Transforming Interactive Systems with Large Language Models: Accelerating Interface Design and Evaluation

Stefano Zeppieri

[1]*Department of Computer Science, Sapienza University of Rome, Rome, Italy*

Abstract

This research investigates how Large Language Models (LLMs) can be integrated into the design and evaluation of interactive systems. It focuses on two main directions: (1) augmenting usability evaluation by simulating user interactions and conducting AI-assisted cognitive walkthroughs, and (2) supporting early-stage design through adaptive persona generation and need-finding. Through a combination of structured prompting, iterative prototyping, and empirical validation, the work explores how LLMs can surface usability issues, suggest design improvements, and complement traditional human-centered methods.

Preliminary findings indicate that LLM-assisted evaluations can offer practical benefits during early design stages, especially when direct access to users is limited. However, challenges remain—such as handling multimodal interfaces, validating AI-generated feedback, and accounting for context-specific usability concerns. To address these, the research adopts a hybrid approach that blends automated analysis with expert review and user testing.

The overarching goal is to build a scalable and interpretable evaluation framework that helps designers identify and address usability problems early, without compromising on the principles of transparency, adaptability, and user focus that define human-computer interaction.

Keywords

LLM, Interactive Systems, Usability Evaluation, Cognitive Walkthrough, AI-Assisted Design, User Simulation, Interface Evaluation

1. Introduction

I am Stefano Zeppieri, a third-year PhD student at the Department of Computer Science, Sapienza University of Rome, under the supervision of Prof. Emanuele Panizzi. My research explores how Large Language Models (LLMs) can be used to support the design and evaluation of interactive systems. This paper, prepared for the CHItaly 2025 Doctoral Consortium, outlines the context, motivation, and current direction of my work, which lies at the intersection of Human-Computer Interaction (HCI) and Artificial Intelligence.

The emergence of powerful LLMs such as GPT-4 has introduced new possibilities for augmenting key stages of the design process. These models can generate text, simulate users, extract structure from unstructured inputs, and reason about interaction flows—capabilities that lend themselves well to tasks such as user modeling, interface evaluation, and early-stage ideation. My research investigates how to make these capabilities practically useful to designers, focusing on how LLMs can generate adaptive personas, simulate user interactions, and provide structured feedback on prototypes.

Over the past two years, I have explored LLM-based tools that assist in evaluating interactive systems through virtual Cognitive Walkthroughs and AI-driven simulations. These tools aim to identify usability issues early in the development cycle, before costly user testing takes place [1, 2]. They are not meant to replace traditional evaluation methods, but to complement them—particularly in the early design stages, where access to users may be limited and iteration speed is critical.

My earlier research focused on implicit interaction in mobile and in-vehicle interfaces [3, 4, 5, 6]. This work dealt with how systems can adapt to user behavior without relying on explicit input, and it shaped my interest in building interactive systems that are context-aware, adaptive, and user-sensitive.

CHItaly 2025: Technologies and Methodologies of Human-Computer Interaction in the Third Millenium, Doctoral Consortium, 6-10 October 2025, Salerno, Italy

zeppieri@di.uniroma1.it (S. Zeppieri)
https://stefanozeppieri.com/ (S. Zeppieri)
0000-0001-8392-2251 (S. Zeppieri)

That foundation now informs my current exploration of how LLMs can serve as reasoning agents that support both the designer and the user during interface development.

This research combines prompt engineering, iterative prototyping, and hybrid validation strategies—where AI-generated feedback is compared against real-user data and expert evaluation. The aim is to ensure that the results are not only scalable and efficient, but also interpretable and reliable in real-world scenarios.

In the final phase of my PhD, I plan to refine these tools and study how they can be embedded into existing design workflows. The goal is to define a practical, generalizable framework for using LLMs in interface design and usability evaluation—one that maintains the core values of human-centered design while leveraging the strengths of generative models.

2. Related Works

User-Centered Design (UCD) has traditionally relied on qualitative methods like interviews, focus groups, and direct observation to gather insights into user needs and behaviors [7, 8]. These methods, while effective, are often time-consuming and hard to scale. Recent advances in Large Language Models (LLMs) offer new ways to support early-stage design tasks, including ideation, need-finding, and usability evaluation [9, 10].

LLMs such as GPT-4 have shown strong capabilities in generating content, analyzing text, and simulating human-like interactions. This has led to applications in several design-related areas: persona generation [11, 12, 1], design ideation support [1, 13, 14, 15, 16], automated usability evaluations [17, 18], and user simulation for testing interfaces [2]. Despite these capabilities, LLMs still struggle with implicit needs, cultural context, and emotional nuance, making it necessary to combine them with more traditional human-centered methods [2].

A key focus in early design is need-finding, which helps define user requirements and shape design directions. Classic approaches such as ethnographies and structured interviews [7, 8] are thorough but often difficult to apply at scale. LLMs offer an alternative by helping synthesize user needs from large datasets and generate potential problem spaces [11, 9].

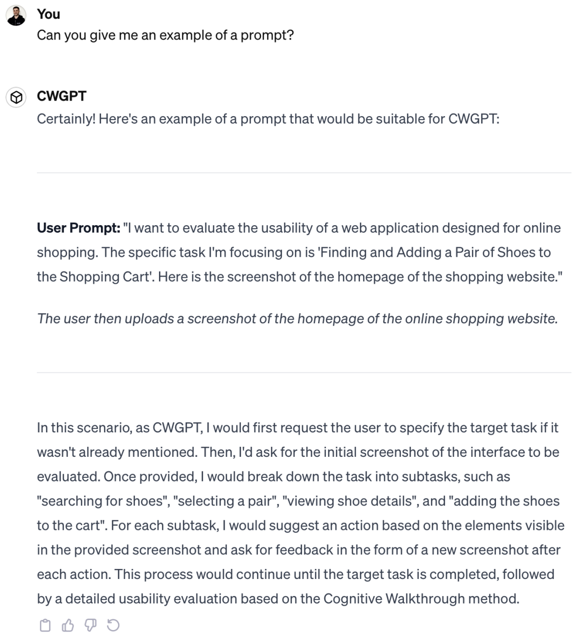
In previous work [1], we explored how ChatGPT-based tools could support need-finding across three case studies. The results showed that LLMs can effectively augment design workflows by surfacing structured insights and analyzing contextual information. However, their output should be interpreted as assistive rather than definitive, since they lack real-world experience and may overlook subtle user expectations.

LLMs have also been used to assist with UI layout generation and prototyping. For instance, they can help define target users, generate design variants, and even produce initial interface code [13, 15, 16]. These tools can be especially helpful for novice designers or in contexts where fast iteration is needed.

On the evaluation side, usability testing remains essential for validating whether interfaces meet user needs. Cognitive Walkthroughs (CWs) are one common method, where experts step through tasks from the user's perspective [19]. While useful, CWs require time and expertise. Commercial platforms now offer remote testing tools like session recordings, A/B tests, and heatmaps [20, 21], but for many designers—especially those with limited experience—the sheer number of tools and data can be overwhelming [22].

To make early-stage evaluations more accessible, our recent work introduced a ChatGPT-based tool that conducts CW-inspired assessments [2]. These evaluations help identify usability issues early in the process, without requiring real users. However, they still have clear limitations: LLMs struggle with visual feedback, interaction timing, and non-verbal cues, which suggests that hybrid approaches combining AI and human input are likely to be more effective.

Lastly, LLMs have also been used for automated GUI testing, where they interact with interfaces and produce structured feedback. While traditional tools focus on task completion or bug reports, LLM-based agents bring a more interpretive layer by explaining their behavior and reasoning [23]. This area shows promise but still requires refinement to better approximate real-world user behavior.



You

Can you give me an example of a prompt?

CWGPT

Certainly! Here's an example of a prompt that would be suitable for CWGPT:

User Prompt: "I want to evaluate the usability of a web application designed for online shopping. The specific task I'm focusing on is 'Finding and Adding a Pair of Shoes to the Shopping Cart'. Here is the screenshot of the homepage of the shopping website."

The user then uploads a screenshot of the homepage of the online shopping website.

In this scenario, as CWGPT, I would first request the user to specify the target task if it wasn't already mentioned. Then, I'd ask for the initial screenshot of the interface to be evaluated. Once provided, I would break down the task into subtasks, such as "searching for shoes", "selecting a pair", "viewing shoe details", and "adding the shoes to the cart". For each subtask, I would suggest an action based on the elements visible in the provided screenshot and ask for feedback in the form of a new screenshot after each action. This process would continue until the target task is completed, followed by a detailed usability evaluation based on the Cognitive Walkthrough method.

Figure 1: Example of a prompt with CWGPT to evaluate an interface [2]

3. Methodology

This research adopts an iterative prototyping approach, combining empirical user testing with expert evaluation to explore how Large Language Models (LLMs) can support interface design and usability assessment. The goal is to develop a practical and interpretable framework for integrating LLMs into early-stage design workflows, particularly when access to users is limited or fast iteration is required.

The first phase focused on prompt design. I created structured prompts to guide LLMs in generating realistic and diverse user personas, which were then used to simulate interviews and focus group responses. These simulations provided early insights into user expectations and design assumptions, helping surface potential problems in low-fidelity prototypes.

Beyond ideation, the same personas were involved in simulated usability walkthroughs. Here, the model reasoned through interaction flows from different user perspectives, flagging issues such as poor labeling, unclear paths, or mismatched expectations. These walkthroughs followed the principles of Cognitive Walkthroughs but adapted the process for LLM-based generation. An example of this prompt structure is shown in Figure 1, where the user specifies a task and uploads a screenshot, prompting the model to iteratively suggest actions and ask for follow-up inputs in a structured loop.

To evaluate the reliability of this approach, I adopted a hybrid validation strategy. LLM-generated insights were compared with real user data and assessed by HCI experts to understand how well they aligned with actual behavior and expectations. This helped reveal both the value and the limitations of using LLMs in early evaluation stages.

Results from this work are reported across several studies. In [2], we introduced a ChatGPT-4-based tool for conducting automated walkthroughs, showing its ability to surface subtle usability issues often missed in early heuristics. In [1], we applied LLMs to need-finding tasks across three design contexts, demonstrating their usefulness in generating structured, relevant insights to support ideation. Finally, in [5], we examined how AI-driven interfaces handle user and system errors, evaluating strategies for real-time recovery and trust maintenance.

This methodology combines the speed and scale of generative models with the grounding of human-centered design. By treating LLMs as assistive tools within a validated pipeline, the approach supports more efficient, informed, and user-aware design decisions.

4. Ongoing Work

The current phase of my research builds on previous work on need-finding and AI-assisted Cognitive Walkthroughs [6, 2], with a focus on improving how LLMs can support the early stages of interface design and evaluation. The main goal is to refine the use of LLMs in simulating realistic user behavior and generating structured feedback that designers can act on.

One area of focus is improving the quality and depth of AI-generated personas and interaction scenarios. This involves designing better prompts and incorporating contextual information from specific domains to produce more nuanced and relevant user simulations. These simulations are then used to perform LLM-based walkthroughs that help identify usability issues such as unclear instructions, navigation friction, or task breakdowns.

At the same time, I am developing hybrid evaluation workflows that combine AI-generated insights with expert and user feedback. This layered approach addresses known limitations of LLMs—for example, their difficulty in interpreting visual layouts, emotional tone, or subtle user signals—and helps ensure that evaluations are grounded in a realistic understanding of user needs.

Despite encouraging early results, some limitations remain. LLM-generated feedback, while useful, lacks the variability and depth of real user responses. These tools also struggle with analyzing complex or dynamic interfaces, especially when multimodal elements are involved. Another issue is the tendency of LLMs to rely on standard usability heuristics, which can cause them to overlook domain-specific design challenges.

To address these issues, I plan to increase the role of human validation in the evaluation pipeline and explore ways to condition LLMs to respond to multimodal input. I am also working on embedding these tools directly into existing design environments. The aim is to support iterative prototyping by allowing designers to run lightweight evaluations as they refine their wireframes and interface components.

Overall, this work moves toward building scalable and accessible evaluation methods that use LLMs not as replacements for human-centered techniques, but as augmentations. The long-term goal is to support both expert designers and less experienced practitioners in creating more usable, user-aware interactive systems.

5. Research Objectives, Contributions, and Dissertation Goals

This research contributes to the field of Human-Computer Interaction (HCI) by exploring how Large Language Models (LLMs) can support the design and evaluation of interactive systems. The main objective is to develop an LLM-assisted evaluation framework that helps designers simulate user interactions and detect usability issues early in the development process. This builds on previous work on AI-guided need-finding, virtual walkthroughs, and error recovery [2, 5, 1, 6].

The goal of this framework is not to replace human evaluation, but to offer a scalable method for generating useful feedback at the prototype stage—especially in situations where traditional methods are too slow, costly, or resource-intensive. By combining prompt-driven simulations, hybrid validation strategies, and expert review, the framework aims to deliver actionable insights that improve interface quality and usability.

So far, the work has resulted in multiple functional prototypes and peer-reviewed publications, showing the potential of LLMs in supporting both ideation and evaluation tasks. Current efforts are focused on extending empirical testing, improving the reliability of model-generated insights, and embedding these tools into design environments for more seamless use. Collaborations with external partners will help evaluate the framework in applied contexts.

The broader aim of the dissertation is to define a practical methodology for integrating LLMs into real-world design workflows—one that respects core HCI values such as transparency, user agency, and interpretability. By bridging AI-driven automation with human-centered design practices, this work hopes to contribute to a more adaptive and accessible future for interface development.

6. Conclusions

This research explores how Large Language Models can support the design and evaluation of interactive systems, with a focus on two key contributions: using LLMs to simulate user behavior for usability evaluation, and supporting early-stage design through persona generation and need-finding.

Initial findings suggest that LLM-assisted methods can improve the speed and reach of interface evaluations, especially during early prototyping. These tools help designers identify potential issues and iterate more effectively, even in the absence of direct user access.

At the same time, there are clear limitations. LLMs can miss domain-specific nuances, struggle with multimodal inputs, and produce feedback that may not hold up under real-world conditions. To address these gaps, this research adopts a hybrid approach that combines AI-generated insights with validation from experts and user studies.

The long-term goal is to develop reliable, interpretable, and accessible tools that help both experienced designers and newcomers create more usable and adaptive interactive systems—without compromising the principles of human-centered design.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4.5 and Grammarly in order to: Grammar and spelling check, paraphrase and reword. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Bisante, S. Zeppieri, D. Venkata Srikanth Varma, G. Trasciatti, E. Panizzi, Assessing Large Language Models Adoption in Need Finding: an Exploratory Study, 2024. Accepted at EISEAIT Workshop, EICS, Cagliari 2024 and soon to be published by Springer in the LNCS series.

[2] A. Bisante, V. Datla, E. Panizzi, G. Trasciatti, S. Zeppieri, Enhancing Interface Design with AI: An Exploratory Study on a ChatGPT-4-Based Tool for Cognitive Walkthrough Inspired Evaluations, in: Proceedings of Advanced Visual Interfaces 2024, AVI '24, Association for Computing Machinery, New York, NY, USA, 2024. doi:10.1145/3656650.3656676.

[3] A. Bisante, E. Panizzi, S. Zeppieri, Implicit Interaction Approach for Car-related Tasks On Smartphone Applications, in: Proceedings of the 2022 International Conference on Advanced Visual Interfaces, 2022, pp. 1–5. doi:10.1145/3531073.3531173.

[4] A. Bisante, V. S. V. Datla, S. Zeppieri, E. Panizzi, Implicit Interaction Approach for Car-related Tasks On Smartphone Applications - A Demo, in: Proceedings of the 2022 International Conference on Advanced Visual Interfaces, 2022, pp. 1–3. doi:10.1145/3531073.3534465.

[5] A. Bisante, A. Dix, E. Panizzi, S. Zeppieri, To Err is AI, in: Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter, CHItaly '23, Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3605390.3605414.

[6] A. Bisante, A. Dix, E. Panizzi, S. Zeppieri, Implicit interactions in proactive systems: Evaluation challenges and adaptations for nielsen's heuristics, 2025.

[7] D. A. Norman, The Design of Everyday Things, Basic Books, Inc., USA, 2002.

[8] J. J. Garrett, The Elements of User Experience: User-Centered Design for the Web and Beyond, 2nd ed., New Riders Publishing, USA, 2010.

[9] A. Schmidt, P. Elagroudy, F. Draxler, F. Kreuter, R. Welsch, Simulating the Human in HCD with ChatGPT: Redesigning Interaction Design with AI, Interactions 31 (2024) 24–31. URL: https://doi.org/10.1145/3637436.

[10] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL: https://arxiv.org/abs/2303.12712. arXiv:2303.12712.

[11] J. Salminen, H. Kwak, J. a. M. Santos, S.-G. Jung, J. An, B. J. Jansen, Persona perception scale: Developing and validating an instrument for human-like representations of data, in: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, CHI EA '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–6. URL: https://doi.org/10.1145/3170427.3188461. doi:10.1145/3170427.3188461.

[12] A. Salminen, B. Smith, C. Johnson, Exploring ai-generated persona profiles for usability evaluation, in: Proceedings of the 2024 ACM Conference on User Modeling, Adaptation and Personalization, 2024, pp. 123–130. doi:10.1145/1234567.8901234.

[13] V. Bilgram, F. Laarmann, Accelerating Innovation with Generative AI: AI-augmented Digital Prototyping and Innovation Methods, IEEE Engineering Management Review 51 (2023) 18–25. doi:10.1109/EMR.2023.3272799.

[14] A. S. Bale, Y. R. Vada, B. E. Oshiojum, U. K. Lakkineni, C. Rao, K. Venkatesh, I. Rani, ChatGPT in Software Development: Methods and Cross-Domain Applications, International Journal of Intelligent Systems and Applications in Engineering 11 (2023) 636–643.

[15] E. York, Evaluating ChatGPT: Generative AI in UX Design and Web Development Pedagogy, in: Proceedings of the 41st ACM International Conference on Design of Communication, 41st ACM International Conference on Design of Communication, Orlando FL USA, 2023, pp. 197–201. doi:https://doi.org/10.1145/3615335.3623035.

[16] A. Schmidt, Speeding Up the Engineering of Interactive Systems with Generative AI, in: Companion Proceedings of the 2023 ACM SIGCHI Symposium on Engineering Interactive Computing Systems, EICS 23, Swansea United Kingdom, 2023, pp. 7–8. doi:https://doi.org/10.1145/3596454.3597176.

[17] Y. Duan, et al., A figma plugin for heuristic evaluation using gpt-4, in: Proceedings of the 2024 ACM Conference on Human-Computer Interaction, 2024. Available on arXiv.org.

[18] C. Hsueh, et al., Evaluating gpt-4 as a usability evaluator: A comparative study, MDPI Journal of Human-Computer Interaction (2024). doi:10.3390/hci2024XXXXX, preliminary findings indicate GPT-4 can identify subtle UX issues.

[19] P. G. Polson, C. Lewis, J. Rieman, C. Wharton, Cognitive walkthroughs: a method for theory-based evaluation of user interfaces, International Journal of Man-Machine Studies 36 (1992) 741–773. URL: https://www.sciencedirect.com/science/article/pii/002073739290039N. doi:10.1016/0020-7373(92)90039-N.

[20] Usertesting, UserTesting human insight platform, https://www.usertesting.com/, 2024. Accessed in February 2025.

[21] Lookback, Lookback, https://www.lookback.com/, 2024. Accessed in February 2025.

[22] S. Kluivert, Usability: Where software testing tools fall short, https://medium.com/@dme_43393/usability-where-software-testing-tools-fall-short-6f56cbb9bf9f, 2021. Accessed in February 2025.

[23] Z. Liu, C. Chen, J. Wang, M. Chen, B. Wu, X. Che, D. Wang, Q. Wang, Chatting with GPT-3 for Zero-Shot Human-Like Mobile Automated GUI Testing, 2023. arXiv:2305.09434.