Usability and User Experience, Vol. 232, 2026, 286–296
https://doi.org/10.54941/ahfe1008057

Application Potential of Large Language Models as Product User Experience Evaluation Tools

Xiaoyue Mao, Jun Zhang, Yijing Yang, and Kaiyang Tang

School of Design of Hunan University, Hunan, 410082, China

ABSTRACT

This study aims to explore the potential of general-purpose Large Language Models (LLMs) in generating User Experience (UX) evaluations during the product verification phase, addressing issues in traditional UX research methods such as difficulties in user recruitment and long scheduling cycles. Using the User Experience Honeycomb model as the theoretical framework, the research selects the off-the-shelf GPT-4o as the experimental model. By combining optimized prompt engineering with multimodal inputs, a comparative analysis is conducted on the similarities and differences between LLMs and human users regarding evaluation coverage rate, language style, and problem perspectives. The experiment employs a deductive-inductive approach to code and analyze the collected evaluation data. The results indicate that the thematic overlap rate between LLM-generated evaluations and human user evaluations reaches 81.05%, demonstrating significant potential in simulating human users to output experience evaluations. Textual analysis reveals that LLM-generated UX evaluations exhibit strengths in systematic analysis, professional expression, and proactive risk identification; however, they show limitations in capturing nuanced emotions and dynamic interaction details. Additionally, the efficiency of UX evaluation is improved by 88.0% compared to human users. The study recommends adopting a Hybrid Intelligence evaluation model, leveraging the systematic analysis capabilities of LLMs while incorporating human users' acute perception of emotions and immediate experiences to enhance both the efficiency and comprehensiveness of UX research.

Keywords: User experience, Large language models, Product verification, User experience Honeycomb, Hybrid intelligence

INTRODUCTION

User Experience (UX) has become a core element of mobile application design and a critical pathway to market advantage (Obrist et al., 2009). In the design process, the concept phase is vital for validating user needs and experience. While traditional methods like user interviews are standard for assessing usability and satisfaction, they suffer from recruitment difficulties, long feedback cycles, and high costs (Law et al., 2009). These inefficiencies are particularly detrimental in the rapidly iterating mobile market, often forcing design teams to slow down or proceed without sufficient validation—leading to accumulated design flaws or severe directional deviations (Basri et al., 2016).

Received December 4, 2025; Revised January 6, 2026; Accepted January 26, 2026; Available online July 20, 2026

The proliferation of Large Language Models (LLMs) is driving a paradigm shift in UX research (Baghela, 2024). By simulating user cognition and emotional responses, LLMs can rapidly generate diverse feedback in the early design stages (Salewski et al., 2023). This high efficiency and flexibility offer distinct advantages for mobile application verification, particularly in scenarios requiring rapid responses to market changes.

However, existing studies primarily focus on theoretical discussions or rely on customized systems, lacking systematic validation of general-purpose LLMs in practical application scenarios (Xiang et al., 2024). Furthermore, the accuracy, stability, and alignment of LLM-generated content with human evaluation remain critical unresolved issues.

Consequently, this study explores the potential of general-purpose LLMs for UX evaluation during the mobile product verification phase. Using Peter Morville's User Experience Honeycomb as the theoretical framework and the widely available GPT-4o model, we conduct a comparative analysis of LLMs versus human users regarding evaluation coverage rate, linguistic features, problem perspectives, and efficiency. This research aims to define the application potential and operational models of LLMs as effective tools for UX evaluation.

APPLICATION STATUS AND CHALLENGES OF LLMS IN THE USER EXPERIENCE FIELD

Traditional User Experience Research Methods

In mature consumer markets, high-quality User Experience (UX) has become a core competency for product development (Obrist et al., 2009). Traditional UX research emphasizes human participation to assess usability and satisfaction. Common methods—including user interviews, usability testing, questionnaires, and focus groups—play distinct roles across development stages (Law et al., 2009, Kujala, 2003). User interviews, a classic qualitative method, are particularly valuable for capturing user motivations and emotional conflicts, revealing implicit logic often missed by quantitative surveys (Krug and Don't Make Me Think, 2014). Empirical studies indicate that interviews help identify "experience gaps" through narrative feedback and uncover cognitive blind spots regarding emerging technologies in complex scenarios (Krug and Don't Make Me Think, 2014, Blandford and Green, 2001).

However, traditional methods face significant limitations. They entail high time and resource costs, with recruitment difficulties increasing sharply with sample size (Krug and Don't Make Me Think, 2014). Subjective evaluations are prone to instability due to memory bias, social desirability effects, or individual cognitive differences (Nielsen and Molich, 1990). Furthermore, users may conceal true needs due to security concerns, making actual behavior difficult to predict (Xiang et al., 2024). In the rapidly iterating mobile application sector, these inefficiencies often fail to meet the demand for high-frequency assessment, leading to delayed iterations and accumulated design issues (Basri et al., 2016).

Potential and Challenges of LLMs in UX Evaluation

LLMs possess unique potential in simulating user behavior and generating feedback. For instance, role-playing prompts can simulate specific demographic language styles and cognitive traits, while in-context learning allows for inferring user emotional intent (Salewski et al., 2023, Xu et al., 2023). LLMs also demonstrate capabilities in emotional analysis and alignment with human perception regarding design elements and aesthetics (Weitl-Harms et al., 2024). Furthermore, they can simulate interactions with mobile apps to generate usability feedback and identify issues in extreme scenarios (Xiang et al., 2024). However, as many existing studies rely on complex, team-specific systems, this study employs a cost-effective, capable general-purpose LLM to explore its practical value in UX evaluation.

LLMs also face application challenges, including limitations in multimodal understanding (image/text), reasoning stability, and sensitivity to prompt engineering (Liu et al., 2021, Zamfirescu-Pereira et al., 2023). This study utilizes optimized prompt engineering and refined input materials to guide the LLM toward generating in-depth evaluations.

Conversely, LLMs may introduce negative impacts, such as reduced credibility due to stochastic evaluation biases or output deviations stemming from inherent cultural and social biases (Kolisko and Anderson, 2023). To mitigate these risks, this research weighs the defects of LLMs against their advantages, exploring the complementarity between LLMs and human UX.

RESEARCH METHODS

First, a coding framework was derived from a literature review of the User Experience Honeycomb model. This framework guides the clustering and extraction of themes from evaluation corpus data to support comparative analysis.

Second, based on this framework, an experimental question outline was designed for both human users and LLMs to collect evaluation data.

Subsequently, a deductive-inductive approach was employed by multiple researchers to code the collected data. Inter-rater reliability was verified using the Kappa coefficient to ensure the robustness of the coding rules.

Finally, a descriptive comparative analysis was conducted to examine the coverage rate and corpus differences (presented as themes) between LLM and human coding results, followed by an analysis of evaluation efficiency.

Evaluation Framework: User Experience Honeycomb

This study adopts Peter Morville's User Experience Honeycomb as the analytical framework, comprising six dimensions: Useful, Usable, Desirable, Findable, Accessible, and Credible (Value was excluded as it relies on market data). This model was selected because: (1) it supports systematic quality assessment of satisfaction (Kim, 2020); (2) it is applicable to the full product lifecycle with independent yet interrelated dimensions (Morville, 2005); and (3) it has been validated across diverse fields including AI, education,

healthcare, and urban planning (Fengyan, 2019) (Meifen, 2017) (Kim et al., 2024, Lee et al., 2021). Thus, the model's versatility and breadth make it an appropriate framework.

Coding Method and Consistency Assessment

A deductive-inductive approach was used to analyze UX evaluations from both data sources (Humans and LLMs) under a unified framework.

During data processing, raw data was cleaned and manually annotated. Coding was performed independently by the research team using a deductive framework based on the Honeycomb model (see Fig. 1). This framework consists of three levels: Category, Raw Data, and Theme. The theme level uses a structured format ("Interface ID - Object - Attribute") to precisely categorize experience issues (McDonald et al., 2019) (Wicks, 2017). The Kappa coefficient was utilized to assess coding consistency and reliability (Lidong et al., 2023).

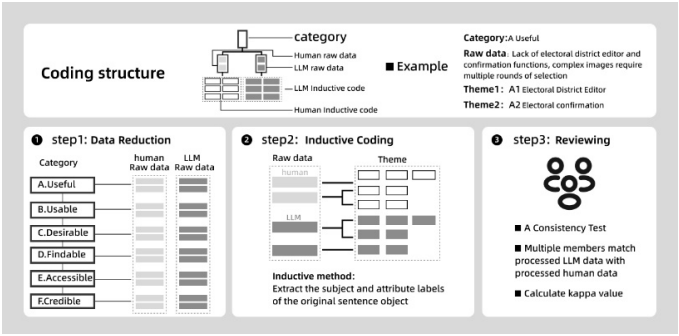

Figure 1: Coding process.

EXPERIMENTAL DESIGN AND IMPLEMENTATION

Experimental Targets and User Personas

This study targets two mobile applications developed by a partner enterprise—AI Call Summary and AI Image Eraser—conducting UX evaluation experiments based on user personas defined by enterprise requirements (see Fig. 2).

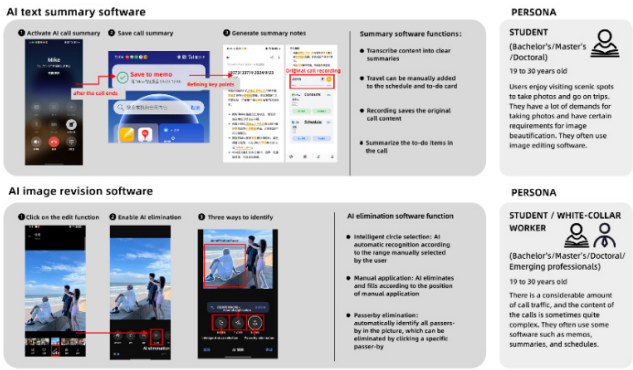

Figure 2: Introduction to experimental objects and user personas.

Experimental Sample Size

For human testing, 12 target users were recruited for each application, totaling 24 participants. The LLM experiment established four sample size gradients ranging from 1 to 4 times the human sample size (totaling 96 experiments, with 48 per application). This setup aims to evaluate the variation patterns of LLM coverage rate and determine the optimal generation scale that balances efficiency with coverage quality.

Human User Experience Evaluation Procedure

Based on the target user groups (see Fig. 2), participants matching the user personas were recruited. The experiment employed structured interviews to identify user pain points (Lidong et al., 2023), utilizing a question set designed around typical task flows and the six dimensions of the User Experience Honeycomb. Users completed basic task flows and answered the question set, with an average experiment duration of 2.4 hours. Researchers recorded raw evaluation data in real-time, classifying it according to the Honeycomb dimensions for subsequent inductive coding.

LLM User Experience Evaluation Design

GPT-4o was selected as the experimental LLM due to its superior performance in zero-shot tasks and text generation (Kalai and Vempala, 2024). In multiple benchmarks, GPT-4o outperforms similar models in accessibility and stability, demonstrating strong reasoning and cross-modal image recognition capabilities for complex tasks without additional training data (Kalai and Vempala, 2024). Furthermore, its market maturity ensures ease of use for designers.

Prompt engineering was optimized through literature review and preliminary experimentation. First, a Chain-of-Thought (CoT) approach was adopted to ensure consistent and in-depth analysis (Wei et al., 2022) (see Fig. 3). Drawing on OpenAI's design principles, a clear, structured prompt framework was formulated to elicit high-quality feedback from limited input (Schulhoff et al., 2024). Additionally, a preset correction step was implemented following the LLM's initial understanding to ensure experimental controllability and accuracy (Liu et al., 2024, Schulhoff et al., 2024). Finally, requiring the LLM to restate its interpretation of input indicators before scoring was found to enhance understanding (Deng et al., 2023).

Based on the sociological information presentation paradigm and the characteristics of the target software (Yang et al., 2024), static image-text materials were created for LLM understanding. During the preparation phase, optimal input forms were determined through multiple iterations, excluding ineffective methods. Ultimately, four types of image-text materials were formed based on the Honeycomb dimensions: flowcharts, output results, input-output correspondence diagrams, and interface screenshots.

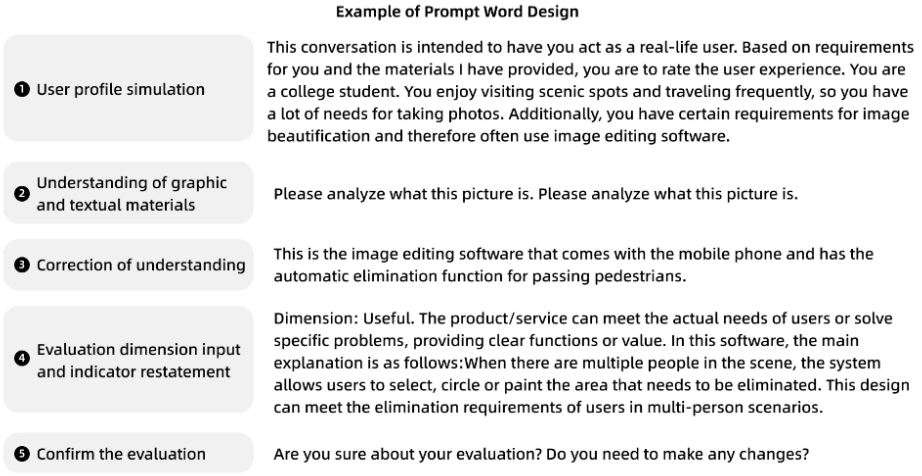

Figure 3: Examples of LLM prompts - AI image eraser software as an example.

LLM Experimental Procedure and Data Collection

First, GPT-4o was selected and its dialogue parameters configured. Subsequently, a structured prompt framework based on the User Experience Honeycomb was designed, employing a zero-shot CoT method to guide multi-dimensional reasoning. On this basis, product information was provided to the model via the multi-type input materials. Finally, evaluation results were collected following a gradient sampling strategy and organized by the Honeycomb dimensions to form a qualitative dataset for subsequent analysis (see Fig. 1).

EXPERIMENTAL RESULTS ANALYSIS

Verification Results of Coding Consistency

To assess coding reliability, inter-rater reliability was rigorously tested using Cohen's Kappa coefficient across 85 valid cases coded by two trained researchers. The analysis yielded a Kappa value of 0.857 ($p < 0.001$). According to Landis and Koch's widely accepted standards, a value between 0.81 and 1.00 indicates "almost perfect" agreement. This statistically significant result confirms the operational robustness and consistency of the coding framework, providing a solid foundation for subsequent analyses (see Table 1).

Table 1: Kappa coefficient results.

	Value	Asymptotic Standard Error (a)	Approximate T (b)
Measure of Agreement Kappa	.857	.038	78.373
N of Valid Cases	85		

Coverage rates were derived by calculating the thematic overlap between LLM and human coding. Analysis reveals that at a sample size 4x that of human users, the LLM achieved a coverage rate of 81.05%. This demonstrates a high degree of similarity to human UX evaluation and validates the LLM's capability to effectively identify experience issues. A systematic analysis of LLM performance across sample sizes (1x to 4x) indicates a logarithmic growth trajectory: coverage increased from 62.1% (1x) to 75.85% (2x), 78.55% (3x), and finally 81.05% (4x) (see Fig. 4). Both the coverage rate and the number of new evaluation topics identified followed this logarithmic trend, indicating diminishing marginal returns. Consequently, the 4x sample size is identified as the optimal generation scale, effectively balancing high thematic coverage (>80%) with generation efficiency.

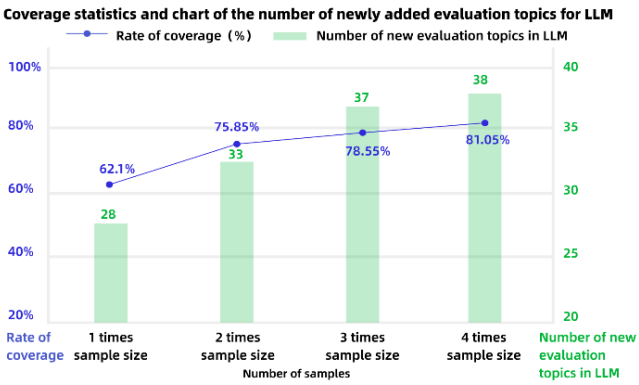

Figure 4: Number of UX evaluation theme codes for LLMs and human users.

Analysis Results of Corpus Differences

From an evaluative perspective, LLMs demonstrated superior systematic analysis, whereas human users focused on specific, immediate pain points. For instance, in the Credible dimension, the LLM noted general "limitations in filling flaws within complex textures or lighting," while humans pinpointed specific failures like "inaccurate passerby" or "hair recognition." Similarly, in the Accessible dimension, the LLM summarized errors broadly under "complex backgrounds," whereas humans highlighted granular details such as "statues misidentified as pedestrians" or "selection areas merging." Furthermore, the LLM exhibited a distinct advantage in proactive risk identification. It predicted potential issues like "error accumulation in long calls" or "impact of speaker count"—risks largely overlooked by human participants.However, the data also highlights the LLM's limitations: it struggles to simulate nuanced emotions and capture dynamic interaction details, resulting in evaluations that cannot fully replicate the human user experience.

Comparative Analysis of Experimental Efficiency

To assess efficiency, time costs were analyzed across preparation and formal experimentation phases (see Table 2).

Table 2: Comparison of experimental efficiency between human users and LLMs.

Comparison Dimension	Human User Experiment	LLM Experiment	Efficiency Difference Analysis
Pre-preparation Time	Not statistically precise (Recruitment, screening, scheduling approx. 6.5 days)	11 hours 48 minutes (Image-text material production)	Human experiments incur significant long-cycle time losses.
Formal Experiment Time	52.6 hours	6.3 hours (including system latency)	LLM execution is 8.4x faster, improving efficiency by approx. 88.0%.
Time Growth Pattern	Linear Growth (Adding 1 subject increases time linearly)	Logarithmic Growth (Low marginal cost for additional samples)	LLMs demonstrate superior efficiency in large-scale verification.

HOW LLMS GUIDE USER EXPERIENCE EVALUATION PRACTICE IN THE RAPID ITERATION STAGE OF MOBILE APPLICATION PRODUCTS

To address critical industry pain points in traditional UX research during rapid mobile application iteration—specifically recruitment difficulties, long cycles, and high costs that lead to accumulated design flaws and directional drift—this study constructs a "Hybrid Intelligence Product UX Evaluation Model" based on the widely accessible general-purpose LLM, GPT-4o. Leveraging the LLM's immediate response and low barrier to entry, this model provides an efficient evaluation paradigm tailored for high-frequency iterations.

Hybrid Intelligence Product User Experience Evaluation Model

This study proposes a funnel-shaped Hybrid Intelligence evaluation model designed to maximize assessment efficacy. Experimental results confirm that LLMs effectively identify UX issues, surpassing average users in systematic analysis and proactive risk identification, although they fall short in capturing nuanced emotions and dynamic interaction details (see Fig. 5).

Consequently, a collaborative, layered mechanism is established:

Phase 1: Utilize the breadth of LLMs for a comprehensive "carpet-style" screening to rapidly perform systematic analysis and proactive risk identification. Using a sample size 4x that of the target human user group is identified as the optimal generation scale, balancing coverage quality with efficiency. The resulting "Experience Issue Diagnosis" provides objective, universal insights, enabling teams to correct fundamental structural and framework defects within minimal cycles.

Phase 2: Recruit a small cohort of human users for interviews. This phase focuses on immediate pain points within complex scenarios, specifically nuanced emotional feedback and dynamic details. Simultaneously, ambiguous issues identified in Phase 1 are converted into specific observation tasks for secondary verification by real users. The results encompass Desirability feedback and logic review, creating a dual assurance of "machine rationality and human perception" to maintain experience precision during rapid iteration.

Retrospective analysis of the experimental corpus empirically validated the effectiveness of this funnel-shaped "rational initial screening by LLMs + emotional re-verification by humans" model.

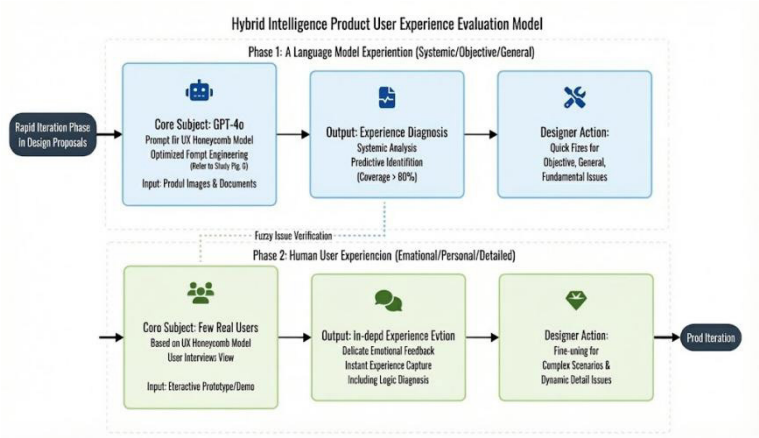

Figure 5: Hybrid intelligence product user experience evaluation model.

CONCLUSION

Based on the User Experience Honeycomb model, this study explored the UX evaluation capabilities of general-purpose LLMs during the mobile application product verification phase, as well as their differences and evaluation efficiency compared to human users, through deductive-inductive and corpus analysis methods. The experimental results indicate that under the condition of 4 times the human sample size, the LLM's theme induction coverage rate exceeded 80%, demonstrating its ability to effectively identify UX issues. It performed particularly better than human users in systematic analysis, professional expression, and risk prediction. Meanwhile, the LLM improved evaluation efficiency by approximately 88.0% compared to human users and demonstrated the advantage of low marginal cost in large-scale verification. Comprehensive analysis suggests that general-purpose LLMs have significant application potential as tools for product UX evaluation.

However, LLMs still have limitations in understanding dynamic interaction details and feeding back nuanced emotions. Their evaluation content differs from real user feedback in terms of emotional feedback, subjective experience description, and capturing dynamic interaction details.

Therefore, the study recommends adopting a Hybrid Intelligence Product UX Evaluation Model in practical applications. First, utilize LLMs for a "carpet-style" broad screening to quickly discover systematic product experience issues and identify proactive risks. Then, use a small number of human users to conduct in-depth verification focusing on nuanced emotions, dynamic interaction details, and vague issues from the previous step. Through this funnel-shaped UX evaluation path of "rational initial screening by LLMs + emotional re-verification by humans," a more comprehensive and efficient UX evaluation pathway is constructed.

REFERENCES

Baghela, R. M. P. D. V. S. 2024. The Future of LLMs in Personalized User Experience in Social Networks. IRE Journals, Volume 8, Issue 5, November-2024, 920–951.

Basri, N. H., Noor, N. L. M., Adnan, W. A. W., Saman, F. M. & Baharin, A. H. A. Conceptualizing and understanding user experience. 2016 4th International Conference on User Science and Engineering (i-USEr), 2016. IEEE, 81–84.

Blandford, A. E. & Green, T. R. 2001. Group and individual time management tools: what you get is not what you need. Personal and Ubiquitous Computing, 5, 213–230.

Deng, Y., Zhang, W., Chen, Z. & Gu, Q. 2023. Rephrase and respond: Let large language models ask better questions for themselves. arXiv preprint arXiv:2311.04205.

Fengyan, Q. 2019. Study on the Strategy of Learning APP User Experience Optimizing in Chinese University MOOC. master, Nanjing University of Posts and Telecommunications.

Kalai, A. T. & Vempala, S. S. Calibrated language models must hallucinate. Proceedings of the 56th Annual ACM Symposium on Theory of Computing, 2024. 160–171.

Kim, D., Park, S. & Choo, J. When model meets new normals: test-time adaptation for unsupervised time-series anomaly detection. Proceedings of the AAAI conference on artificial intelligence, 2024. 13113–13121.

Kim, N.-H. 2020. User experience validation using the honeycomb model in the requirements development stage. International journal of advanced smart convergence, 9, 227–231.

Kolisko, S. & Anderson, C. J. Exploring social biases of large language models in a college artificial intelligence course. Proceedings of the AAAI Conference on Artificial Intelligence, 2023. 15825–15833.

Krug, S. & Don't Make Me Think, R. 2014. A common sense approach to web usability. New Riders, 3.

Kujala, S. 2003. User involvement: a review of the benefits and challenges. Behaviour & information technology, 22, 1–16.

Law, E. L.-C., ROTO, V., Hassenzahl, M., Vermeeren, A. P. & Kort, J. Understanding, scoping and defining user experience: a survey approach. Proceedings of the SIGCHI conference on human factors in computing systems, 2009. 719–728.

Lee, Y., Chung, J. J. Y., Song, J. Y., Chang, M. & Kim, J. Personalizing ambience and illusionary presence: How people use "study with me" videos to create effective studying environments. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021. 1–13.

Lidong, Q., Liping, Y., Jiamin, C., Dengpan, Z., Mei, S. & Xihe, L. 2023. The Inter-Coder Consistency in Qualitative Research. Psychological Science, 46, 760–767.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L. & Chen, W. 2021. What Makes Good In-Context Examples for GPT-$3 $? arXiv preprint arXiv:2101.06804.

Liu, Y., Wang, Y., Sun, L. & Yu, P. S. 2024. Rec-gpt4v: Multimodal recommendation with large vision-language models. arXiv preprint arXiv:2402.08670.

Mcdonald, N., Schoenebeck, S. & Forte, A. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. Proceedings of the ACM on human-computer interaction, 3, 1–23.

Meifen, C. 2017. *Study on Relations Between UserExperience and Learning Motivationof Massive Online Courses*. doctor.

Morville, P. 2005. Ambient findability: What we find changes who we become, " O'Reilly Media, Inc."

Nielsen, J. & Molich, R. Heuristic evaluation of user interfaces. Proceedings of the SIGCHI conference on Human factors in computing systems, 1990. 249–256.

Obrist, M., Roto, V. & Väänänen-Vainio-Mattila, K. 2009. User experience evaluation: do you know which method to use? CHI'09 Extended Abstracts on Human Factors in Computing Systems.

Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E. & Akata, Z. 2023. In-context impersonation reveals large language models' strengths and biases. Advances in neural information processing systems, 36, 72044–72057.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H. & Schulhoff, S. 2024. The prompt report: A systematic survey of prompting techniques. arXiv preprint arXiv:2406.06608, 5.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V. & Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824–24837.

Weitl-Harms, S., Hastings, J. D. & Lum, J. Using LLMs to establish implicit user sentiment of software desirability. 2024 International Conference on Machine Learning and Applications (ICMLA), 2024. IEEE, 1645–1650.

WIcks, D. 2017. The coding manual for qualitative researchers. Qualitative research in organizations and management: an international journal, 12, 169–170.

Xiang, W., Zhu, H., Lou, S., Chen, X., Pan, Z., Jin, Y., Chen, S. & Sun, L. SimUser: Generating Usability Feedback by Simulating Various Users Interacting with Mobile Applications. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024. 1–17.

Xu, B., Yang, A., Lin, J., Wang, Q., Zhou, C., Zhang, Y. & Mao, Z. 2023. Expertprompting: Instructing large language models to be distinguished experts. arXiv preprint arXiv:2305.14688.

Yang, Y., Li, Z., Dong, Q., Xia, H. & Sui, Z. 2024. Can Large Multimodal Models Uncover Deep Semantics Behind Images? arXiv preprint arXiv:2402.11281.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B. & Yang, Q. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. Proceedings of the 2023 CHI conference on human factors in computing systems, 2023. 1–21.