Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing[1]

Book Report

1. Introduction

1.1 Research Background

With the proliferation of digital products, the importance of user experience (UX) has become increasingly prominent. Usability testing serves as a key method to enhance UX by analyzing user behavior in interactive systems to identify potential issues. However, analyzing usability testing videos is a complex and resource-intensive task that requires evaluators to observe both user behavior and audio signals simultaneously, quickly identifying usability problems across multiple tasks. In industrial settings, limitations in time and resources can lead to the omission or misunderstanding of critical information. Although collaborative analysis by multiple individuals can improve the reliability and completeness of the analysis, the high cost of collaboration has hindered UX evaluators from adopting this method to some extent. Given the rapid development of AI technology, researchers have begun to explore how AI-driven analysis can provide a supplementary perspective for UX evaluators, especially through human-machine collaborative analysis enabled by natural language interfaces.

1.2 Research Objectives and Importance

This study focuses on the application of proactive conversational AI assistants in UX evaluation, exploring the impact of their automatic suggestions at different time points (before the problem occurs, simultaneously, and after the problem occurs) on the analysis behavior of UX evaluators. The significance of studying this issue lies in understanding the optimal timing of suggestions to guide the design and functionality of future tools, thereby enhancing AI-assisted UX evaluation decisions.

2. Related Research

2.1 Research on Using AI to Detect Usability Issues

2.1.1 Automated Methods:

Automated methods include machine learning, pattern recognition, audio and video analysis, and natural language processing. These methods alleviate the burden of manual analysis to some extent. However, the limitation of automated methods is their difficulty in fully replacing human evaluators' intuition and contextual understanding capabilities. There is still a lack of recognition regarding ChatGPT's effectiveness in identifying usability issues. Our strategy for using ChatGPT to generate usability issue suggestions requires an assessment of the quality of its output.

2.1.2 Human-Machine Collaboration Methods:

AI provides information to humans through an "algorithm-in-the-loop" process, supporting rather than replacing human decision-making. However, existing UX evaluation tools primarily provide non-interactive visualizations, limiting evaluators' ability to ask questions and seek explanations.

2.2 Research on Proactive Dialogue

Proactive dialogue can remind users of potentially missed information and provide suggestions to assist in decision-making. In various fields such as education, health management, and decision support, proactive dialogue has been proven to improve user satisfaction and trust. However, inappropriate proactive interactions can distract users or generate distrust. Therefore, it is necessary to carefully design the timing and content of interactions to ensure that the messages are timely and reasonable.

3. Research Methods

Among the commonly used research methods, this paper employs experimental, survey, and literature review methods. Here, we mainly discuss the applicability and limitations of the first two. The experimental method includes AI-generated problem suggestion methods and the Wizard of Oz method:

The former is suitable for scenarios requiring the rapid generation of a large number of problem suggestions, such as identifying usability issues in UX research. It can process natural language text, analyze transcripts, and provide potential problem points for researchers. However, GPT can only access users' verbal content and cannot obtain visual information from videos, which may result in suggestions being inaccurate or incomplete. Moreover, it may have difficulties in recognizing complex or context-dependent problems, requiring manual proofreading and editing.

The latter is suitable for research requiring real-time interaction with participants, such as real-time interaction in UX research. It ensures consistent responses and focuses on investigating the impact of specific variables (such as the timing of automatic suggestions). However, it introduces human intervention factors, which may affect the naturalness of the research. Additionally, it has high requirements for the host, needing pre-set scripts and categorization of potential questions from participants. The experiment also has high contingency due to the small sample size.

The experimental designs of the three methods are rigorous, collecting a large amount of data. All indicators have quantitative calculations based on improvements over previous theories, considering variables overlooked by predecessors and being highly targeted. However, the relatively small sample size of participants (24 participants) may affect the universality of the results. Moreover, it is also mentioned that due to related costs in terms of time, resources, and energy, evaluators have limited adoption of collaborative practices. Therefore, future efforts should strive to balance efficiency and robustness.

4. Experimental Results

4.1 Existing Experimental Results and Findings

The study found that the timing of suggestions had no significant impact on the number of issues identified by evaluators, but suggestions after the problem occurred significantly improved trust and efficiency. Most evaluators preferred suggestions after the problem occurred as they helped verify their analysis.

In addition, evaluators generally agreed with the suggestions generated by ChatGPT but deemed them not comprehensive enough. For highly consistent suggestions, evaluators tended to directly confirm them; for low-consistency suggestions, evaluators sought clarification or ignored them.

Although ChatGPT can identify some usability issues, it missed most of the issues identified by evaluators (58.8%). This is mainly because ChatGPT relies solely on text information spoken by users and lacks visual and interactive information from videos.

4.2 Implications and Discussion

The findings of this study emphasize the importance of human-machine collaboration in UX evaluation, especially in leveraging AI to improve evaluation efficiency and reliability. At the same time, the results also point out the limitations of current AI tools (such as ChatGPT) in UX evaluation, suggesting that future research should focus on how to combine multimodal data (such as video and audio) to improve AI's analytical capabilities.

Researchers have empirically validated the potential of proactive conversational AI assistants in UX evaluation, particularly providing valuable insights into the impact of suggestion timing on evaluator behavior. However, the study also reveals the limitations of current AI tools, especially their inadequacies when handling complex UX evaluation tasks. Future research should further explore how to combine multimodal data to enhance AI's analytical capabilities and develop more intelligent and user-friendly human-machine collaboration tools.

4.3 Possible Future Directions and Challenges

Based on this study, future research can focus on the following areas:

(1) Multimodal Data Analysis: Combining video, audio, and text data to leverage deep learning technology to improve AI's ability to identify UX issues.

(2) More Personalized Suggestions: Providing personalized suggestion timing and content based on evaluators' experience and preferences to improve collaboration efficiency.

(3) Considering Scalable Design: Adapting to longer usability testing videos.

(4) Balancing Efficiency and Robustness: Providing detailed explanations for AI suggestions to

help evaluators understand the reasons behind them, thereby enhancing trust. Providing training for evaluators to help them better understand and utilize AI suggestions and collecting feedback to continuously optimize the tools.

5. Insights

Besides the quality of AI technology itself, the timing of its responses also has a significant impact on the credibility of suggestions. This study does not directly contribute to improving AI performance but provides a path for optimizing AI from a human-computer interaction perspective. In addition, the complexity and comprehensiveness of this study exceeded my expectations. It covers psychology, statistics, computer science, and many other disciplines, making numerous improvements over previous research and providing convincing analysis and calculations of survey data.

In the rapid development of AI, many fields it involves are also undergoing rapid updates, with more and more research being conducted. Human-computer interaction is an important aspect of AI, and therefore research on AI's human-computer interaction is also booming. I believe that in the future, the interaction between artificial intelligence and humans will reach the level of human-to-human interaction, or even surpass it.

References

[1] Kuang Emily, Li Minghao, Fan Mingming, Kristen Shinohara. Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing[C]. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24), 2024: 1-16.