AI system's uncertainty and complexity can lead to unpredictable errors that damage UX and result in undesired societal impact.. Gray's finding, encapsulated in the phrase *"more of a mindset than a method"* also highlighted that UX practitioners often follow a codified method rather than adopting a "mindset" that aligns with their design process, practice context, and the specific design problem [34]. All these works suggest that new tools and approaches are needed to support researchers and practitioners in developing their UX evaluation plans.

2.2 The Potential of LLM-Supported Ideation

Large Language Models (LLMs) can process vast amounts of multimodal information and generate context-aware, coherent text, making them ideal tools for evaluation-related tasks in various domains. A group of researchers developed novel interaction prototypes using LLMs for evaluating generative outputs [19, 47, 83]. For example, Kim et al. developed EvalLM [47], an interactive system for refining prompts by evaluating multiple outputs on user-defined criteria. This helps developers compare prompt outputs to diagnose weaknesses in their prompts. Regarding LLMs' support for scientific tasks, recent studies have showcased LLM-powered methods for retrieving and recommending relevant literature [84], reviewing and critiquing research drafts, writing research manuscripts, and augmenting research ideation and brainstorming for scientific writing by leveraging scientific evidence and large datasets [66]. Recent advancements in HCI and AI research also have also aimed to support UX practitioners with AI-enabled tools for usability testing (e.g., LLMs are employed to help human evaluators identify additional usability issues in usability test video recordings). All these research highlights the significant potential of LLMs to support research ideation and provide valuable feedback for task evaluation. However, little is known about how researchers with varying levels of experience can benefit from using LLM-supported tools to develop their UX evaluation plans.

3 EvAlignUX

To address the research gap, EvAlignUX was designed as a tool to aid researchers in exploring evaluation metrics for developing UX plans for Human-AI research. The system was developed through an iterative process, informed by both the literature review

[14, 41, 77, 95, 110] and input from a focus group. Leveraging the authors' expertise in HCI research and practices, we began with an initial process of developing a UX evaluation plan, e.g., proposing a research topic, identifying relevant UX metrics for evaluation, and framing the expected research contribution. The design was refined over two months through multiple weekly feedback sessions with an HCI research group composed of, consisting of nine graduate students and faculty members from the HCI, Informatics, and Computer Science departments (n=9). All of the participants have worked on at least one HCI project where they designed or developed systems. Their applications covered diverse contexts: vision-based question-answering, ideation and writing, accessibility, immersive technologies, and intelligent assistants. Based on the UX evaluation challenges reported in prior literature [38, 50], and after piloting the initial pipeline with this group, we established three design goals (DG). We then iteratively searched for additional literature and sought feedback through weekly focus group meetings.

- **DG1:** Recommend UX metrics relevant to a project;
- **DG2:** Explain how the metrics are used in existing literature;
- **DG3:** Inform researchers about potential risks associated with their design and evaluation plans.

3.1 System Design

Based on the DG2, EvAlignUX comprises with three main panels, as shown in Figure 1. To illustrate interactions in EvAlignUX , we present a walk-through with Sarah, a UX researcher designing an AI counseling chatbot for alcohol addiction. The user journey is likened to a "shopping" process, where Sarah selects ideal evaluation plans for her proposed system. See Figure 2 for **User Journey 1**.

*3.1.1 Project Ideation Panel (**DG1**).* This panel extracts relevant indexes to categorize the work within the existing literature. In the rapidly evolving field of UX research, evaluators often face challenges in understanding the context and relevance of newly proposed systems [50]. To enhance the UX evaluator's understanding of the context in which a system operates and its significance within the broader field of UX research, we built on categorizations from prior literature reviews and key guiding literature for UX in AI design practices [45, 56, 95, 110]. We leverage these established frameworks to integrate critical factors into the system output. This output is designed to automatically map the proposed system based on the provided narrative, into categories including paradigms, application domains, modalities, system features, design novelty, design methods, human-AI relationship types, stakeholders, social scales, and theoretical frameworks. See Appendix 10 for details. See Figure 3 for **User Journey 2**.

*3.1.2 Metric Explorer Panel (**DG1, DG2**).* This panel recommends relevant evaluation metrics and related literature while enabling users to explore additional metrics through both a list-view and graph-view interactions, helping them formulate a comprehensive evaluation plan. See Figure 4 for **User Journey 3**.

Prior literature has highlighted that users are often excluded from the process of defining evaluation criteria for a system, with most participants consulted only at the end of the design process to assess the system's effects, rather than being actively involved in setting these criteria. This exclusion contradicts the goal of democratizing the design process, which seeks to empower all stakeholders to contribute meaningfully throughout the design life-cycle [10, 109]. Inspired by this challenge, we propose a metric recommendation feature that integrates UX evaluation metrics into design materials as inputs for design or evaluation. Given that graph-based interaction is a commonly used and effective approach for recommendation systems [4], we propose incorporating a graph-view alongside the traditional list-view. This graph-view will allow users to visualize UX metrics and their related metrics, fostering inspiration and enabling more informed filtering and selection of metrics. Furthermore, prior literature outlines various methods for collecting metrics, such as system logs, surveys, and interviews [110]. To enhance user understanding, we will provide context and examples of how these metrics have been derived and applied.

*3.1.3 Outcomes and Risks Panel (**DG2, DG3**).* Assists users in framing their research contributions based on previous studies and identifying potential pitfalls by referencing databases of prior AI risks and incidents. Technologies have positively transformed society, but they have also led to undesirable consequences not anticipated at the time of design or development. Previous HCI research has explored using insights from past undesirable consequences to help researchers and practitioners gain awareness of potential adverse effects and reflect on their own experiences with technology [72]. Inspired by this, we propose a panel for generating related research outcomes and risks based on existing research and real-world cases. We sourced case studies from a database indexing real AI incidents documented by journalists — the AI, Algorithmic, and Automation Incident and Controversy (AIAAIC) repository, which previous papers have used [51]. This approach aims to strengthen UX evaluation metrics by incorporating literature-based research outcomes and relevant risks grounded in real-world incidents.

> **User Journey 4**: As Sarah adds and removes metrics from her shopping cart, she notices that the results in the Outcome feature update dynamically. She selects outcomes from prior studies that catch her interest. Then, she explores potential risks associated with her proposed UX evaluation plan by clicking on the Risk feature. *"All set!"* said Sarah. She then exports the data and conducts a final review of her generated UX evaluation plan and UX outcomes report.

3.2 System Backend and Implementation

In this section, we describe the process of preparing the data repository, system backend, and implementation, as well as the techniques used for prompting and making recommendations. The frontend of the system is implemented using ReactJS[1], TypeScript and CSS. The backend uses Python with FastAPI[2] as the RESTful API server framework. The data is represented as a Knowledge Graph using the Neo4j[3] Graph Database to model the complex relations among research papers [43, 94, 102]. We use the GPT-4o model as our LLM engine and set the temperature to 0.7 for all components. The embeddings in the system are generated using text-embedding-3-small,

[1] https://github.com/facebook/react
[2] https://github.com/tiangolo/fastapi/
[3] https://github.com/neo4j

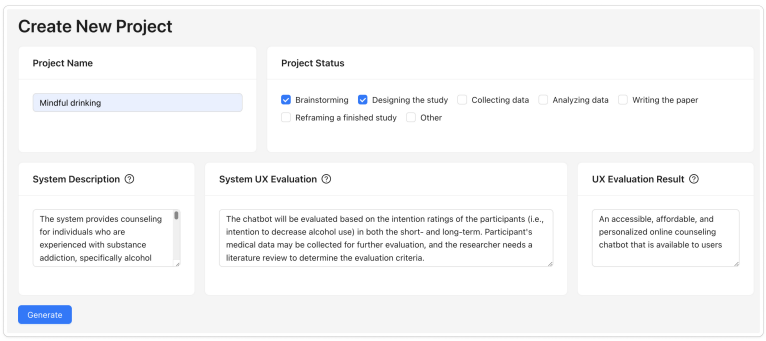


Figure 2: EvAlignUX Landing Page

User Journey 1:
Upon landing on EvAlignUX , Sarah inputs her project details. She names the project *"Mindful Drinking"*, selects "brainstorming" and "designing the study" as the current project statuses, and provides a project description, evaluation plans, and details the expected outcomes. After entering all details, she clicks "Generate" and proceeds to the main interaction page.

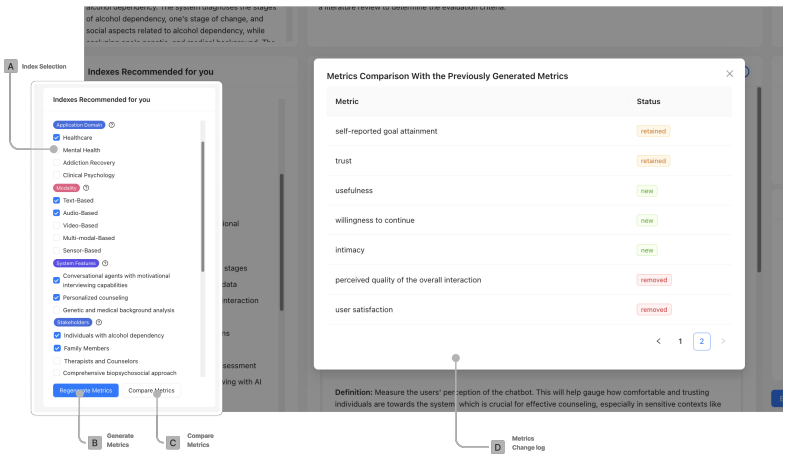


Figure 3: Project Ideation Panel

User Journey 2:
Sarah views a list of indexed recommendations from EvAlignUX , noting that some are selected while others are not. She realizes that "Addiction Recovery" isn't checked under "Application Domain," so she selects it (A). She also checks "Family Members" as the evaluation target stakeholder, which she had previously overlooked. After making these changes, she clicks "Regenerate" (B) followed by "Compare Metrics" (C). This process updates her list, showing added, retained, and removed metrics (D), sparking her curiosity about the new metrics.

with FAISS[4] as the vector database for data retrieval. Additionally, Langchain[5] is used for prompting the GPT-4o model.

3.2.1 Data Preparation. In this section, we present how the scientific database was prepared. First, we established an **initial repository of evaluation metrics** based on the definitions specified in prior literature (e.g., [110]). Soon, we recognized that the literature review focused on conversational AI interactions; thus two authors expanded the UX evaluation metrics to include interactions in other modalities by incorporating additional metrics that were widely applied beyond conversational interactions [41, 95]. For example, we expanded the repository by incorporating metrics related to ethics and risks, such as "communication fairness," "user control," and "privacy concerns," which address broader dimensions of responsible interaction design. We do not claim that the repository is complete or exhaustive. It will continue to grow as new HCI literature and UX evaluation metrics emerge. Definitions for all metrics, including the newly added ones, were either directly sourced from these surveys or developed collaboratively by the authors to ensure completeness and coherence.

After building the repository of evaluation metrics, we created a **paper database** from the same survey papers and utilized GPT-4o [69] to generate more fine-grained information associated with the evaluation metrics. We prompted GPT with the full text of each paper and asked it to provide a narrative of the system described in the papers, along with the challenges or the research gaps that the papers addressed. We provided index definitions and asked GPT to generate a list of ten descriptive indices based on the content of the papers. These descriptive indices include: interaction paradigms,

[4]https://github.com/facebookresearch/faiss
[5]https://github.com/langchain-ai/langchain

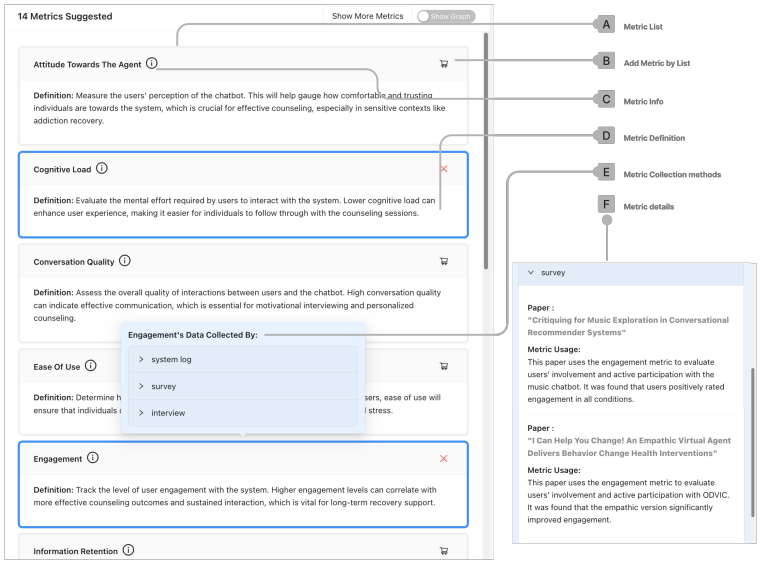


Figure 4: Metrics Explorer Panel

User Journey 3:
Sarah starts by reviewing the suggested metrics in the list (A) and is intrigued by the "goal attainment", defined as assessing users' perceptions of achieving their recovery goals through the system (D). She clicks the metric info button (C) to explore how this metric is measured—through surveys, system logs (E), and interviews—and reviews related papers on each method's application and findings (F). Interested, she adds it to her shopping cart (B) and uses the Metric Graph feature to explore additional metrics.

application domain, modality, system features, design novelty, design methods, human-AI relationship types, stakeholders involved, social scale, and theoretical frameworks used. The definitions for these indices can be found in Appendix A.1.

Between the metrics and the papers in the database, we also created usage information for each metric in the relevant papers. For example, a metric usage is formatted as follows: *"This paper uses the [X] metric to evaluate users' [Y] towards [Z technology]. It was found that [S findings related to Y]." [X]* represents the metric identified in the paper, *[Y]* is the aspect of user interaction being evaluated, *[Z]* is the technology or system being evaluated, and *[S]* is the findings related to the user interaction being evaluated.

We then crawled metadata and appended it to each paper. The metadata includes the paper's UID, title, authors, abstract, publication date, UIDs of papers that cited it, keywords, publication venue, publisher, and authors' affiliations. Besides, we also leverage the existing AI Incident Database as a repository for risks [60] and extracted relevant incidents and risks.

EvAlignUX 's backend recommends UX metrics based on an initial *System Description*. After a series of user interactions, it generates a new Evaluation Plan and UX Outcome using an LLM and data from research papers via a Knowledge Graph, which represents complex entity relations derived from source documents.

3.2.2 Index Generation and derivation of Knowledge Graph. Our system, EvAlignUX , generates a list of indexes using the initial system description provided by the user. The indexes are intended to standardize the system description into fixed categories. Indexes are generated by providing the System Description, examples of other generated indexes, and the corresponding definitions of each index as input to the LLM. Subsequently, the system recommends three additional index values for each category. These indexes and the initial system description are used to retrieve a list of metrics from our Knowledge Graph.

The Knowledge Graph, built using the Neo4j Graph Database, serves as the foundation for recommending metrics. We structured our data as a graph to capture the intricate relationships within our dataset [43, 94, 102], enabling efficient information extraction and retrieval of global contextual information [22]. The backing dataset of research papers is organized into distinct clusters of communities [28] based on their index similarity, citations, and common metrics, using the Louvain Community detection algorithm [17]. The graph consists of three different types of nodes, each characterized by distinct properties. Table A.2 in the Appendix provides a detailed overview of the distinct nodes and their associated properties. The three unique nodes are as follows:

- **Paper Node:** This node contains details about each unique research paper.
- **Metric Node:** The node includes attributes specific to a UX metric. This node is attached to the Paper and the Outcome node via several edges. The node is shared between two paper nodes if they measure the metric.
- **Outcome Node:** It is attached to the Paper Node and the Outcome Node via several edges. The node is shared between two Paper Nodes if they measure the same metric. Since the metric node could be shared between multiple paper nodes, this node stores the unique information of a metric corresponding to each metric as used in a research paper.

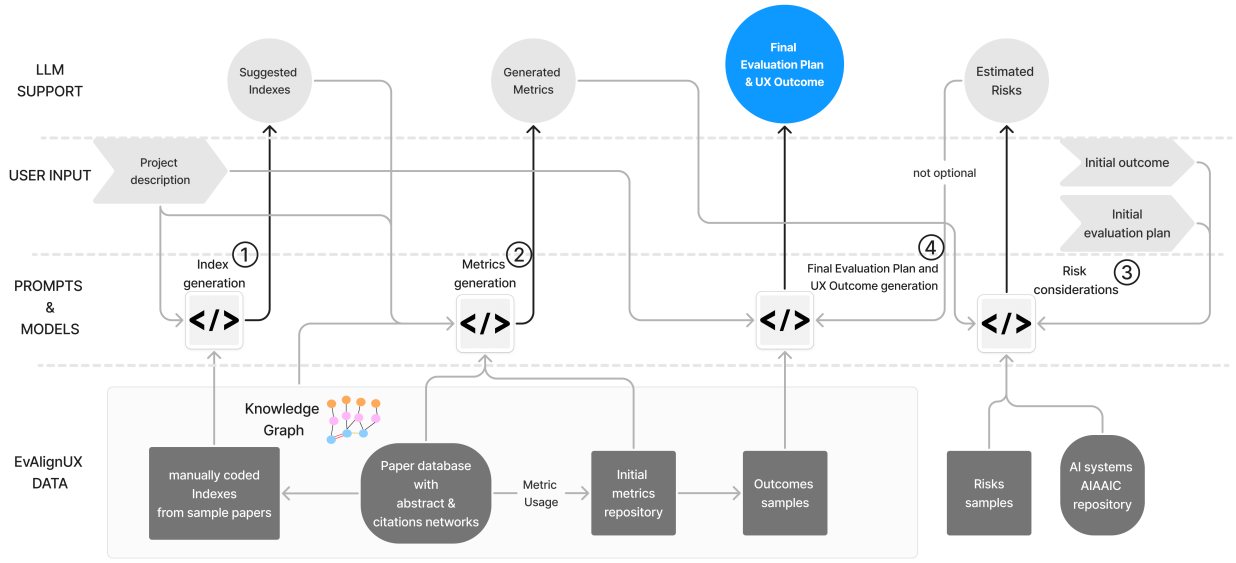


Figure 5: In EvAlignUX , the user provides a project description, an initial evaluation plan, and an initial outcome. ① EvAlignUX uses the project description to generate a list of indexes using the index generation prompt and a manually coded index; ② Based on either system-suggested or user-selected indexes and the initial project description, metrics are extracted from the Knowledge Graph and filtered through a metrics generation prompt to recommend UX metrics; From the user-selected metrics, a group of potential outcomes appears using information from the Knowledge Graph; ③ Additionally, based on the user's project description, initial evaluation plan, initial outcome, and information from risk samples and AI Risk repository, several risks are displayed; ④ A UX evaluation plan and UX Outcome are generated based on the selected UX Metrics, project description, initial evaluation plan, initial outcome, and risks.

All nodes are unique and do not duplicate. Furthermore, weighted edges exist between nodes. For instance, there exist three edges between each pair of paper nodes, namely, [SHARED_METRIC], [CITES], and [CITED_BY]. The [SHARED_METRIC] edge exists between two paper nodes if they share at least one common metric. Similarly, if any paper A cites another paper B, there exists a [CITES] edge from paper node A to paper node B, and a [CITED_BY] edge from paper node B to paper node A. Edges of type [CITES] and [CITED_BY] are assigned relatively higher weight, compared to [SHARED_METRIC] edge. This is achieved by multiplying the edge weight with a constant value, due to stronger relation among the papers because of the citation relation, based on an increase in the *modularity score* as discussed in [8]. The weight assigned to these edges is influenced by the relative importance of each index category in evaluating the similarity between research papers. Index categories that indicate greater similarity receive higher weights. These weights are then multiplied by the similarity score between index values of the categories for each paper, contributing to the overall edge weight. Additionally, each Paper node is connected to a Metric node via a non-weighted edge of type [HAS_METRIC], which is then connected to an Outcome node via another non-weighted edge of type [HAS_OUTCOME].

3.2.3 Metrics Recommendation. The provided system description and its corresponding generated indexes are used to extract metrics from the Knowledge Graph. The system description and indexes are vectorized, and the most similar paper is determined using *cosine similarity* on the existing Knowledge Graph. The community for the most similar paper is determined from the aforementioned Knowledge Graph and consequently, all metrics linked to the paper nodes belonging to that community are extracted. These metrics, along with the user system description and generated indexes, are used to further filter metrics, by prompting an LLM. Additionally, each metric is further classified into three metric collection methods, based on its usage in previous research papers, using the Knowledge Graph. These metrics are displayed in a list view and a graphical view to the user.

3.2.4 Potential risks and metric outcomes. Each metric in the Knowledge Graph is associated with outcomes corresponding to the research paper. Subsequently, these results are retrieved from the Knowledge Graph based on the final metrics list and displayed to the user so that they can filter them based on relevance to the user's system description. Each outcome is specific to a research paper

and is mapped to all metrics associated with that paper. Consequently, these outcomes are further considered when generating the Potential Evaluation Plan and the Potential UX Outcome.

In addition to the recommended metric outcomes, our system generates potential risks associated with the user system using the AIAAIC repository. We extract the top three incidents that have a *Euclidean distance* of less than 0.5 by performing a vector-based similarity search, using *FAISS* between the user system description and the AI system description present in the AI Incident dataset. Subsequently, we extract risks associated with these AI systems, and provide them along with the user system description to a LLM to filter the most relevant risks for the system. The identified risks are contextualized based on the initial system description, and corresponding URLs for related incidents are also provided to the user.

Once the metrics have been finalized by the user, along with the corresponding metric outcomes, a Potential Evaluation Plan and Potential UX Outcome are recommended to the user. The Potential Evaluation Plan includes detailed instructions on the usage of each metric and the methods for measuring them based on the user's system description and initial evaluation plan. The potential evaluation plan is generated by providing the user's system description, a list of selected metrics, the specified metric outcomes, and the methods previously used to evaluate the UX metrics to an LLM, and then prompting the LLM to generate the plan. The Potential UX Outcomes are generated by considering the list of selected metrics, the desired results from the metric outcomes, the initial System Description, and the risks associated with the system.

3.3 Technical Evaluation

We conducted a small-scale evaluation to compare EvAlignUX 's understanding of evaluation metrics with human understanding and to assess the quality of its metric suggestions based on a given narrative. Figure 6 illustrates the process used to extract UX metrics applied in HCI publications and the method used to evaluate the accuracy of metric recommendations with EvAlignUX .

3.3.1 Evaluating Metrics Extraction. To assess GPT's capability in extracting evaluation metrics in research papers from a predefined list of candidate metrics, we adopted a similar strategy as in [72] and randomly selected 15 papers from our initial pool of research articles. The evaluation metrics used in these papers were extracted using our GPT prompt. One author reviewed each identified metric and rated it from three aspects: whether the metric was actually measured, whether the metric was in the candidate list, and whether the reasoning or citation provided makes sense. A second author verified these assessments. A metric was considered accurate if it fulfilled the above requirements.

We found that out of the 102 metrics identified in the 15 papers, 94% (96/102) were indeed measured, 92% (94/102) were in the candidate list, and 88% were accurate (90/102), noting that two metrics overlapped between the first and second criteria. 100% (102/102) of the reasoning and citations made sense. Overall, the results suggest that with our prompt, GPT's ability to identify metrics is largely reliable.

A common source of hallucination is that the metric extracted was not actually measured in the evaluation section of the paper (i.e., hallucination [42]). More specifically, there were three types of hallucination: a) GPT inferred metrics based on descriptions from the paper, but these metrics were not explicitly measured in the evaluation process; b) GPT synthesized metrics from the results, but these metrics were not explicitly measured in the evaluation process; and c) GPT made recommendations for a metric that should have been used but was not actually used, based on the context of the paper. The case for (a) and (b) was often that GPT often selected a statement or conclusion from a qualitative result described in the paper. However, such qualitative results were synthesized from interviews or surveys, and therefore the identified metric was not explicitly measured in the paper. For example, GPT identified "attitude towards the agent" as an evaluation metric from [97], quoting *"students made positive comments about the interaction with CAs related to the perceived level of enjoyment and interaction."* While attitude towards the agent was not explicitly measured in the paper, it was described in the findings and was thus misidentified. For (c), an example is GPT identifying "privacy concern" as a metric from [53], quoting *"because chatbot use can result in users' self-disclosure of sensitive topics"* from the discussion section. Although it makes sense to measure "privacy concern" based on the context of potential self-disclosure of sensitive information, this metric was not explicitly measured in the paper, so it was marked as inaccurate.

It was possible that the extracted metric was not in the list of candidate metrics. There were two reasons: (a) GPT reworded the metric in the candidate list while preserving its meaning, and (b) GPT invented a metric that was not included in the candidate list. For (a), one example is GPT identifying "utterance length" from [92], while the candidate list actually uses "length of utterance". Regarding (b), GPT identified "transparency" as a metric from [73], even though "transparency" is not part of the candidate list.

Additionally, we observed two scenarios in which certain metrics were presented in the papers but not extracted by the model. First, as noted in [101], GPT struggled to accurately extract metrics presented in tables or figures and omitted metrics such as "fear of being judged" and "usefulness." Second, GPT omitted metrics that were described as 'measured' in the original text but were not explicitly reported in the final results. For example, in [46], "usage motivation" was "measured" in the questionnaires but not reported. Similarly, in [97], "technology acceptance" was measured but not reported; instead, "intention to use a conversational agent" was reported.

3.3.2 Evaluating Metrics Recommendation. We evaluated the performance of the metrics recommendation using a sample of 40 papers, including papers from CHI, CSCW, DIS, UIST, etc., in emerging technologies such as AI, wearable technology, VR/AR, conversational agents, and similar fields. The metrics for these papers were extracted using GPT-4o by providing the complete text of the papers to the LLM engine and they were manually verified by two authors to serve as the ground truth.

Our system is supported by GPT-4o, and we also compared our system against the same LLM engine to recommend UX metrics using precision, recall, and F1 score [26, 90]. The baseline GPT-4o and our system were given the same set of inputs, including the system description and the corresponding list of indexes. Unlike metrics extraction, where the entire text of the paper was provided,

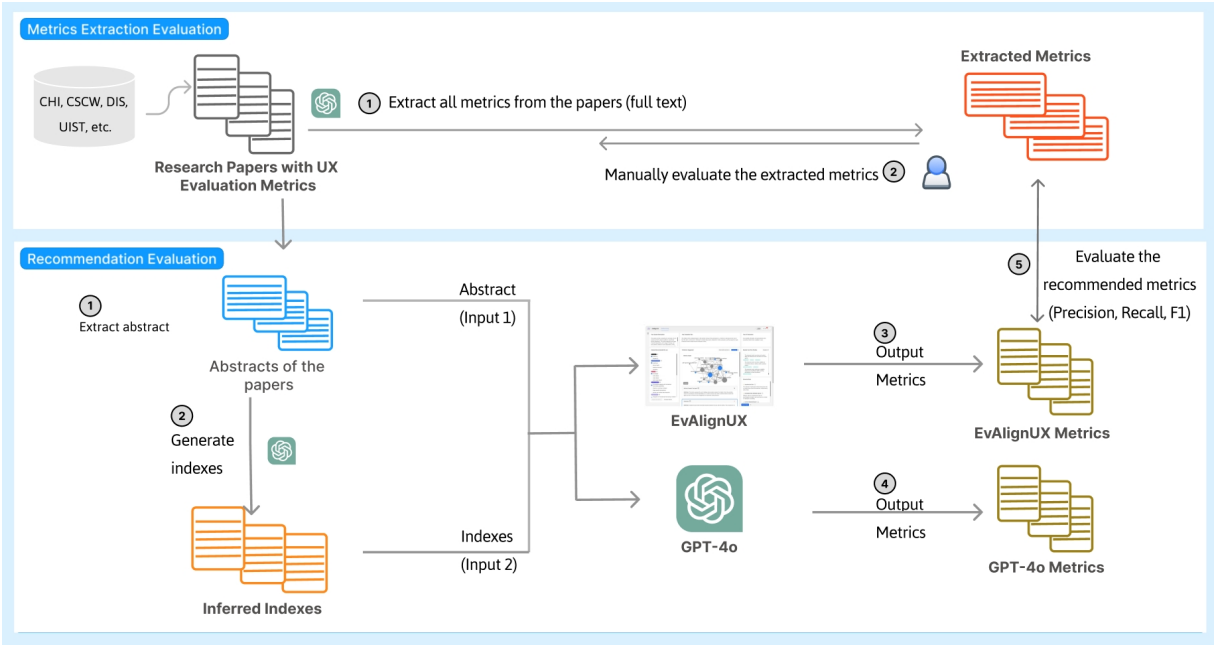


Figure 6: An overview of the workflow for evaluating *Metrics Extraction* (illustrated in the upper box, details provided in Section 3.3.1) and *Metrics Recommendation* (illustrated in the lower box, details provided in Section 3.3.2). For example, the metrics extracted from the full text and manually validated by researchers served as the ground truth. To evaluate the performance of the metrics *Recommendation*, we compared the metrics recommended by GPT-4o (baseline) with those recommended by our EvAlignUX tool, using abstracts from sample papers published in HCI conferences such as CHI, CSCW, DIS, and UIST. The evaluation was based on precision, recall, and F1 scores calculated using the ground truth metrics.

the evaluation against both EvAlignUX and the baseline LLM (GPT-4o) was conducted using only the system description as the input. This approach reflects the likelihood that EvAlignUX users would typically provide only a brief description of their system. Given the system description, a corresponding list of indexes was inferred using an LLM for each paper, as described in section 3.2.2. We then compared EvAlignUX and the baseline GPT-4o based on the number of metrics they recommended that were actually used in the papers, as determined by referring to the ground truth. For instance, Wan et al. [98] used "response quality", "conversation quality", "average response time", and "agent behavior" to measure users' satisfaction with NPC responses, evaluate the coherence and contextual relevance of NPC dialogues, determine the response latency, and assess the emotional authenticity of NPC interactions, respectively. Given the system description and identified indexes for this work, we assessed EvAlignUX 's and the baseline's ability to recommend relevant and coherent UX metrics.

For implementation, we sampled each input three times using our system and the baseline to generate a list of metrics. The results obtained from both systems were generalized to reflect the naming standards used in the test dataset. To perform the final comparison, we calculated the precision, recall, and F1 score for each subsequent run, and the final metric scores obtained were averaged to obtain the overall mean scores.

Table 2 demonstrates that our system significantly outperforms the baseline GPT-4o, with a 6% increase in precision, a 10.1% increase in recall, and a 7.4% increase in F1 score. We further conducted a pairwise t-test by processing each sample 10 times on both systems, to compare the metric mean between EvAlignUX and the baseline ($t = 8.32, p = 0.014^{*}$). The t-test demonstrates that there is a statistically significant difference between the performance (precision, recall, and F1 score) of the two models on the same dataset. Additionally, we observed that our system recommends more relevant UX metrics for a given system description than the baseline GPT-4o. For instance, when evaluating Tan et al.'s work [91], our system recommends using *Perceived Naturalness* to measure how naturally the AI-generated visuals integrate with the audiobook narrative.

However, when evaluating the baseline GPT-4o, none of the generated metrics explicitly address the naturalness of the integration between visuals and audio. Furthermore, when testing the system on Wang et al. [103], our system recommends evaluating

Model	Mean Precision	Mean Recall	Mean F-1 Score
GPT-4o	0.096	0.203	0.121
EvAlignUX	0.156	0.304	0.195

Table 1: Comparison for metric recommendation between GPT-4o and EvAlignUX

Emotion Awareness to monitor the system's ability to recognize and appropriately respond to users' different emotional states, which is also measured in the research. In contrast, none of the metrics recommended by the baseline model address the evaluation of the varied emotional states of the user. Additionally, when using the approach from [89], the system recommends measuring *Perceived Intelligence* to assess how users perceive the system's ability to generate coherent motions, which the baseline model does not explicitly recommend. The results indicate that our system outperforms the baseline GPT-4o in recommending UX metrics to users.

4 User Study

In this section, we outline how we evaluate EvAlignUX 's user experience and explore the following key research questions.

RQ1: *How effectively does the use of EvAlignUX impact researchers when developing their UX evaluation plans?*

RQ2: *How do users leverage the EvAlignUX features to enhance their thought processes when developing UX evaluation plans?*

RQ3: *How do users' backgrounds impact their use of EvAlignUX ?*

4.1 Participant

Our study was approved by the university's IRB. We recruited 19 participants from social media. All participants reported that they had extensive experience with HCI/UX research: 11 participants (57.9%) had 3–4 years of experience, 7 participants (36.8%) had 1–2 years of experience, and only one participant had less than one year of UX experience. Regarding participants' use of AI in research studies, the majority used AI to help with simple tasks, such as coding and writing (17, 89.5%), literature review (7, 36.8%), and data analysis (11, 57.9%). Notably, we found that few participants used AI for tasks that require rigor and creativity, including research ideation (7, 36.8%), design prototyping (7, 36.8%), research evaluation plan development (3, 15.8%), and usability testing (1, 5.3%). The background information of participants is shown in Appendix 6. Each user study lasted 1.5 to 2 hours. Participants received 20 USD per hour as compensation for the study.

4.2 Procedure

Figure 7 demonstrates the study procedure. ① The study began with collecting background information, including participants' UX experience levels. Researchers then introduced the study context and explained the tasks. Then, they used EvAlignUX 's landing page to spend about 10 minutes entering their project descriptions, evaluation plans, and expected outcomes. After that, they completed an initial survey that captured their perceptions of the proposed project plans (survey 1, initial plan). Then, participants used the Zoom Whiteboard to draw their thought processes for developing the UX evaluation plan and engaged in a think-aloud to ensure their ideas were communicated clearly.

In EvAlignUX , participants were guided through the system features and asked to refine their project plans using all available features. After experimenting with the features and achieving satisfactory results, participants used the "Export Data" button to review, regenerate, and re-edit their project plans. Upon finalizing their plans, they completed exit surveys (survey 2, revised plan; survey 3, EvAlignUX) to evaluate their perceptions of the generated plans and their experience with the system. During the EvAlignUX process, the think-aloud method was employed to observe and record participants' interactions and thoughts. Next, participants used Zoom Whiteboard to revise their thought processes for developing the UX evaluation plan if they felt changes were needed. Finally, participants were interviewed to delve deeper into their overall experiences and gather additional insights. This step provided a comprehensive understanding of user perceptions and allowed the collection of feedback on both the proposed plans and the system features.

4.3 Measures

We recorded the following data to assess participants' performance and perceptions during the study: task logs, thought process diagrams, think-aloud data, Likert-scale questionnaire responses, and interview transcripts.

4.3.1 System Logs. For system log data, we first collected user's input to the system, including the project description (i.e., *What are the main objectives of your project?*), UX evaluation plans (i.e. *How do you plan to evaluate the UX of your project?*), and expected outcomes (i.e. *What outcomes do you expect from this evaluation?*) as well as any updated ones. Additionally, we logged their usage behavior data, including (1) the number of times a user regenerated indexes, (2) the total number of metrics selected, (3) the number of metrics added through the graph, (4) the number of outcomes clicked including both selected and deselected outcomes, and (5) the number of risks clicked. We also logged their revised system description, UX evaluation plans, and expected outcomes.

4.3.2 Thought Process Diagrams. We collected users' thought process visualizations for developing a UX evaluation plan using Zoom Whiteboard before and after using the EvAlignUX to see how AI-based assistance influenced their thinking.

4.3.3 Questionnaire. The questionnaire is divided into three main parts:

- Background information: Users answered questions related their past UX experiences, AI literacy, and AI usage for research purposes.
- Perceptions of initial and revised UX study plan (survey 1 and 2): We collected user perceptions before and after they explored EvAlignUX to determine if there were significant differences between their perceptions of their UX research plans. The survey items are adapted from main schemes for analyzing UX evaluation methods [18, 93, 95], including novelty, relevance, efficiency, completeness, clarity, concreteness, quality, feasibility, rigorous, risks considerations, and confidence.
- System feature perceptions: For each metric, we asked participants to rate their perceptions of each feature across 8 Likert-scale items (survey 3): enjoyment of use [15], intention to use it in the future [23], perceived usefulness [16], trust in the suggestions and decisions provided [23], perceived cognitive load in understanding the information [70], inspiration for exploring and developing UX research [13], the system's contributions to the UX evaluation plan, and the system's contributions to the UX outcome.

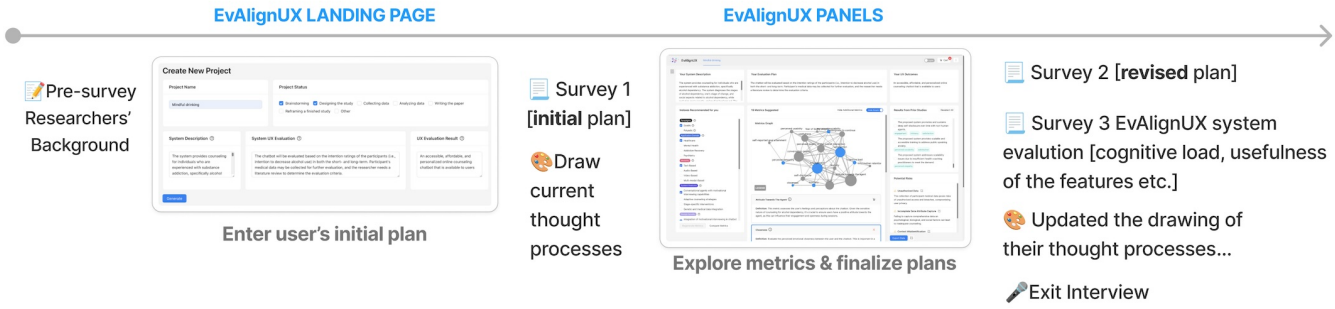


Figure 7: The Study Procedure. Participants took four surveys at different times during the study. They were asked to provide initial project ideas to EvAlignUX and explore its three panels to select metrics and finalize their evaluation plans. They were also asked to draw their thought-processes before and after using the EvAlignUX panels.

4.3.4 Think Aloud. We recorded participants' think-aloud protocols while they used the system features as well as their thought process diagrams. For example, we asked them to provide the reasoning behind their decision-making process.

4.3.5 Interview. Finally, we conducted a short follow-up semi-structured interview. We asked questions in three main categories, supplemented by follow-up questions based on the interviewee's responses.

- Perceptions of features and their role in decision-making for proposing a UX plan: *Which feature is the most impressive? Which feature is the least impressive? How do these features support your decision-making?*
- Overall user experience and potential improvements: *After using the system, how useful do you feel about the system? Were there any features or information that you felt were lacking? How do you view your original and revised plans?*
- Follow-up questions about their research mental model: *What differences do you notice between your previous research process and the process of interacting with EvAlignUX ? How do you perceive the strengths and weaknesses of these strategies?*

4.4 Data Analysis

We analyze each participant's ① thought process diagrams, ② quantitative data from surveys and logs, and ③ user decision-making reasoning through think-aloud protocols, interview transcripts, and UX research plans.

4.4.1 Thought Process Diagrams Analysis. We standardized the thought process diagrams drawn by participants by coding each illustrated step. Modifications, such as envisioning AI assistance or adding new steps, were identified. These changes were further analyzed by examining the diagrams alongside the think-aloud data. We summarized the benefits participants identified where AI could assist in proposing a UX research plan.

4.4.2 Quantitative Data Analysis. We analyzed participants' responses to the two post-task surveys. For all these measures, we conducted a Shapiro-Wilk test [31] and found that the data were not normally distributed. Given the nature of our data, we applied Kendall's tau correlation to explore relationships between user experience levels, system usage, and perceptions of changes in plans and system features [9, 44]. We also reported effect sizes using standardized methodologies [20]. All statistical analyses were conducted using Python 3.12.

4.4.3 Behavioral Analysis through Think-Aloud and Interviews. A thematic analysis [11] was performed on data from think-aloud sessions and interviews. Two researchers collaboratively identified common themes through an iterative process. This analysis helped map how EvAlignUX features address the challenges stemming from participants' thought processes while also outlining the pros and cons of these features.

5 Findings

This section presents findings on how participants' perceptions of their proposed UX evaluation plans changed through the use of EvAlignUX (RQ1) in Section 5.1, their feature usages and perceptions (RQ2) in 5.2, and how these perceptions varied based on their UX research background (RQ3) in Section 5.3. Participant quotes are indicated by *quotation marks*, while `system features` are formatted in code style.

5.1 EvAlignUX Significantly Improves the UX Evaluation Plans (RQ1)

In this section, we compared participants' ratings of their original and revised UX plans across 11 measurement criteria. Of these, nine measurements showed significant improvement in the revised plans compared to the original. Five participants fully accepted the system-generated evaluation plan without modification. On average, 60.1% (SD = 0.33) of the generated UX evaluation plan and 59.6% (SD = 0.38) of the generated UX outcome were incorporated into their final evaluation plans. The results are shown in Figure 8, with detailed information available in Appendix Table 7.

Overall, the revised plans were more specific. The increased scores in Concreteness (W = 24.0, p = .032*, r_b = 0.74), Completeness (W = 16.0, p = .004**, r_b = 0.79), Rigor (W = 15.0, p = .009**, r_b = 0.75), and Quality (W = 11.0, p = 0.004**, r_b = 0.82) suggest that participants added more detailed information to their revised

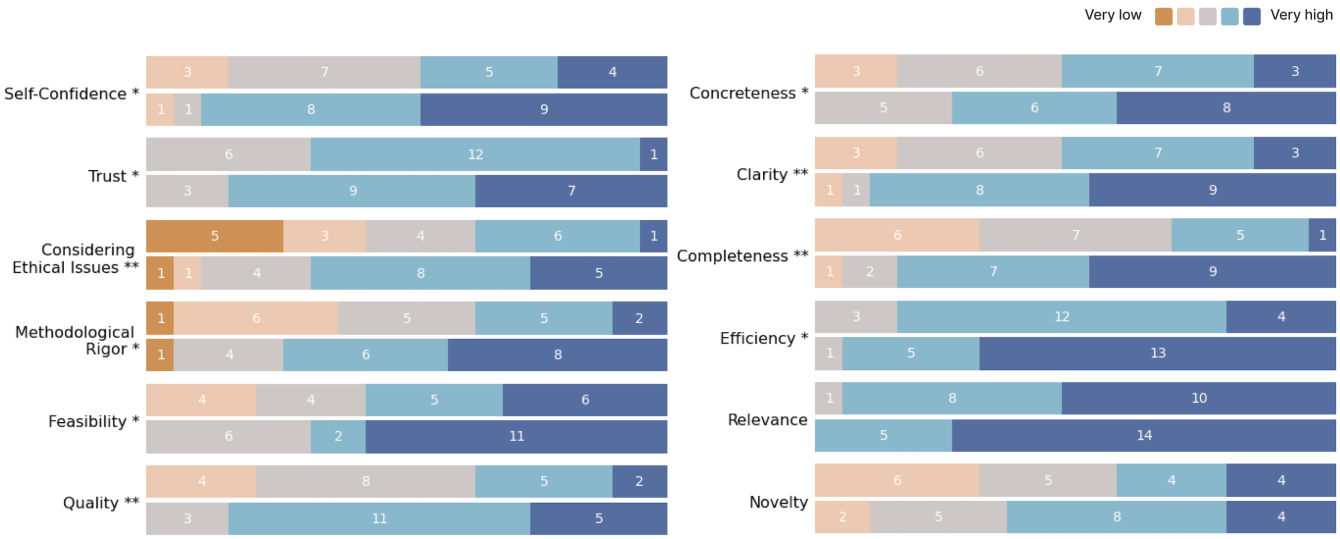


Figure 8: Distribution of participants' ratings on their proposed UX plan before (top of each item) and after (bottom of each item) using EvAlignUX . Participants felt that the updated plans are more efficient in achieving their research goals, more complete, clear, concrete, high-quality, feasible, rigorous, considerate ethical issues, trustworthy, and enhanced their self-confidence compared to their original proposed plans.(*:p<.05, **:p<.01). The effect size for all significant findings ranges from r_b= 0.51 to 0.90.

plan. Several participants only had a vague idea about which variables to measure (e.g., P7, P14, see Table 2). With EvAlignUX , they noted how it helped them improve the details in the UX plans. For example, P14 later *strategically* specified six measurements by questionnaires and system logs, with each in great detail; and P7 *copied* a holistic UX evaluation plan using a survey, system logs, and interview protocols, each with specific details added.

Participants not only added more details to their initial plans but also perceived them as more **concise and relevant**, as both Clarity (W = 4.0, p = 0.005**, r_b = 0.90), and Efficiency in achieving research goals (W = 12.0, p = .012*, r_b = 0.74) also demonstrated a considerable improvement. Participants made meaningful revisions for clarity. For instance, P5 removed text related to assessment from their initial system description, as they found it more suitable as a part of the evaluation plan instead of putting it into the system narration. Also, several participants praised the `"Metrics list"` feature for clarifying metric relevance, noting it showed *"how this metric can be relevant in the context (P3)"* and *"encouraged the reasoning of using this metric (P17)"*. P8, during a think-aloud session, explained, *"I just selected 'information retention' because my project is about memory, and the definition reminds me of the importance of evaluating how well my participants retain information to measure the effectiveness of counseling sessions in my project."*

Participants also felt **more confident** and believed their plans **were feasible** as Feasibility (W = 16.5, p = 0.035*, r_b = 0.64) and Self-Confidence (W = 18.0, p = .014*, r_b = 0.70) as indicated by significant increases. For example, participants felt they could know how to *cite (P12, P15)*, *"measure (P1, P3, P14)"*, *"design questions (P1, P9, P17)"*, *"learn previous findings (P8, P2)"* for metrics using the metric feature, making the UX plan tangible, and even *"how to collect the data specifically"* (P13, see Table 2). They also gained confidence, especially when noticing prior work used the same metric selections. As P3 expressed, *"I'm really happy someone else also uses' improvement in social skills in this context."* The validation further reinforced their self-confidence in the chosen metrics and plans.

Participants also significantly revised their plans to include more **risk considerations** (W = 10.0, p = 0.007**, r_b = 0.81). For example, building upon their original UX outcomes, P11 expanded, *"There alongside the expectations the research outcome may not align with education effect the system may not be generalizable. However, these issues could be partially reflected in the objective metrics and be acknowledged in the limitations."* P11 took inspiration from concerns about generalizability from the system's output (see Table 2, where "Generalization Failure" was mentioned, stating, *"Thinking about the potential biases made me rethink how I approach user data to ensure that I don't overlook important issues."*

However, there were no significant changes found in "Novelty" (W = 22.0, p = 0.083, r_b = 0.51) or UX plans' "Relevancy" to research questions. Several participants felt that the system was hard to change. P4 noted, *"The core idea is already set,"* while P6 said, *"It's challenging to introduce major changes."* Participants also reflected that "*It is hard to justify the novelty without referring to the actual paper and verifying it on my own*" (P4).

5.2 EvAlignUX Feature Use and Perceived Value (RQ2)

In this section, we discuss how features of EvAlignUX helped overcome various challenges in developing evaluation plans while also streamlining the thought process. In particular, three features,

Input	ID	Initial Plan	Revised Plan
UX Plan	P7	To conduct a semi-structured interview with experts to see how they collaborate with our system to analyze the video	Use a survey to measure the trust experts have in the AI's analysis capabilities. Questions could focus on reliability, perceived competence ... Discuss their perceptions of the assistant's expertise, any concerns about the analysis, and instances where the assistant's suggestions were particularly effective or ineffective.
UX Plan	P14	Ask users to experience this tool and other similar ones and then report usability, satisfaction, and others with questionnaires.	Collection of short open-ended responses from participants to analyze qualitatively. Ask pre and post survey questions about their decision confidence (how accurate they feel about their prediction of what the AI will do), perception of how clear the AI visualization was, and how accurate their prediction was. Do these before and after each action.
UX Outcome	P13	Low mental load, higher satisfaction, higher quality of subfield extraction.	I can expect the tool to provide accurate and useful summaries for both lay people and HCI researchers, leading to high user satisfaction and effective functionality. The interactive visualizations will likely be well-received for their ability to save time and allow quick exploration.
UX Outcome	P11	It outperforms the baseline techniques in the education effect and subjective preferences. Achieved this results through a more engaging experience, which also lowered participants' understanding difficulty.	(...Same with initial...) There are potential risks alongside the expectations that the research outcome may not align with the intended education effect and the system may not be generalizable. However, these may could be partially reflected in the objective metrics and be acknowledged in the limitations.

Table 2: Participants' Initial and Final UX Evaluation Plans

metrics graph, metrics list, and outcomes, were regarded as the most effective.

5.2.1 The Value of EvAlignUX Features in Enhancing UX Evaluation. We observed how the use of individual features and the combined use of different features in EvAlignUX enhanced the user experience. In Figure 9, we summarized a set of prototypical questions that UX researchers asked themselves while developing their UX evaluation plans, annotated from think-aloud data collected during participants' thought processes, which highlighted five perceived benefits of AI: justify, inform, support self-critique, streamline, and vision.

Participants mentioned that AI could be used to **inform** them by providing knowledge from prior literature, e.g., to quickly *"suggesting relevant literature*" (P15, 18), *"summarizing existing evaluation methods based on the research context"* (P3, P11). AI also supported researchers in understanding appropriate metrics by offering *"definitions"* (P8), *"citations"* (P3), *"how it can be used"* (P16), and information on "constructs adopted" (P17). For example, P19 combined the **`Index`** and **`Metrics List`** features. P9 initially felt that EvAlignUX did not accurately capture their intentions. They proposed a version control system, but EvAlignUX recommended metrics that were more likely to measure "machine learning efficiency," such as accuracy. However, after modifying a few indexes (e.g., adding two design novelties: graph-based version control, and automated design documentation), they found that EvAlignUX suggested new metrics, such as interaction experience, overall satisfaction, perceived usability, and task success, that were much more accurate and aligned with their research.

Many participants recognized the value of using AI to visualize literature networks, which help them better ***justify*** metric choices to align with research goals. This capability was particularly appreciated because it facilitates *"metric selection across various venues and disciplines"* (P17), *"highlight connections between metrics"* (P16) to determine whether two metrics are "interrelated or distinct" (P13) and supports decisions based on *"popularity and relevance in the field"* (P9) in a way that *"aligns with their (my) research goal"* (P6). For example, P9 used **`Metrics Graph`** and **`Outcome`** together. P9 designed a conversational agent for judges. After reviewing the graph, they found that user satisfaction, accuracy, and task completion were connected, which motivated them to think about the relationship between these three metrics. However, they found that while certain metrics aligned with their research goals, not all findings were applicable to their specialized context. They remarked that, no emotions or diversity should be included as my evaluation outcomes; otherwise, the judges would be easily biased.

Participants recognized the value of using AI to help them better ***self-critique*** during the development of their UX evaluation plans. This capability was particularly appreciated as it facilitates *"prompt feedback"* (P1) and provides "new perspectives" (P15), acting as a *"teammate"* (P6, P16) or *"partner"* (P18) in scholarly settings by commenting on the *"effectiveness and limitations"* of their proposed plans (P13). One participant also mentioned its potential to support

	General Questions	Questions Related to AI Support
RESEARCH GOAL	• *What is the research question?* • *What is the hypothesis?*	
LITERATURE REVIEW	• *What are common metrics in this context?* • *What gaps exist in previous evaluations?* • *How do others measure similar goals?* • *How can my UX plan contribute to literature?*	• *What are the similar terminologies used?* • *Are the keywords suggested relevant to my contexts?* • *How to search with the keywords combinations?* • *Is there any new perspectives (e.g., stakeholders)?* • *What's the meaning of this new keyword?*
EVALUATION METRICS	**Selection:** • *What is the rationale for using these metrics?* • *How can this metric apply to my context?* • *How to identify domain-specific metrics?* • *How's the novelty of the selected metric?* • *How does the metric impact my entire plan?* • *What aspects are missing from current metrics?* **Collection:** • *How will data be collected (e.g., interviews, surveys, logs)?* • *Do I add other types of data?* • *Which constructs should I use?* • *How do I design conditions and task?* • *Are there any challenges in collecting data?* • *How much time and money do I need?* **Analysis:** • *How do I analyze the data to achieve my goal?* • *Do I have a baseline?* • *What is the effect size of my analysis?*	• *What is AI's definition of this metric in my context?* • *How do AI suggest to use this metric in my context?* • *What are the metrics under [domain] among them?* • *Can these metrics be grouped by their commonality?* • *Which domains do this [metric] apply to previously?* • *What are the evaluation methods that have been used in similar contexts?* • *Why some metrics have larger nodes?* • *Why some metrics are used under the same context?* • *Why some metrics are collected differently?* • *How and why some metrics are suggested or removed after changing the system keywords?* • *What are the survey constructs or items for these metrics?* • *What is the paper that has the highest citation using this construct?* • *Does the AI adapt the survey items to my contexts correctly?*
RESEARCH IMPACTS	**Expected Contributions:** • *What is the expected outcome for these metrics?* • *What is the expected contribution using the approach?* **Potential Risks:** • *What are the risks of relying on these metrics?* • *How can I potentially mitigate these risks?*	• *What are the outcomes related to general contributions that apply to all user-centered systems?* • *What are the outcomes that are specific to a context (e.g. healthcare, e-commerce)?* • *How do I filter results from prior literature?* • *Why does AI suggest this risk category to me?*
EXTERNAL FEEDBACK	• *Do I need to add new types of data?* • *What's the problem with the plan?* • *Do experts see any biases in this plan?*	• *Do I understand the limits of AI's generation?* • *What are the new insights generated by AI?* • *Can I trust AI's result?*
SYSTEM PROTOTYPE	• *How feasible is implementation to achieve my goal?* • *[If not] How do I design the WoZ instead?* • *What functions to be prioritized to evaluate the UX?*	

Figure 9: Question Bank for Developing UX Evaluation

"overcoming social inhibitions, such as hesitancy in seeking advice or posing questions to peers too frequently" (P16). As P4 commented that, *"It's nice also to have EvAlignUX where you can throw ideas around, and then it will guide you directly in your research project, what to think about—not just a chat with somebody, which can sometimes be embarrassing."*

Participants also recognized the value of using AI to reduce cognitive load, which helps them better ***streamline*** their UX evaluation development and align it with dynamic project timelines. This capability was particularly appreciated as it *"simplify decision-making processes in UX design"* (P16) to determine whether new updates are "directly contributing to the UX" (P13) of their original research goal or not. For example, P4 said, *"Generally if I need to propose a UX plan, I probably need to spend like two days at least to collect all the papers. I believe your system can actually help us [HCI researchers] to shorten the time. Because every paper here comes from valid sources, so we can trust the results."* Furthermore, participants also shared the value of using the system for ***re-structuring their thought process*** of the UX plan development. Figure 10 is an example illustrating how P2 modified the literature review process by using AI to conduct the literature review and brainstorm metrics earlier in the process.

Nearly all participants recognized the value of using AI to promote early detection of risks, which helps them better ***envision*** their UX planning. This capability was particularly appreciated as it facilitates *"spotting potential risks before they turn into real*

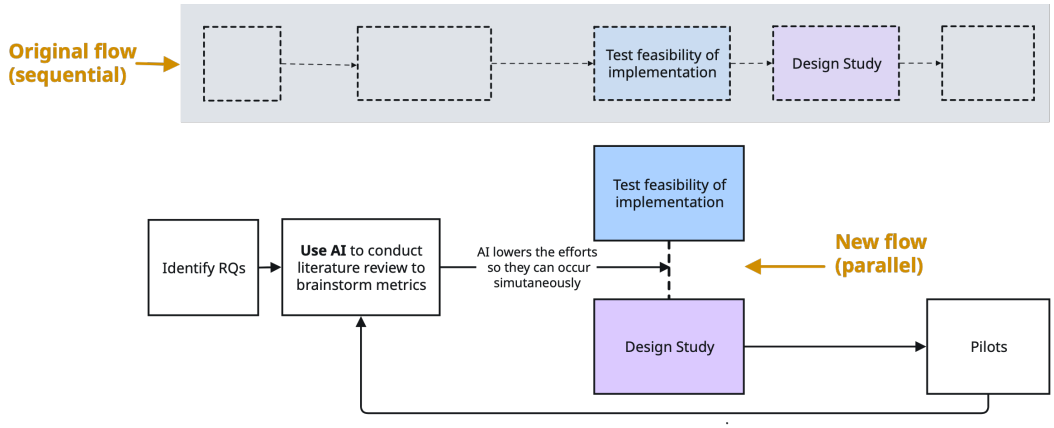


Figure 10: P2's Thought Process. P2 emphasized that AI "lowers the efforts so that the two steps can now occur simultaneously" so that *"rapid updates to evaluation plans as new features are added."*

headaches after the study is done" (P5), and supports decisions based on *"minimizing unexpected challenges"*(P9). As P10 said mentioning the **`Risk`** feature, *"It doesn't simply show a list of risks but encourages users to click the associated link to review the actual report."*

5.2.2 Enriched Thought-Processes with the Support of EvAlignUX . Besides how participants used the EvAlignUX features differently, they also demonstrated diverse thought processes during their development of UX evaluation plans. When updating their thought processes, they all revised their literature review process by incorporating relevant EvAlignUX features. More than half of the participants expressed that AI facilitates ***iterative loops*** during the development of an evaluation plan (8 participants). For instance, Figure 11 illustrates P13, a UX researcher with 1-2 years experience, revised thought process after using EvAlignUX , showing AI's role as a supportive teammate in this process. The AI-suggested metrics, as highlighted by the purple boxes, prompted participants to critically assess the relevance and appropriateness of AI's recommendations, ensuring alignment with system design goals and expected results. This iterative reflection, where P13 evaluates the added value of AI's input, emphasizes AI's role in fostering critical thinking and refinement. Similarly, Figure 12. demonstrates how P11, with 3-4 years of UX experience, views AI as beneficial to the iterative metric selection process.

We also observed that UX risks and outcomes were incorporated into participants' thought processes, **regardless** of their experience level. Overall, 42% of participants incorporated new components into their UX evaluation development thought processes after using EvAlignUX . Notably, these included "Propose measurements" (3 junior UX-level participants, mixed research stages), "UX risks" (mixed UX levels, mixed research stages), and "Research outcomes" (mixed UX levels, mixed research stages). These additions reflect a newly gained awareness of potential challenges and impacts in their UX evaluation plan development. For example, Figure 13 shows how P7, a UX researcher with 1-2 years of experience, added two new considerations: defining the ideal outcome of the evaluation and identifying potential risks associated with the chosen measurements to their thought process after using EvAlignUX .

5.2.3 The Significant Use of Hypothesized **`Outcome`** *and* **`Metrics Exploration`** *Features.* Based on these observations, we further analyzed the relationship between users' system log variables (e.g., the number of times indexes were regenerated, the number of metrics selected via the Metrics List, the number of metrics selected or deselected via Metric Graphs, the number of outcomes clicked, and the number of risks clicked) and their perceptions of five different features across 11 perception variables (e.g., enjoyment, intention, usefulness, cognitive load, inspiration), see Section 4.3.3 for measurements. A total of 55 pairs were examined, with 6 showing significant results. All significant results are presented in Table 7.

The results showed that selecting more potential outcome items (i.e., previous research findings) is associated with higher user ratings of the **`Outcome`** feature, both enjoyment (r = 0.57, 95% CI [0.28, 0.82], $p < 0.01^{**}$) and intention to use (r = 0.40, 95% CI [0.05, 0.66], $p < 0.05^{*}$). Additionally, users rated this feature more positively in terms of its contribution to revising their project idea (r = 0.45, 95% CI [0.05, 0.76], $p < 0.05^{*}$), UX evaluation plan (r = 0.44, 95% CI [0.08, 0.71], $p < 0.05^{*}$), and UX outcomes (r = 0.49, 95% CI [0.16, 0.75], $p < 0.05^{*}$). The findings from relevant HCI domains encouraged participants to consider questions such as *what demographic is left in the original plan?* (P5) *What factor was not considered yet?* (P14) These thoughts prompted them to refer back to their original project idea, UX evaluation plan, and UX outcomes. P4 selected 15 potential outcomes that were derived from prior work, they commented that for instance, *"measuring trust and confidence in the chatbot's abilities and context understanding"* was *"exactly what they were looking for and this point really hit the point"*. P4 further commented, they liked this feature because they felt "*the generation understands my initial thoughts about the project... It helps me finalize the details of my plan.*"

Additionally, the results also indicated that adding more metrics to the graph was associated with a reduction in perceived cognitive load (r = 0.40, 95% CI [0.07, 0.72], $p < 0.05^{*}$). Participants found

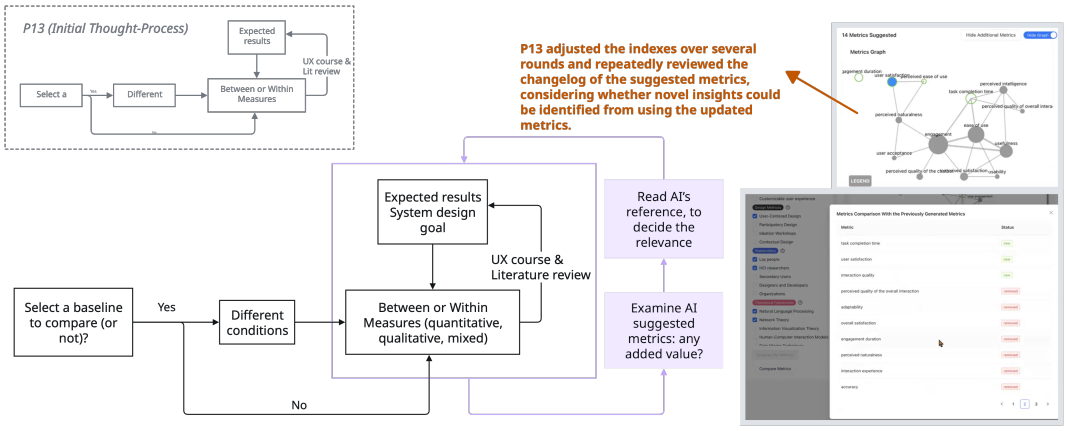


Figure 11: P13's evolving thought process when viewing AI as a teammate supporting interactive self-critique in developing UX evaluation plans for `Index` and `Metric Graph` feature. The upper left shows their initial thought process, while the purple-shaded areas indicate the perceived benefits AI introduces in the revised thought process.

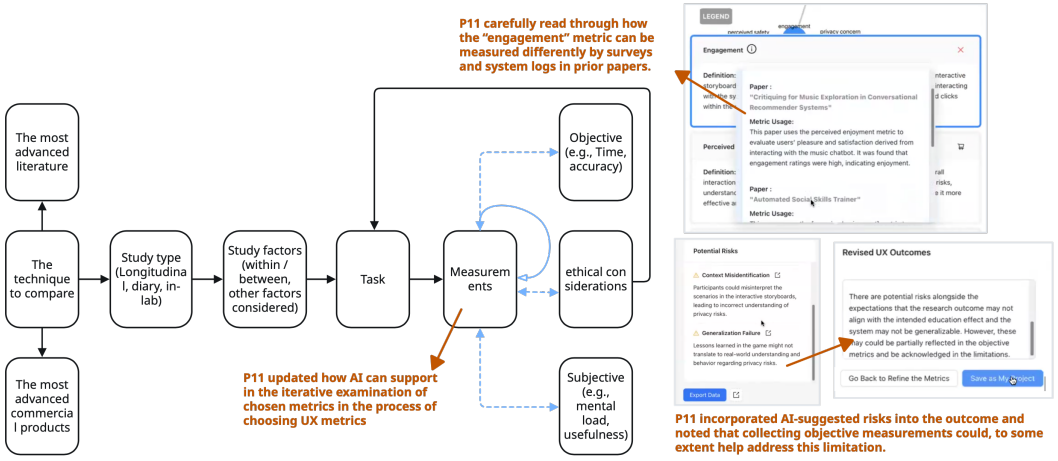


Figure 12: P11's evolving thought process when viewing AI as a teammate supporting interactive self-critique in developing UX evaluation plans for `Metrics List` feature. The blue lines shows their perceived AI's benefits in the iteration on metric selection.

	Dependent Variable				
Feature Usage Log	**`Feature`: Perception Scores**	**Mean**	**SD**	***r***	**95% CI**
# of metrics added by graph	`Metrics Graph`: Cognitive Load *	2.74	1.15	-0.40	[-0.72, -0.07]
# of selected results	`Outcome`: Enjoyment **	3.32	1.16	0.57	[0.28, 0.82]
	`Outcome`: Intention to Use *	3.63	1.07	0.40	[0.05, 0.66]
	`Outcome`: Contribution to UX evaluation*	3.42	1.39	0.44	[0.08, 0.71]
	`Outcome`: Contribution to UX outcomes**	3.37	1.50	0.49	[0.16, 0.75]

Table 3: Examining the interaction between participants' feature uses (based on the system log) and their perceived value of the features (N = 19). Significant values: $^{**}p < 0.01$, $^{*}p < 0.05$. Note that questions about fear of over-reliance were reversed and phrased as an attention check question.

using the feature of **`Metrics Graph`** simple – not only showing *"an aerial view of how these metrics tie together"* (P6) but also the usage of each metric in prior research, which is *"easy to understand"* (P8).

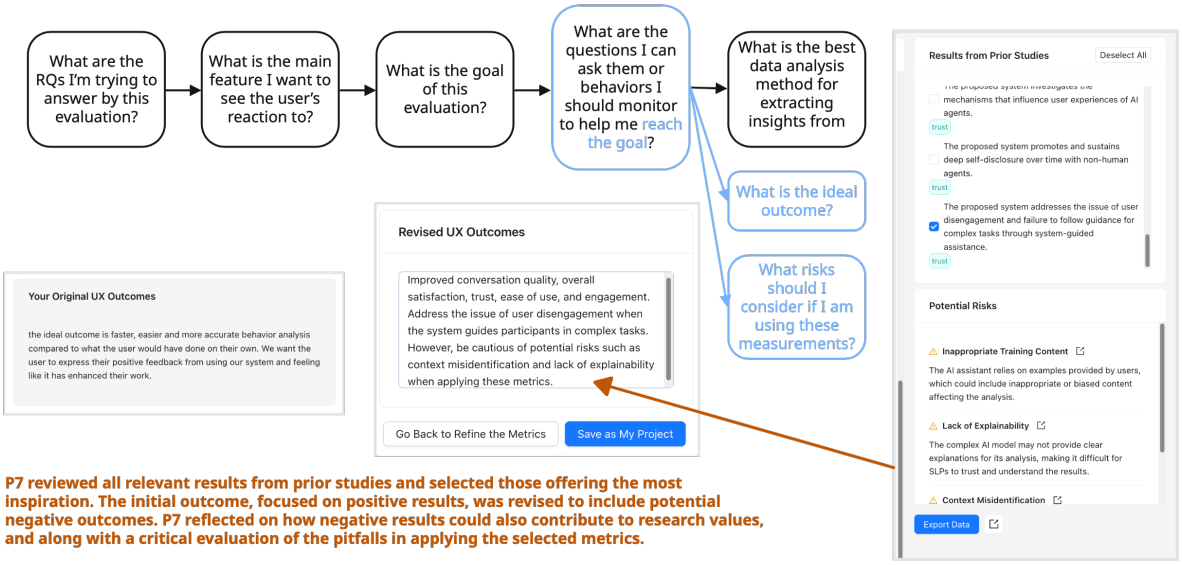


Figure 13: P7's evolving thought process by adding a new component of outcomes and risks inspired by the `Outcome` and `Risk` features.

5.3 EvAlignUX Perception Varying by Research Background (RQ3)

In this section, we examined the relationship between users' backgrounds and their perceptions of five different features across nine perception variables (e.g., enjoyment, intention, usefulness, cognitive load, inspiration, see measurements in Section 4.3.3), aiming to identify which users benefit the most from our system and which users require additional support. Seven significant results were found. All significant results are shown in Table 4.

5.3.1 The Extent of Perceived Inspiration Varies by Users' Research Experience. Overall, the perceived value of providing inspiration through EvAlignUX was significantly negatively correlated with participants' background, with the exception of the **`Outcome`** feature. **Specifically, participants with less UX experience provided higher scores for perceived inspiration, except with the `Outcome` feature.** Despite this, all participants rated the inspirational measure highly (M = 3.58–4.00, SD = 0.94–1.33).

For **`Index`**, some participants with less experience shared that the feature helps to structure research ideas using relevant terminologies as sometimes they *"just have a vague thought without knowing the exact terms"* (P9). P6, with 1 year of UX experience, shared that the feature is a *"very good at pinpointing important aspects, like user-centered design and stakeholders (researchers). It did a good job of checking what the user intended. I wouldn't say anything was wrong or needs changing. Everything was spot on. There are certain things you can add, but they're already in line with what was checked before."* For experienced researchers like P2, she mentioned that the feature helps to distinguish *"umbrella terms"* when they only have a vague understanding of what is being measured: *"It lowered the effort to identify the literature that is actually related. For instance, I saw three terms, knowledge acquisition, knowledge retention, and knowledgability. If I just use the wrong term, I may think that no one has done it before, but it turns out people have already done it."* While experienced users acknowledged its utility for combining suggested indexes in their paper searches, they felt it lacked novelty, as many of the indexes were already familiar (P2), *"I am looking for something that I have no knowledge about."* Instead, they tend to *"self-validate whether they are on the right track"* (P4) rather than seek inspiration for relevant knowledge.

For **`Metrics List`** and **`Metrics Graph`**, less experienced users valued these features for revealing *"metrics [they] hadn't considered"* (P7). As P3 (with 1-2 years of UX experience) mentioned, "*Because generally if I need to do that (on my own), I probably need to spend like 2 days at least to collect all the paper, which is really bad. I think EvAlignUX can actually help us to shorten the time. Because every paper here is coming from valid sources, such as top conferences, so we can trust the result.*" Similarly, P6 said that the graph gives her serendipity, *"An aerial view of how these evaluation schemes tie together, like from usefulness to ease of use, helps when writing the paper or designing an evaluation scheme. Certain things in usefulness and trust are often correlated, as shown by the thickness of the line, which is inspiring."* Participants shared that the graph helped them explore metrics that were *"novel"* (P3), as they noticed smaller nodes, disconnected nodes, or nodes with unfamiliar names. The graph made them feel more interested in discovering *"under-studied"* (P6) metrics in a domain; and for commonly used metrics, they felt they gained insights into common practices (P7, P12). As P7 shared, *"I feel like changing my selections based on the graph resulted in better and more related contributions."*

This pattern extended to the **`Risk`** feature. P6, with 1-2 years of UX research experience, found it interesting that the feature helped identify *"risks that directly related to the system."* More experienced researchers, like P1, mentioned that *"The risks mentioned here are not*

	Dependent Variable				
Background	**Feature: Perception Scores**	**Mean**	**SD**	*r*	**95% CI**
UX experience	`Index`: Perceived inspiration *	4.00	0.94	-0.44	[-0.73, -0.05]
	`Metrics list`: Perceived inspiration **	4.11	0.94	-0.67	[-0.84, -0.43]
	`Metrics Graph`: Perceived inspiration *	3.89	1.10	-0.49	[-0.77, -0.13]
	`Outcome`: Perceived inspiration (*n.s.*)	3.68	1.00		
	`Risk`: Perceived inspiration*	3.68	1.33	-0.42	[-0.72, -0.04]

Table 4: Examining the interaction between participants' background (based on the survey Likert 1-5) and their perceived value of the features (N = 19).Significant values: $^{**}p < 0.01$, $^{*}p < 0.05$.

hardl ines, meaning that if users are not followed, then the system will fail. For instance, if the evaluation plan doesn't receive IRB approval, they cannot continue to the next step; whereas the potential risk information here is more likely to be soft suggestions that may not contribute to actual revision."

Interestingly, inspiration from the **`Outcome`** feature did not vary with UX experience, aligning with participants ***drawing*** of thought-process, where outcomes were often missing from initial UX plans regardless of participants' UX experience. The overall high mean score for rated inspiration from `Outcome` suggests that the feature resonated across all users, though for different reasons based on experience level. Several participants with less UX experience found the **`Outcome`** feature useful for refining their research plans, noting how it helped them *"rephrase [their] outcome to make it more detailed"* (P13). P7 said, *"I haven't thought about it before, but now that I saw it here, that can be an useful contribution... It's really cool that they come from actual papers and aren't fabricated or made up."* In contrast, P4, with 3-4 years of UX experience, noted that *"without going into too much like spending a day thinking about it (UX outcomes). I think that it (your suggested UX outcomes) is a really great restatement of my original one."*

6 Discussion

In this section, we discuss methods for fostering a mindset-focused UX evaluation in HCI education and practice, the potential risks of using Large Language Models to support UX evaluation, and the design implications of our findings.

6.1 Shifting UX Evaluation from *Method-Centric* to *Mindset-Centric*

Gray's prior work [34] highlighted that many UX practitioners view UX as a *"rigid set of methods"* rather than as a *"mindset"*. This method-centric perspective often emphasizes the application of techniques without fully aligning with project-specific goals. Similarly, prior research [32, 35, 61, 79] has critiqued academic discussions for focusing on method creation and testing, neglecting real-world implementation contexts. This disconnect creates a *"projected practice community"* [34, 35], where academic conceptions fail to align with practitioners' realities.

The interaction design and evaluation of EvAlignUX demonstrated that it empowered UX researchers to conduct UX evaluation more holistically. First, providing metric definitions, relevant literature, and examples that enabled for researchers to reflect on the implications of these metrics enhanced both the perceived quality of their proposed plans and their confidence in their UX evaluation plans (RQ1). Second, combining different EvAlignUX components to refine UX plans allowed participants to self-validate, critique, and rationalize their evaluation choices (RQ2). Third, participants' thought processes and perceived inspirations from the system varied based on their UX experience and evolved from asking general questions to leveraging AI for iterative reasoning about their chosen metrics, including previously unconsidered aspects, such as outcomes and AI risks (RQ3).

6.1.1 Mindset-Centric Evaluation: Encouraging Reflections on Insights. Unlike method-centric approaches, mindset-centric evaluation emphasizes critical reflection on why, what, and how evaluations are conducted, aligning choices with broader implications–goals, values, and the insights researchers and practitioners aim to achieve. This paradigm shift acknowledges the complexity and adaptability required in UX evaluation, bridging the gap between academic research and practical applications. EvAlignUX facilitates this shift by not only helping define metrics, enabling users to understand their underlying concepts, relevance, and alignment with other research areas, but also by connecting researchers with example research papers and real-world applications. This integration encouraged them to move beyond focusing solely on specific methods and to instead reflect on the insights they aim to achieve, such as fostering engagement or building trustworthiness. By shifting their question from *"What method should I use?"* to *"What insights am I seeking?"*, EvAlignUX helped researchers adopt a more holistic perspective. Additionally, the inclusion of AI risk considerations encouraged users to evaluate their UX plans through new lenses, adding depth and broadening the implications of their work. This shift enriched the researchers' mental models, underscoring EvAlignUX 's value in promoting a mindset-centric approach to UX evaluation.

6.1.2 Implications for HCI Practice and Education. This mindset-centric shift holds significant implications for both HCI practice and education.

First, it aligns with the increasing diversity and sophistication of UX evaluation approaches [74]. While numerous reviews outline UX evaluation methods [14, 45, 74, 77, 95], few interaction designs directly support the development of comprehensive evaluation plans. For example, scholars have stressed the need for adaptable, context-sensitive approaches: Vermeeren et al. [95] stress that no single method fits all projects; Rivero et al. [77] advocate for iterative frameworks that incorporate feedback; and Darin et al. [14] call for continuous critical reflection throughout the evaluation process.

Our design aligns with the intention to democratize the design process: when EvAlignUX users select metrics from the recommended options, click "more" to explore additional metrics, edit the index to update the recommendations, or adjust their selection after identifying potential risks, they provide feedback to the system. In the final UX evaluation export, users can further refine their selected metrics. Our findings demonstrate an effective interaction design that fosters a mindset-centric perspective, informing and elevating HCI systems research.

Second, our findings reveal that EvAlignUX offers five key benefits for developing UX research plans. However, participants exhibited diverse usage patterns (RQ1, RQ2). Some enriched system-generated plans with detailed context; others prioritized applying the application of specific metrics; and some adopted the plans without critical evaluation. These patterns were, to some extent, linked to participants' UX experience (RQ3). For example, less experienced UX researchers used EvAlignUX to explore unfamiliar metrics, domains, and terminologies, thus finding it more inspirational. In contrast, experienced researchers primarily leveraged the system as a validation tool rather than a source of inspiration. This suggests that EvAlignUX 's primary value evolves with users' expertise—serving as an exploratory and instructive tool for novices while functioning as a confirming and refining aid for experts. Prior work emphasizes that HCI education should extend beyond teaching predefined methods to fostering adaptability, improvisation, and critical thinking skills [34]. EvAlignUX embodies these principles by enabling researchers to tailor their approaches to the unique demands of each project. Rather than enforcing rigid processes, future AI-driven UX evaluation systems and educational frameworks should encourage researchers to treat metrics as core design materials—fundamental elements that shape and guide evaluation. By adopting this mindset-centric perspective, researchers and practitioners can better navigate evolving complexities, ensuring their UX evaluations remain meaningful, impactful, and contextually relevant.

Third, the literature-based UX repository is a growing collection that could expand to include industry case studies and metrics, broadening its applicability beyond academic research. While academic UX research emphasizes systematic methodologies, theoretical frameworks, and empirical validation, industry UX case studies often prioritize practical insights, real-world constraints, and agile decision-making processes. Integrating industry contributions would make the repository more reflective of diverse UX practices, enabling designers, engineers, and customers to generate tailored plans and share them for collaborative decision-making. Future studies could explore how to structure this integration effectively, curating a broader spectrum of metrics from both peer-reviewed sources and real-world implementations, while maintaining clarity and usability. Additionally, incorporating industry case studies could introduce new ways of evaluating UX beyond traditional academic benchmarks, offering a more holistic perspective on how different sectors define and measure UX.

6.2 Towards Responsible Use of LLMs in Interdisciplinary and Cross-Disciplinary UX Research

Previous literature has highlighted the ethical concerns and potential harms of over-reliance on LLMs in various domains. For instance, Ahmad et al. [3] found that excessive dependence on AI in education led to students becoming passive, diminishing their decision-making abilities, and negatively impacting their cognitive skills. Similarly, in human-AI co-creative writing, AI outputs are heavily influenced by the distribution of their training datasets [5, 80]. Without considering users' unique needs, over-reliance on these outputs can result in research homogeneity and a lack of creativity [52, 76]. Additionally, issues such as hallucination, lack of transparency, and algorithmic bias have been identified as major barriers to effective AI collaboration [108]. To mitigate these risks, future iterations of EvAlignUX and similar tools could be designed with adaptive modes or adjustable feedback mechanisms that align with a user's experience level. Such tailoring would help novices develop core competencies over time while allowing more experienced users to refine advanced strategies, ultimately supporting meaningful skill growth rather than hindering it. Given the limitations of LLMs, future studies need to carefully consider and address the ***ethical concerns*** raised, ensuring that UX researchers remain critical, reflective, and ultimately self-reliant evaluators. Potential research directions include:

6.2.1 Mitigating Potential Over-Reliance and Triggering Deeper Thought Processes. Questions about potential over-reliance have also emerged. We position the design of such a tool as an educational resource that potentially bridges formal education and real-world practice, helping novices transition into more advanced roles while also supporting experts in adopting new perspectives. It is designed to foster a mindset-centric approach, which inherently encourages reflection and critical thinking, counteracting de-skilling risks. One potential direction is to use AI systems that provide hints, or to use a *Question Bank* (see Fig. 9). This approach may aim to augment researchers' reasoning by prompting reflection and encouraging deeper engagement with their work. Essential features include controllability and explainability, alongside clear disclaimers that warn users about the risks of over-reliance on generative AI [56]. Previous research highlights that transparent systems with clear AI outputs foster better human-AI collaboration by encouraging users to engage more critically with the tool [30, 99]. Furthermore, studies indicate that embedding mechanisms that encourage active thinking—such as presenting alternative viewpoints or highlighting uncertainties—can help mitigate the risk of over-dependence on AI-generated solutions and foster deeper cognitive engagement [7].

6.2.2 Using AI to Support Self-Critique. Our findings suggest that AI may assist in addressing challenges through self-critique mechanisms. For example, EvAlignUX uses its literature base to provide users with relevant information that encourages self-critique. However, unlike human experts, AI systems often fail to identify nuanced flaws in research plans that stem from researchers' misunderstandings. Previous studies have shown that AI's capacity for self-critique is limited, particularly when dealing with complex, contextual errors that require domain-specific knowledge or critical

thinking [7, 100]. While AI's ability to critique existing research is promising, it lacks the contextual depth found in human-to-human collaboration [100, 106]. Further research is needed to explore how AI can more closely mimic human critical evaluation in the research process.

6.2.3 Using AI to Mitigate UX Risks. A key design implication often overlooked is the concept of "impact visioning." While only one participant in our study considered potential risks during the initial development of their UX plans, previous studies have shown that the negative consequences of new technologies can often be predicted in early design stages [51, 72]. AI systems, by drawing on cross-disciplinary insights, may also help researchers anticipate and mitigate potential risks in UX evaluation, raising awareness of emerging concerns in technology use and design. Prior studies in AI ethics and risk management suggest that AI can be instrumental in identifying unintended consequences and risks in various domains, including UX design, by leveraging data across disciplines to foresee issues that might not be immediately apparent [27, 96]. As the body of evaluation metrics and AI-related risks expands in the rapidly evolving landscape of human-AI interaction, practitioners can benefit from recommendations that continuously evolve.

6.3 Limitations and Future Work

We acknowledge several limitations that can affect EvAlignUX 's performance and the findings. First, as an exploratory system, it incorporates a limited selection of papers, primarily focused on human-AI interaction. Also, for the EvAlignUX backend, recommendations are confined to metrics that have been predefined in our database. Participants proposing studies in interdisciplinary fields may find the system's suggestions less relevant or accurate; however, we do not claim that the current metrics repository is exhaustive. Second, although we observed differences in participants' perceptions and usage behaviors, the majority of participants were doctoral students, which may limit the generalizability of our findings. Our user evaluation was conducted by HCI researchers, but in the future, we can include industry practitioners who may lack formal UX research training but need to conduct effective system evaluations. Expanding the system's design to better support these groups in conducting effective evaluations could be an important direction for future development. Finally, EvAlignUX scaffolds the conceptual phase of early-stage UX research planning. It focuses on ideation, metric selection, and critical reflection, providing a foundation for subsequent stages, such as methods, logistics, and participant considerations. For instance, selecting metrics upfront facilitates targeted discussions about participants and data collection methods. However, we acknowledge that detailed logistical support is beyond the tool's scope and could be addressed in future enhancements.

7 Conclusion

User experience is vital to any Human-Computer Interaction research, yet many UX designers and practitioners often rely on rigid evaluation methods rather than adopting a holistic approach. In this paper, we introduced EvAlignUX , an LLM-based tool designed to support UX evaluation development. Through an evaluation with 19 HCI researchers, we found that centering UX metrics in the evaluation process fostered a mindset that balanced holistic UX impact with contextual factors, evaluation practices, and research value. This approach significantly improved participants' perceived plan quality and confidence, while also prompting their thought processes to consider risks when evaluating systems. For practical implications, we offer a *Question Bank* to guide future UX evaluation efforts, and suggest that future researchers can enhance EvAlignUX by incorporating domain-specific literature to further refine their evaluation planning. In a world where experience defines impact, a UX-driven mindset serves as the key to meaningful and lasting designs evaluation.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant #2119589. The work is also supported under the AI Research Institutes program by the National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award #2229873 - AI Institute for Transforming Education for Children with Speech and Language Processing Challenges. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education. Additionally, this project is made possible in part by the Institute of Museum and Library Services RE-252329-OLS-22. The views, findings, conclusions or recommendations expressed in this article do not necessarily represent those of the Institute of Museum and Library Services. The authors acknowledge the National Artificial Intelligence Research Resource (NAIRR) Pilot (NAIRR240177) for contributing to this research. We sincerely thank the anonymous reviewers for their constructive suggestions, which have significantly strengthened this work.

References

[1] Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, et al. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digital Medicine* 7, 1 (2024), 82.

[2] Khlood Ahmad, Mohamed Abdelrazek, Chetan Arora, Muneera Bano, and John Grundy. 2023. Requirements practices and gaps when engineering human-centered Artificial Intelligence systems. *Applied Soft Computing* 143 (2023), 110421.

[3] Sayed Fayaz Ahmad, Heesup Han, Muhammad Mansoor Alam, Mohd Rehmat, Muhammad Irshad, Marcelo Arraño-Muñoz, Antonio Ariza-Montes, et al. 2023. Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications* 10, 1 (2023), 1–14.

[4] Zafar Ali, Guilin Qi, Pavlos Kefalas, Waheed Ahmad Abro, and Bahadar Ali. 2020. A graph-based taxonomy of citation recommendation models. *Artificial Intelligence Review* 53 (2020), 5217–5260.

[5] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (2023), 337–351. doi:10.1017/pan.2023.2

[6] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[8] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[9] Sorana-Daniela Bolboaca and Lorentz Jäntschi. 2006. Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences* 5, 9 (2006), 179–200.

[10] Claus Bossen, Christian Dindler, and Ole Sejer Iversen. 2016. Evaluation in participatory design: a literature survey. In *Proceedings of the 14th Participatory Design Conference: Full papers-Volume 1*. 151–160.

[11] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[12] Beenish Moalla Chaudhry. 2024. Concerns and Challenges of AI Tools in the UI/UX Design Process: A Cross-Sectional Survey. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 83, 6 pages. doi:10.1145/3613905.3650878

[13] Hui-Chun Chu, Gwo-Jen Hwang, Chin-Chung Tsai, and Judy CR Tseng. 2010. A two-tier test approach to developing location-aware mobile learning systems for natural science courses. *Computers & Education* 55, 4 (2010), 1618–1627.

[14] Ticianne Darin, Bianca Coelho, and Bosco Borges. 2019. Which Instrument should I use? Supporting decision-making about the evaluation of user experience. In *Design, User Experience, and Usability. Practice and Case Studies: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part IV 21*. Springer, 49–67.

[15] Shayn S Davidson, Joseph R Keebler, Tianxin Zhang, Barbara Chaparro, James Szalma, and Christina M Frederick. 2023. The development and validation of a universal enjoyment measure: The enjoy scale. *Current Psychology* 42, 21 (2023), 17733–17745.

[16] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[17] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. 2011. Generalized louvain method for community detection in large networks. In *2011 11th international conference on intelligent systems design and applications*. IEEE, 88–93.

[18] Douglas L Dean, Jill Hender, Tom Rodgers, and Eric Santanen. 2006. Identifying good ideas: constructs and scales for idea evaluation. *Journal of Association for Information Systems* 7, 10 (2006), 646–699.

[19] Michael Desmond, Zahra Ashktorab, Qian Pan, Casey Dugan, and James M Johnson. 2024. EvaluLLM: LLM assisted evaluation of generative outputs. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*. 30–32.

[20] Wilfrid J Dixon and Frank J Massey Jr. 1951. Introduction to statistical analysis. (1951).

[21] Anil R Doshi and Oliver P Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* 10, 28 (2024), eadn5290.

[22] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).

[23] Andrea Everard and Dennis F Galletta. 2005. How presentation flaws affect perceived site quality, trust, and intention to purchase from an online store. *Journal of management information systems* 22, 3 (2005), 56–95.

[24] Lawal Ibrahim Dutsinma Faruk, Mohammad Dawood Babakerkhell, Pornchai Mongkolnam, Vithida Chongsuphajaisiddhi, Suree Funilkul, and Debajyoti Pal. 2024. A review of subjective scales measuring the user experience of voice assistants. *IEEE Access* (2024).

[25] Lawal Ibrahim Dutsinma Faruk, Debajyoti Pal, Suree Funilkul, Thinagaran Perumal, and Pornchai Mongkolnam. 2024. Introducing CASUX: A Standardized Scale for Measuring the User Experience of Artificial Intelligence Based Conversational Agents. *International Journal of Human–Computer Interaction* (2024), 1–25.

[26] Zeshan Fayyaz, Mahsa Ebrahimian, Dina Nawara, Ahmed Ibrahim, and Rasha Kashef. 2020. Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *applied sciences* 10, 21 (2020), 7748.

[27] Luciano Floridi and Mauro Chiriatti. 2021. Ethical challenges of AI in biomedicine. *AI in medicine* 84 (2021), 1–5.

[28] Santo Fortunato. 2010. Community detection in graphs. *Physics reports* 486, 3-5 (2010), 75–174.

[29] Mark Glickman and Yi Zhang. 2024. AI and generative AI for research discovery and summarization. *Harvard Data Science Review* 6, 2 (2024).

[30] Katrin Glinka and Claudia Müller-Birn. 2023. Critical-Reflective Human-AI Collaboration: Exploring Computational Tools for Art Historical Image Retrieval. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33.

[31] Elizabeth González-Estrada and Waldenia Cosmes. 2019. Shapiro–Wilk test for skew normal distributions based on data transformations. *Journal of Statistical Computation and Simulation* 89, 17 (2019), 3258–3272.

[32] Elizabeth Sarah Goodman. 2013. *Delivering design: Performance and materiality in professional interaction design.* University of California, Berkeley.

[33] Arthur C Graesser, Zhiqiang Cai, Max M Louwerse, and Frances Daniel. 2006. Question Understanding Aid (QUAID) a web facility that tests question comprehensibility. *Public Opinion Quarterly* 70, 1 (2006), 3–22.

[34] Colin M Gray. 2016. " It's More of a Mindset Than a Method" UX Practitioners' Conception of Design Methods. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4044–4055.

[35] Colin M Gray, Erik Stolterman, and Martin A Siegel. 2014. Reprioritizing the relationship between HCI research and practice: bubble-up and trickle-down effects. In *Proceedings of the 2014 conference on Designing interactive systems*. 725–734.

[36] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[37] Marc Hassenzahl, Sarah Diefenbach, and Anja Göritz. 2010. Needs, affect, and interactive products–Facets of user experience. *Interacting with computers* 22, 5 (2010), 353–362.

[38] Serena Hillman, Samira Jain, Craig M Macdonald, Elizabeth F Churchill, Carolyn Pang, Jofish Kaye, and Erick Oduor. 2023. Understanding and Evaluating UX Outcomes at Scale. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 466–469.

[39] I-Han Hsiao, Shuguang Han, Manav Malhotra, Hui Soo Chae, and Gary Natriello. 2014. Survey sidekick: Structuring scientifically sound surveys. In *Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings 12*. Springer, 516–522.

[40] Michael E Hyland, Sidney H Irvine, Clive Thacker, Peter L Dann, and Ian Dennis. 1989. Psychometric analysis of the Stunkard-Messick Eating Questionnaire (SMEQ) and comparison with the Dutch Eating Behavior Questionnaire (DEBQ). *Current Psychology* 8, 3 (1989), 228–233.

[41] Aulia Inan Nur, Harry B. Santoso, and Panca O. Hadi Putra. 2021. The method and metric of user experience evaluation: a systematic literature review. In *Proceedings of the 2021 10th International Conference on Software and Computer Applications*. 307–317.

[42] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[43] Mayank Kejriwal. 2022. Knowledge graphs: A practical review of the research landscape. *Information* 13, 4 (2022), 161.

[44] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1-2 (1938), 81–93.

[45] Suzanne Kieffer, Luka Rukonic, Vincent Kervyn de Meerendré, and Jean Vanderdonckt. 2019. Specification of a UX Process Reference Model towards the Strategic Planning of UX Activities.. In *VISIGRAPP (2: HUCAPP)*. 74–85.

[46] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[47] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[48] Emily Kuang, Minghao Li, Mingming Fan, and Kristen Shinohara. 2024. Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[49] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction.* Morgan Kaufmann.

[50] David Ledo, Steven Houben, Jo Vermeulen, Nicolai Marquardt, Lora Oehlberg, and Saul Greenberg. 2018. Evaluation strategies for HCI toolkit research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–17.

[51] Hao-Ping Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[52] Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1321–1340. doi:10.1145/3630106.3658975

[53] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2021. Exploring the effects of incorporating human experts to deliver journaling guidance through a chatbot. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[54] James R Lewis. 1992. Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the human factors society annual meeting*, Vol. 36. Sage Publications Sage CA: Los Angeles, CA, 1259–1260.

[55] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the*

2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590

[56] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[57] Q Vera Liao, Mihaela Vorvoreanu, Hari Subramonyam, and Lauren Wilcox. 2024. UX Matters: The Critical Role of UX in Responsible AI. *Interactions* 31, 4 (2024), 22–27.

[58] Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.

[59] Nicolai Marquardt, Steven Houben, Michel Beaudouin-Lafon, and Andrew D Wilson. 2017. HCITools: Strategies and best practices for designing, evaluating and sharing technical HCI toolkits. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 624–627.

[60] Sean McGregor. 2020. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. arXiv:2011.08512 [cs.CY] https://arxiv.org/abs/2011.08512

[61] Hilary McLellan. 2000. Experience design. *Cyberpsychology and behavior* 3, 1 (2000), 59–69.

[62] DTK Ng, WY Luo, HMY Chan, and SKW Chu. 2022. An examination on primary students' development in AI Literacy through digital story writing. *Computers & Education: Artificial Intelligence* 100054 (2022).

[63] Jakob Nielsen. 1992. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 373–380. doi:10.1145/142750.142834

[64] Jakob Nielsen. 1994. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 152–158.

[65] Jakob Nielsen. 1994. *Usability engineering*. Morgan Kaufmann.

[66] Hitesh Nigam, Mayur Patwardhan, Lovekesh Vig, and Gautam Shroff. 2024. Acceleron: A Tool to Accelerate Research Ideation. *arXiv preprint arXiv:2403.04382* (2024).

[67] Don Norman. 2007. *Emotional design: Why we love (or hate) everyday things*. Basic books.

[68] Stavroula Ntoa, George Margetis, Margherita Antona, and Constantine Stephanidis. 2021. User experience evaluation in intelligent environments: A comprehensive framework. *Technologies* 9, 2 (2021), 41.

[69] OpenAI, Josh Achiam, Steven Adler, and et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[70] FGWC Paas, Paul Ayres, and Mariya Pachman. 2008. Assessment of cognitive load in multimedia learning. *Recent Innovations in Educational Technology That Facilitate Student Learning, Information Age Publishing Inc., Charlotte, NC* (2008), 11–35.

[71] Fred GWC Paas. 1992. Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *Journal of educational psychology* 84, 4 (1992), 429.

[72] Rock Yuren Pang, Sebastin Santy, René Just, and Katharina Reinecke. 2024. BLIP: Facilitating the Exploration of Undesirable Consequences of Digital Technologies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[73] Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A model of social explanations for a conversational movie recommendation system. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. 135–143.

[74] Ingrid Pettersson, Florian Lachner, Anna-Katharina Frison, Andreas Riener, and Andreas Butz. 2018. A Bermuda triangle? A Review of method application and triangulation in user experience evaluation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–16.

[75] Frederick F Reichheld. 2003. The one number you need to grow. *Harvard business review* 81, 12 (2003), 46–55.

[76] Jeba Rezwana and Mary Lou Maher. 2023. User Perspectives on Ethical Challenges in Human-AI Co-Creativity: A Design Fiction Study. In *Proceedings of the 15th Conference on Creativity and Cognition* (Virtual Event, USA) *(Creativity and Cognition '23)*. Association for Computing Machinery, New York, NY, USA, 62–74. doi:10.1145/3591196.3593364

[77] Luis Rivero and Tayana Conte. 2017. A systematic mapping study on research contributions on UX evaluation technologies. In *Proceedings of the XVI Brazilian symposium on human factors in computing systems*. 1–10.

[78] Kerry Rodden, Hilary Hutchinson, and Xin Fu. 2010. Measuring the user experience on a large scale: user-centered metrics for web applications. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2395–2398.

[79] David J Roedl and Erik Stolterman. 2013. Design research at CHI and its applicability to design practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1951–1954.

[80] Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. Managing Bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 539–544. doi:10.1145/3308560.3317590

[81] Caitlin Sadowski, Margaret-Anne Storey, and Robert Feldt. 2019. A software development productivity framework. *Rethinking Productivity in Software Engineering* (2019), 39–47.

[82] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108*. (2017).

[83] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya G Parameswaran, and Ian Arawjo. 2024. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. *arXiv preprint arXiv:2404.12272* (2024).

[84] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. ConvXAI: Delivering Heterogeneous AI Explanations via Conversations to Support Human-AI Scientific Writing. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 384–387. doi:10.1145/3584931.3607492

[85] Ben Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–31.

[86] Dominik Siemon, Taras Rarog, and Susanne Robra-Bissantz. 2016. Semi-automated questions as a cognitive stimulus in idea generation. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 257–266.

[87] Lazar Stankov, Jihyun Lee, Wenshu Luo, and David J Hogan. 2012. Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and individual differences* 22, 6 (2012), 747–758.

[88] Åsne Stige, Efpraxia D Zamani, Patrick Mikalef, and Yuzhen Zhu. 2023. Artificial intelligence (AI) for user experience (UX) design: a systematic literature review and future research agenda. *Information Technology & People* (2023).

[89] Haowen Sun, Ruikun Zheng, Haibin Huang, Chongyang Ma, Hui Huang, and Ruizhen Hu. 2024. LGTM: Local-to-Global Text-Driven Human Motion Diffusion Model. In *ACM SIGGRAPH 2024 Conference Papers*. 1–9.

[90] Yan-Martin Tamm, Rinchin Damdinov, and Alexey Vasilev. 2021. Quality metrics in recommender systems: Do we calculate metrics consistently?. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 708–713.

[91] Felicia Fang-Yi Tan, Peisen Xu, Ashwin Ram, Wei Zhen Suen, Shengdong Zhao, Yun Huang, and Christophe Hurter. 2024. AudioXtend: Assisted Reality Visual Accompaniments for Audiobook Storytelling During Everyday Routine Tasks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[92] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2020. Expressions of style in information seeking conversation with an agent. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1171–1180.

[93] John Venable, Jan Pries-Heje, and Richard Baskerville. 2012. A comprehensive framework for evaluation in design science research. In *Design Science Research in Information Systems. Advances in Theory and Practice: 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, 2012. Proceedings 7*. Springer, 423–438.

[94] Shilpa Verma, Rajesh Bhatia, Sandeep Harit, and Sanjay Batish. 2023. Scholarly knowledge graphs through structuring scholarly communication: a review. *Complex & intelligent systems* 9, 1 (2023), 1059–1095.

[95] Arnold POS Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. 2010. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries*. 521–530.

[96] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone D Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications* 11, 1 (2020), 1–10.

[97] Thiemo Wambsganss, Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2020. A conversational agent to improve response quality in course evaluations. In *Extended Abstracts of the 2020 CHI conference on human factors in computing systems*. 1–9.

[98] Hongyu Wan, Jinda Zhang, Abdulaziz Arif Suria, Bingsheng Yao, Dakuo Wang, Yvonne Coady, and Mirjana Prpa. 2024. Building LLM-based AI Agents in Social Virtual Reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[99] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in*

computing systems. 1–6.
[100] Di Wang, Ming Yin, and Gloria Mark. 2021. Do you trust me to trust you? Mental model transparency and personalized explainability in human-AI collaboration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[101] Jinping Wang, Hyun Yang, Ruosi Shao, Saeed Abdullah, and S Shyam Sundar. 2020. Alexa as coach: Leveraging smart speakers to build social agents that reduce public speaking anxiety. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
[102] Ruijie Wang, Yuchen Yan, Jialu Wang, Yuting Jia, Ye Zhang, Weinan Zhang, and Xinbing Wang. 2018. Acekg: A large-scale knowledge graph for academic data mining. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1487–1490.
[103] Zhan Wang, Lin-Ping Yuan, Liangwei Wang, Bingchuan Jiang, and Wei Zeng. 2024. Virtuwander: Enhancing multi-modal interaction for virtual tour guidance through large language models. In *Proceedings of the CHI conference on human factors in computing systems*. 1–20.
[104] GB Willis and JT Lessler. 2013. Question appraisal system: QAS 99. National Cancer Institute.
[105] Gordon B Willis. 2004. *Cognitive interviewing: A tool for improving questionnaire design*. sage publications.
[106] Horvitz Eric Zeng Eric Xu, Wesley and Deirdre K Mulligan. 2022. Demystifying AI among end users: A taxonomy of AI literacy. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.
[107] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.
[108] Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11, 1 (2024), 28.
[109] Qingxiao Zheng and Yun Huang. 2023. " Begin with the End in Mind": Incorporating UX Evaluation Metrics into Design Materials of Participatory Design. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.
[110] Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 2022. UX research on conversational human-AI interaction: A literature review of the ACM digital library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

A Appendix

A.1 Definitions for Index

(1) **Paradigms:**
- *Dyadic:* One-on-one interaction between human and AI. Dyadic human-AI interaction, where end users interact with AI systems, and end users are not AI creators.
- *Polyadic:* Multiple social interactions mediated by AI. Where AI sits between two or many types of end users, and these end users are not AI creators either.

(2) **Application domain:** The specific field or sector in which the system is applied, such as education, healthcare, finance, urban planning.
(3) **Modality:** The form or channel through which users interact with the system, including Text-Based, Voice-Based, Multi-Modal (Text and Voice Based), Visual Based, Sensor Based, Haptic (touch) Based.
(4) **System Features:** The characteristics and attributes of the system, including embodiment, level of intelligence, autonomy, adaptability, and learning capabilities.
(5) **Design Novelty:** The uniqueness and innovativeness of the system's design, indicating how novel the system is compared to existing technologies.
(6) **Design methods:** Refers to the systematic approaches, strategies, and techniques employed in the creation and development of designs. These methods guide designers in the process of generating ideas, solving problems, and producing effective, functional, and aesthetically pleasing solutions. Example design methods can be:
- Framework-based design (proof-of-concept designs guided by prior frameworks, models, or theories)
- Participatory design
- Need-finding interviews
- Ideation workshops
- User-Centered Design
- Research through design
- Design fiction
- Human-Centered Design
- Contextual design
- Service design
- Inclusive design
- Ethnographic design
- Critical design
- Reflective design
- Sustainable design
- Wizard of Oz

(7) **Human-AI relationship types:** The nature of the interaction and relationship between humans and the system, ranging from assistant, collaborator, tool, to advisor.
(8) **Stakeholders:** Individuals or groups who have an interest in, are affected by, or can influence the design, development, deployment, and use of the system. This includes:
- Primary Users: Main direct users.
- Secondary Users: Occasional or indirect users.
- Tertiary Users: Affected non-users.
- Designers and Developers: System creators and maintainers.
- Organizations: Entities managing or funding the system.
- Policymakers: Regulators setting guidelines.
- Support Teams: Maintenance personnel.
- Researchers: Those studying and evaluating the system.

(9) **Social scale:** The scale at which the system operates, such as individual, group, organizational, societal.
(10) **Theoretical Frameworks:** The underlying theories and concepts that support the development and application of the system, which may include theories from cognitive science, social science, computer science.

A.2 Graph

Table 5 is the table for the graph database nodes and their attributes.

A.3 User Study Survey

(1) Background collection
- What's your name?
- What's your email address?
- What's your age? (optional)
- How do you identify your gender? Male, Female, Prefer not to say, Other, please specify
- What degree program are you currently enrolled in? Bachelor's Degree, Master's Degree, Doctorate (PhD or equivalent), Other, please specify

Node Type	Attributes	Explanation
Paper	Title	The research paper title
	Narrative	The narrative/abstract of the research paper
	Indexes	The results corresponding to each index category
	CommunityID	The communityID to which the research paper is classified into
	Narrative Embedding	The embedding of the research paper narrative
	Index Embedding	The embedding of each index value
Metric	Metric Name	The metric name category to which the metric is assigned
	Metric Type	
Outcome	Outcome Achieved	Metric outcome/result
	Paper Title	The research paper title
	Paper Citation	The reason for using the metric
	Metric Method	Method used to evaluate the metric
	Metric Usage	How the metric is used in the research paper
	Metric Challenges	The challenges or research gap identified in the research paper

Table 5: Graph Database Nodes and Their Attributes

- UX research experience: How many years of UX research experience do you have? None, less than 1 year, 2-3 years, 3-4 years, 5+ years
- Coding experience: How many years of coding experience do you have? None, less than 1 year, 2-3 years, 3-4 years, 5+ years
- AI literacy [62]: How strongly do you agree or disagree with the following statements?
  - I can operate AI applications in everyday life.
  - I can use AI meaningfully to achieve my everyday goals.
  - I can assess what the limitations and opportunities of using an AI are.
- AI Assistance in UX Research: Currently, what types of research tasks do you use AI (e.g., ChatGPT) to assist with? Research Ideation, Literature review, Evaluation plan development, Data analysis, Prototyping and design, Usability testing, Coding, Writing, Other please specify

(2) Survey 1 and 2 Initial/Final UX Study Plan: We are going to ask about your perceptions of the proposed UX plans for your project.
- How do you feel about the following statements about your proposed UX plans? Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree
  - Novelty [18]: My proposed UX plans are novel.
  - Relevance [18]: My proposed UX plans are relevant to my research goal.
  - Efficiency [93]: My proposed UX plans are efficient in supporting my research goals.
  - Completeness [18]: My proposed UX plans are complete.
  - Clarity [18]: My proposed UX plans are clear.
  - Concreteness [18]: My proposed UX plans are concrete.
  - Quality [18]: My proposed UX plans are of high quality.
  - Feasibility [18]: My proposed UX plans is workable and can be easily implemented.
  - Rigorous [93]: My proposed UX plans are methodologically rigorous.
  - Risk considerations [93]: My proposed UX plans are in consider of potential issues.
  - Confidence [87]: My proposed UX plans are making me feel confident about them.

(3) Survey 3 System Features: We are going to ask about your perceptions of each feature in our system. How do you feel about the following statements regarding the Index/ Metric Graph/ Metric List/ Outcome/ Risk feature? (We repeated the same set of questions for each feature.) Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree
- I enjoy using this feature. [15]
- I will use this feature in the future if it is publicly available [23].
- I find this feature useful [16].
- I trusted the suggestions and decisions provided by it [23].
- It took significant mental effort to understand this feature [71].
- I was more interested in exploring and developing my research projects/UX evaluation plan/UX outcomes after using it [13].
- This feature contributes to revising my UX evaluation plan.
- This feature contributes to revising my UX outcomes.

A.4 Demographic

Table 6 is the demographic table.

A.5 Data Analysis

Table 7 is the data analysis for changes in user perceptions towards the proposed UX plan result.

Table 6: Participants Demographic. LR = Literature Review, RI = Research Ideation, EPD = Evaluation Plan Development, PD = Prototyping & Design, DA = Data Analysis, C = Coding, W = Writing, UT = Usability Testing.

ID	Education Level	UX Experience	Coding Experience	AI Literacy [62]	AI Usage in Research
P1	Master's	3-4 years	1-2 years	4.33	RI, EPD, C, W.
P2	Master's	3-4 years	5+ years	5.00	RI, EPD, PD, DA, C, W.
P3	Doctorate	1-2 years	3-4 years	4.67	LR, PD, DA, C, W.
P4	Master's	3-4 years	5+ years	4.67	RI, LR, DA, C, W.
P5	Postdoc	1-2 years	3-4 years	4.00	LR, UT, C, W.
P6	Doctorate	1-2 years	3-4 years	5.00	PD, DA, C, W.
P7	Doctorate	1-2 years	5+ years	4.00	DA, C, W.
P8	Doctorate	3-4 years	5+ years	5.00	LR, DA, C, W.
P9	Doctorate	1-2 years	3-4 years	4.00	RI, C, W.
P10	Doctorate	3-4 years	3-4 years	4.00	DA, C, W.
P11	Doctorate	3-4 years	5+ years	3.67	RI, DA, C, W, Other
P12	Master's	1-2 years	3-4 years	5.00	RI, LR, EPD, PD, DA, C, W.
P13	Doctorate	1-2 years	3-4 years	4.00	C, W.
P14	Bachelor's	3-4 years	3-4 years	5.00	LR, PD, DA, C, W.
P15	Doctorate	3-4 years	5+ years	4.00	C, W.
P16	Doctorate	< 1 year	5+ years	4.33	Other
P17	Doctorate	3-4 years	5+ years	4.67	Other
P18	Doctorate	3-4 years	5+ years	5.00	RI, LR, PD, DA, C, W.
P19	Master's	3-4 years	5+ years	4.00	PD, C, W.

Table 7: Descriptive Analysis and Wilcoxon Signed Rank Test Results for Changes in User Perceptions Towards a Proposed UX Plan (N = 19). Significant values: $^{**}p < 0.01$, $^{*}p < 0.05$.

Constructs	Before Median	Before IQR	After Median	After IQR	Test Statistic	Effect Size r_b
Novelty	3	2	4	1	22	0.51
Relevance	5	1	5	0.5	8	0.56
Efficiency*	4	0	5	1	12	0.74
Completeness**	3	2	4	1	16	0.79
Clarity**	4	1	4	1	4	0.90
Concreteness*	4	1	4	1.5	24	0.74
Quality**	3	1	4	0.5	11	0.82
Workability*	4	2	5	2	16.5	0.64
Rigor*	3	2	4	1.5	15	0.75
Ethics**	3	2.5	4	1.5	10	0.81
Trust*	4	1	4	1	5	0.82
Confidence*	3	1	4	1	18	0.70