Towards Realistic Artificial User-Guided Usability Test Executions Using GenAI

Mathis Arends[1,2], Hendrik Winkelmann[1], Marie Griesbach[3], Norman Lahme-Hütig[1,2] and Christian Grimme[3]

[1]viadee Unternehmensberatung AG
[2]FH Münster
[3]University of Münster

Abstract

Usability testing in human–computer interaction is effective but time-consuming. Large language models allow for natural specifications of simulated user interactions to accelerate early-stage evaluation. This paper introduces `vizron`, a system that automates qualitative usability testing on web applications through persona-guided LLM agents acting as artificial users. Building on the Browser-Use framework, `vizron` combines visual, structural, and textual inputs to simulate realistic user navigation and generate think-aloud feedback. Initial findings reveal a tension between goal-oriented automation and persona realism, motivating a revised architecture that separates technical execution from persona guidance. Additionally, we argue for artificial users based on dynamic mental models rather than static personas only.

Keywords

usability testing, user experience, generative artificial intelligence, large language model, simulating human users, artificial users

1. Introduction

Usability testing is an important activity in human-computer interaction research and practice. It relies on human participants performing tasks while usability experts observe behaviors, collect verbal feedback (for instance, using think-aloud commentary from usability test participants), and analyze interaction traces to identify usability findings [1]. However, these methods are time-consuming and costly, particularly in early design phases where rapid iteration is critical. In addition, suitable human users are not always readily available, and they can only participate once for each application being tested, since their knowledge of the application changes in the course of task completion and cannot be explicitly reset.

Recent advances in large language models (LLMs) and multimodal AI have renewed this interest by enabling systems that can generate user personas, reason about interfaces, simulate user behavior, and generate qualitative feedback [2]. This paper focuses on a tool that employs LLMs to simulate and thereby test user interactions within web applications. It utilizes visual inputs and personas to evaluate how human users might interact with the web application to accomplish tasks specified in a usability test. To this end, the tool integrates six major functionalities: (1) creating personas, (2) simulating web interactions, (3) generating think-aloud comments using persona-based LLM agents, (4) inferring usability findings, (5) viewing a recording or live stream of individual runs, and (6) the ability to chat with the respective LLM agent after the execution of usability tasks. Particularly, functionalities (1), (2), and (3) are the focus of this work as they involve creating a realistic simulation of real-world usability testing.

ACM CHI'26 Workshop on Responsible Use of AI Personas in Human-Centered Design and Research, April 13, 2026, Barcelona, Spain.

Mathis.Arends@viadee.de (M. Arends); Hendrik.Winkelmann@viadee.de (H. Winkelmann); marie.griesbach@uni-muenster.de (M. Griesbach); norman.lahme-huetig@fh-muenster.de (N. Lahme-Hütig); christian.grimme@uni-muenster.de (C. Grimme)
0000-0002-7208-7411 (H. Winkelmann); 0009-0008-9319-3521 (M. Griesbach); 0000-0002-1794-7095 (N. Lahme-Hütig); 0000-0002-8608-8773 (C. Grimme)

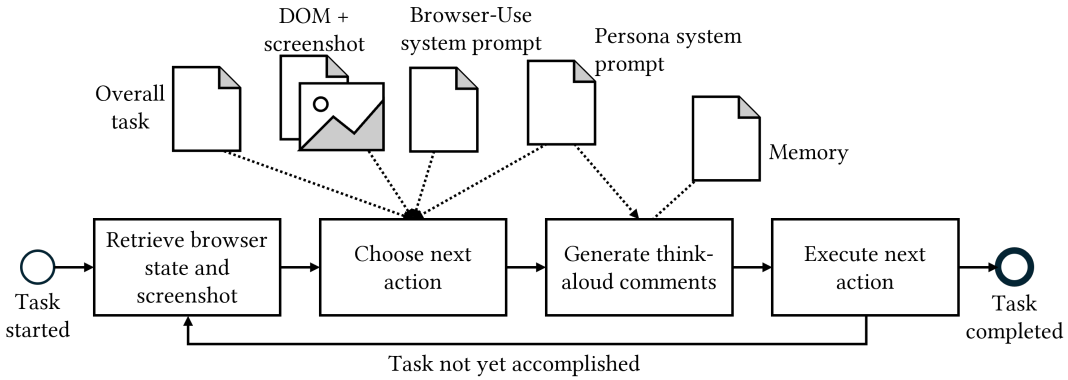

Figure 1: An abstract view on the workflow, inputs, and outputs of `vizron`.

This work is developed within the same line of research as UXAgent [3] and SimUser [4]. Both focus on launching AI agents that, given a task, navigate an application or mock-up and highlight usability concerns. UXAgent introduces an LLM-driven agent that simulates users by generating personas, navigating live websites through a browser connector, and producing, e.g., action logs, memory streams, and video recordings [3]. UXAgent demonstrates that several preparatory and execution-oriented aspects of usability testing - such as persona generation, task execution, and data capture - can be automated. In turn, SimUser combines textual descriptions of a mock-up of an application with persona-driven user agents that generate expectations, perform interactions, and provide think-aloud feedback grounded in usability standards (e.g. SUS, UEQ, NASA-TLX). Empirical results suggest a substantial overlap between SimUser-generated feedback and human user feedback, indicating that persona-based simulation can approximate certain aspects of human usability evaluation [4]. However, UXAgent and SimUser possess technical limitations: Agents primarily perceive (simplified) HTML [3] or textual descriptions of the GUI [4] rather than visual representations of the currently displayed application state, have a constrained action space [3, 4], or are not applicable to all kinds of web pages [4]. Furthermore, while SimUser does not study individual user journeys, the authors of UXAgent note that persona-informed agent executions were deemed unrealistic [3].

2. An Abstract Overview of Persona-Guided User Simulation in `vizron`

We developed a custom tool [5] to overcome several of these limitations. Its latest prototype has evolved into the system now called `vizron`. The tool is designed to carry out usability evaluations by interacting autonomously with target websites and making decisions based on visual input. When provided with a usability task and the corresponding URL, `vizron` navigates the website and employs its available features to complete the usability task's objective. To facilitate this process, `vizron` integrates Browser-Use[1], a robust framework that enables AI agents to autonomously interact with a web browser. This simplifies the cumbersome agent-browser interaction as we build upon a state-of-the-art library. Similar to UXAgent, it leverages Document Object Model (DOM) information to help the AI decide on the next interaction needed to reach a goal. The abstract workflow of `vizron` is depicted in Figure 1.

The main loop fetches the browser state and uses it, together with the DOM of the website, a screenshot of the current webpage state, and a system prompt to decide on the next action. `vizron` has been adapted to more closely resemble a human action space, e.g., by constraining the set of possible actions to those visible within the current viewport of the website and also taking an annotated screenshot of the website into account for deciding on the next action. Furthermore, the system prompt is extended by a persona. After deciding on a next action, think-aloud comments are generated and fed into a memory component, which in turn are taken into account for the subsequent generation of

[1] https://github.com/browser-use/browser-use

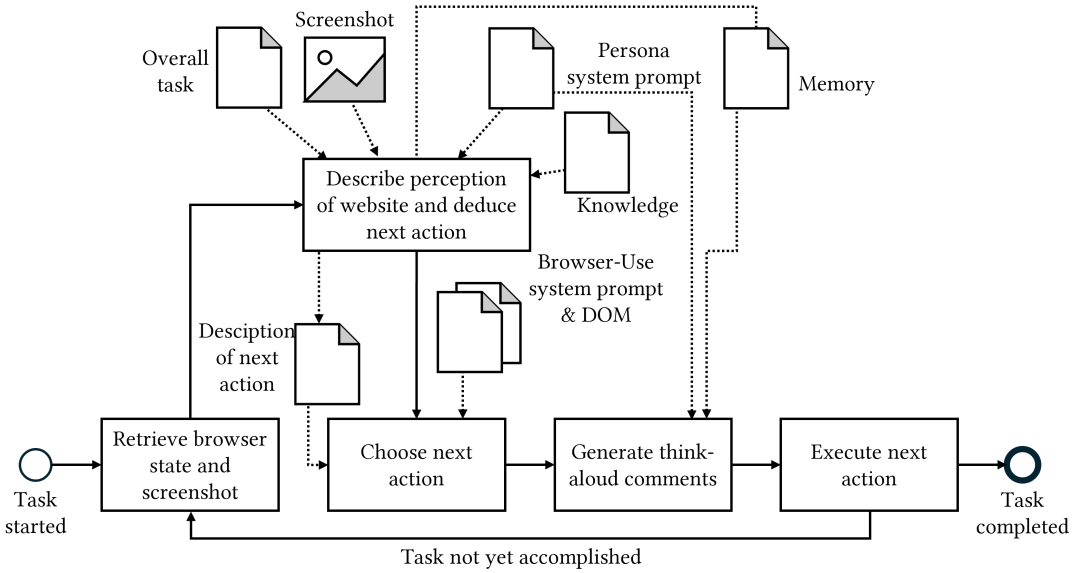

Figure 2: An abstract view on the new aspired version of `vizron`.

think-aloud comments. These steps are iterated repeatedly until the AI agent perceives that the task is completed.

An automated usability test of a web application using `vizron` is conducted with multiple LLM agents, each based on a different persona, to simulate the behavior of different types of users. Personas are models of typical or target users of the application [6]. In `vizron`, they are created by providing a textual description of the relevant factors. These factors include general data such as a description, the name, age, occupation, and the experience within the domain of the persona. Moreover, typical navigation preferences are provided, along with several examples of natural wordings from thinking aloud protocols. The collection of the relevant persona data is outside the scope of our tool, so it can be combined with classical user research or tools for automated persona generation, e. g. [7].

3. Lessons Learned & Research Directions

In agreement with the findings of past research, we found that `vizron` is capable of identifying usability issues regarding test websites [4, 3, 5]. While adopting Browser-Use increased the robustness for navigating various web pages, the effect of including personas proved to be ineffective in terms of chosen navigation paths in various exemplary runs. One reason for this are the conflicting goals of the Browser-Use system prompt and the persona system prompt extension: While the former is designed to realize interactions between the LLM and the browser as efficient and goal-oriented as possible, the latter strives to incorporate the subjective tendencies and capabilities of a human user. In the exemplary case of a persona representing a senior citizen, the think-aloud comments might reflect that there is a lot of small text and a highly complex UI, while the framework nevertheless navigates to the next best option seamlessly.

3.1. Architectural Adaptations

To counteract this, we strive to divide the technical interaction of an LLM agent and the persona-guided evaluation, as shown in Figure 2. In this new version of `vizron`, the persona-guided evaluation will be used to describe the next action to take in a comment. Thereafter, the Browser-Use library decides how to perform this action. In consequence, the persona can guide the decision and is solely responsible for finding authentic next actions for completing a task. This modification is considered to resolve the conflicting goals of the two system prompts while preserving Browser-Use's robustness.

For mimicking actual users, multiple factors must be considered. We hold that a static prompt describing a persona does not suffice. When imitating real users in usability tests, the **static persona** prompt must be complemented by dynamically enriched components (dynamic mental model). In this context, we define an *artificial user* as any attempt of mimicking a real user. In the planned new version of `vizron` personas still provide the static data to distinguish different user types. On the other hand, artificial users are the active entities solving the tasks in the course of a usability test session. In contrast to the definition of *synthetic users* [8], we do not require artificial users to be generated using AI. Furthermore, while a synthetic user is described to be *alive* in the sense that, e.g., "you can chat with it" [8], we suggest a set of components, capturing the notion of *aliveness* for the use case at hand. This is our suggested reference model for *artificial users for usability testing*.

First, a **memory component** (see, e.g. [9]) preserves and weights new experiences made during the present run. Furthermore, a human user might utilize pre-knowledge that might alter their scanning and decision behavior. This includes prior application knowledge, i.e., how the application worked in the past, or domain knowledge, i.e., general knowledge on the task. Take, for instance, an internal web application within a company. While the underlying process is not known to an LLM due to its proprietary nature, employees are well-versed in it. Conducting a usability test on such an application may require domain knowledge, for example by linking documentation to provide the necessary information to the LLM. Moreover, such domain knowledge potentially implies expectations towards an application supporting or implementing the process of work. In consequence, the addition of **mental models** [10] to the personas, encompassing aforementioned types of knowledge and experiences, seems necessary.

Finally, the artificial user's action space might be further specified (not depicted in Figure 2). In the case of web applications, a screen magnifier, a screen reader, or other accessibility tooling might be used if the situation demands it.

3.2. Evaluation Adaptations

The next challenge is how to evaluate the generated output to ensure that it truly represents the intended set of test participants. While comparing usability-test results from artificial users with those from actual users is an obvious approach for an evaluation, it does not scale. Moreover, when applied in practice, this comparison is a circular approach since manual usability tests is what we strive to postpone. Other research thus rather employs the judgment of usability experts. This, however, too can be rather expensive.

Another idea for evaluation is the well-established LLM-as-a-judge approach in which an LLM might evaluate the think-aloud comments of artificial users and argue whether these artificial users properly represent their personas. In the light of the divergence between think-aloud comments and actual exhibited behavior, we furthermore argue that, aside from this comparison of think-aloud comments, the actual behavior should be considered. To this end, Lugoloobi et al. [11] illustrate a promising research direction. In their research, they analyze different traces generated by LLMs controlling the browser. These traces then are evaluated to see whether a model can be generated that predicts which LLM is controlling the browser. In future work, this approach can be adapted to not analyze whether different LLMs controlling the browser can be identified, but rather if different artificial users representing different personas, behave meaningfully differently.

4. Conclusion

Forcing an LLM-driven artificial user to consistently adopt an assigned persona while making interaction decisions and generating thoughts remains challenging [12], yet it is crucial to achieve valuable usability testing results. One of the core challenges lies in forcing the LLM to base its interaction decisions on the assigned persona rather than on solving the task perfectly. Conversely, the second part, which involves simulating users' thoughts and behavior based on personas, remains a problematic subject. The challenge here lies not so much in the capability to generate natural-sounding output but in the ability

to obtain helpful usability testing outputs (interaction paths and thoughts) that represent different personas. This gives rise to challenges regarding the creation of personas in terms of both techniques (steering and prompting) and attributes (included factors), as well as how to evaluate the observable behavior and think-aloud comments of the artificial user to ensure it accurately matches the underlying persona.

vizron strives to use AI to improve the usability of software for human users. This is an important distinction from research where the focus rather lies on automating activities typically performed by humans using AI. To achieve this, we nevertheless draw on the possibilities offered by automating tasks using LLMs. We argue that the technology can be used to enhance existing methodology and enable the incorporation of a dynamic mental model of an artificial user in addition to its static persona to explicitly represent its mental state during a usability test session and base interactions on this mental model. Such detailed mental models would enable the creation of useful reports to improve software applications to better meet the needs of the relevant users represented by the personas.

Such automated testing allows for the inclusion of a larger and more diverse set of test participants than in smaller real-life usability testing scenarios, which presents the challenge of creating a representative set of artificial users. This includes deciding which and how many attributes to include in personas, how to aggregate them, and how to incorporate these attributes into the LLM (e.g. activation steering vs. prompting and different prompting techniques). Besides these advantages, our approach comes with the risk of assuming the complete validity of feedback if no validation is performed on human users. We therefore propose that vizron should be used for usability tests, yet, a validating final usability test should be conducted with human users.

Moreover, creating a methodology for modeling mental states of persona-based artificial users carries the risk of misuse. Spear phishing campaigns (see, e.g., [13]) could utilize similar techniques to orchestrate persuasive attacks. Finally, our approach might also be used to circumvent agent identification measures (see, e.g., [11]).

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-5 and Deepl 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] M. Hertzum, Usability testing: A practitioner's guide to evaluating the user experience, Synthesis Lectures on Human-Centered Informatics, Springer Nature, 2022.

[2] E. Kuric, P. Demcak, M. Krajcovic, J. Lang, Systematic literature review of automation and artificial intelligence in usability issue detection, arXiv preprint arXiv:2504.01415 (2025).

[3] Y. Lu, B. Yao, H. Gu, J. Huang, Z. J. Wang, Y. Li, J. Gesi, Q. He, T. J.-J. Li, D. Wang, Uxagent: An llm agent-based usability testing framework for web design, in: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706599.3719729. doi:10.1145/3706599.3719729.

[4] W. Xiang, H. Zhu, S. Lou, X. Chen, Z. Pan, Y. Jin, S. Chen, L. Sun, Simuser: Generating usability feedback by simulating various users interacting with mobile applications, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, Association for Computing Machinery, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3613904.3642481. doi:10.1145/3613904.3642481.

[5] M. Griesbach, J. Lütke Stockdiek, H. Winkelmann, C. Grimme, Towards simulating user behavior for automating usability tests by employing large language models, in: T. Bui, R. Sprague (Eds.),

Proceedings of the 59th Hawaii International Conference on System Sciences, Hawaii International Conference on System Sciences, Maui, Hawaii, 2026. In press.

[6] A. Cooper, R. Reiman, D. Cronin, About face 3: The essentials of interaction design, John Wiley Sons, 2007.

[7] A. Schuller, D. Janssen, J. Blumenröther, T. M. Probst, M. Schmidt, C. Kumar, Generating personas using LLMs and assessing their viability, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '24, Association for Computing Machinery, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3613905.3650860. doi:10.1145/3613905.3650860.

[8] M. Rosala, K. Moran, Synthetic users: If, when, and how to use AI-generated "research", https://www.nngroup.com/articles/synthetic-users/, 2024. Accessed: 2026-05-28.

[9] Z. Tan, J. Yan, I.-H. Hsu, R. Han, Z. Wang, L. Le, Y. Song, Y. Chen, H. Palangi, G. Lee, A. R. Iyer, T. Chen, H. Liu, C.-Y. Lee, T. Pfister, In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 8416–8439. URL: https://aclanthology.org/2025.acl-long.413/. doi:10.18653/v1/2025.acl-long.413.

[10] N. A. Jones, H. Ross, T. Lynam, P. Perez, A. Leitch, Mental models: an interdisciplinary synthesis of theory and methods, Ecology and society 16 (2011).

[11] W. Lugoloobi, S. Marro, J. Magomere, J. Wright, C. Russell, Known by their actions: Fingerprinting llm browser agents via ui traces, 2026. URL: https://arxiv.org/abs/2605.14786. arXiv:2605.14786.

[12] P. Bhandari, N. Fay, M. J. Wise, A. Datta, S. Meek, U. Naseem, M. Nasim, Can LLM agents maintain a persona in discourse?, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 29213–29229. URL: https://aclanthology.org/2025.emnlp-main.1487/. doi:10.18653/v1/2025.emnlp-main.1487.

[13] F. Heiding, S. Lermen, A. Kao, B. Schneier, A. Vishwanath, Evaluating large language models' capability to launch fully automated spear phishing campaigns: validated on human subjects, arXiv preprint arXiv:2412.00586 (2024).