That's Enough about AI Replacing Users in User Research

Ian Arawjo
Department of Computer Science and Operations Research (DIRO)
Université de Montréal
Montreal, Quebec, Canada
ian.arawjo@umontreal.ca

Abstract

The title of the present workshop, to develop standards for LLMs to simulate participants in user research, implies that such usage is inevitable, even widespread—that the question is no longer whether, but when, how, and so on. I take issue with this premise, especially as it regards scientific research in HCI. I argue that the jury is still out on LLM simulations, and that LLMs have very little, if any, place in replacing users in UX research, especially in an academic context. Although LLMs can help designers identify *potential* user behaviors and usability defects—serving an analogous role to previous, "discount" analytic methods but at even cheaper cost—LLMs are not "research participants," and cannot ever serve as empirical evidence of user behavior. I contextualize LLMs as merely the latest attempt to shortcut quality UX research, argue that one must conduct human subjects research anyway whenever one uses LLMs to simulate users (and thus LLMs are not worth the trouble), and finally, that while LLMs can support the process and preparation of user research, replacing humans in human subject studies with simulations (and then presenting those simulations as "proof" of user behavior or meeting user needs) is where we as a community must draw the line.

CCS Concepts

• **Human-centered computing** → **User studies**; *Interaction design theory, concepts and paradigms*; Usability testing; • **Computing methodologies** → Natural language processing.

Keywords

user research, large language models, evaluation methods, simulated users, UX research methodology, empirical studies

ACM Reference Format:
Ian Arawjo. 2026. That's Enough about AI Replacing Users in User Research. In *Proceedings of The Workshop on Developing Standards and Documentation For LLM Use as Simulated Research Participants at the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26 Workshop Paper).* ACM, New York, NY, USA, 4 pages.

1 Introduction

The present workshop aims to consider "developing standards and documentation for LLM use as simulated research participants." As an HCI researcher with years of experience conducting user research, I understand the temptation to use LLMs to simulate users. Quality user research is hard, time-consuming, and costly. Yet I am less optimistic that language models are the solution, and I am worried about framings which imply that LLM-based user simulation is a foregone conclusion for UX research—an existing practice to "standardize" and "document," rather than critique or forbid. Rather than lying down on the tracks and letting the AI hype train run us over, I believe the jury is still out on what place, if any, LLMs have in user research, especially as it concerns replacing human subjects studies in research publications at our HCI venues. Before we ask "in what format is this practice acceptable," we should consider whether it is ever acceptable at all.

In this polemic, I argue that the risks and complications of LLMs as simulated research participants so far outweigh the potential benefits that I don't think that HCI researchers should use language models to replace real human beings in user research at all—that is, to treat simulations as "proof" of user behavior or addressing real user needs. I make three major points (which, it should be said, are far from the only points one could make [1, 7]):

(1) People and companies have always looked for ways to shortcut user research, and LLMs are not special in this regard
(2) It is impossible for peer reviewers to validate 'findings' from synthetic user research without the authors resorting to conducting some real user research anyway
(3) At best, LLMs can support the process of *preparing* to conduct user research, and can serve a similar role to discount analytic methods of the past, but where we should draw the line is using them to replace human subjects

The implication of this argument is that *under no conditions* should a system paper be published at an HCI venue with the sole "proof" of evaluation being a simulation of user behavior. While LLM simulations can have a role in estimation, similar to computational models of the past like game theoretic analyses, these models are not "user research," and no one should claim so. In what follows, I lay out these points in detail, in the hopes that those who wish to proceed with conducting (and evaluating) "synthetic" user research consider the complications with this dream and the broader history of attempts at shortcutting or automating user research.

2 Why is this happening? The social and economic context of repeated dreams to shortcut user research

Predicting likely user behavior with LLMs is merely the latest attempt to expedite and automate user research [1]. Running user studies—collecting data, interacting with messy, real people—all of this takes time, money, and training. Human-centered design researchers across academia and industry have long sought to reduce the costs involved: whether crowd-sourcing on platforms like MTurk, or running 'design sprints' where research is quick-and-dirty [9, 10]. UX researchers have always had a precarious position

CHI '26 Workshop Paper, Barcelona, Spain
2026.

in industry, with managers and researchers alike looking for shortcuts to the time and money required to run quality user studies [4, 9]. While the current tendency in academia and industry is to treat LLMs as uniquely special disrupters, I see LLMs as merely a natural continuation of this longstanding search by industry and academia to find ever-cheaper ways to conduct, and especially to circumvent, user research.

Consider the early history of usability research. Heuristic evaluations were introduced by Nielsen around 1989 under the term "discount" usability [12]. Nielsen framed this new method explicitly as a way to cut company costs for potentially large gains, making a series of business arguments that heuristic evaluation cuts down the cost of full-blown user research from $128,330 to $10,500, ultimately producing a user-centered design that saves the company $500,000 at only $10,500 expense [13].

Yet even in these papers on "discount" usability, Nielsen does not claim the human can be foregone entirely—the goal is rather to provide ever-more structured, shortened, and/or informalized procedures, to around 5 users, to gain quick feedback. Even in heuristic evaluations, the human participants need to be "usability experts"—trained in the principles of a formalized coding scheme. The goal of heuristic evaluations was from the start to reduce the costs of usability testing at the price of accuracy and alignment with real human behavior. While the rare academic paper might report a heuristic evaluation, no one would claim analytic methods serve to substitute for research participants and especially not as empirical evidence to ground scientific claims. When they are reported, analytic methods are treated more informally, matching the informality and tenuous nature of their conclusions.

While Nielsen's work was positioned in an industry context, academics are also continually enamored by shortcuts to quality user research. In computing, there is a growing practice of "performative" user studies that "result from reviewers' expectation that 'papers should have some evaluation,' not from careful thought about the value and usefulness of the studies themselves" [5]. Performative user studies involve authors tacking on a user study to satisfy reviewers and get papers published, regardless of the quality of the study or its goals—e.g., giving a NASA TLX survey to participants "solely because reviewers expect to see a table of numbers" [5]. Despite calls for appreciating other methods of evaluation [3, 14], the predominant standard of systems evaluation at conferences like UIST and CHI remains a controlled usability study, an obligation that can cause authors to fall prone to similar performativity. Qualitative research also faces continued, even increased, skepticism from reviewers who hold positivist perspectives on what constitutes an acceptable contribution (focusing on quantification, statistical generalizability, and so on) [2, 11, 17]—further pushing authors into performing quantification to justify value.

LLMs provide an even easier outlet for people to "perform" user research. The positivist allure is strong: if one needs more numbers to perform quantitative evidence, one can now conjure them up with the click of a button. As to what these numbers mean or how they should be interpreted, who cares if it means one gets their publication accepted or if one gets the UI design past their manager? The only difference is that LLMs are even cheaper and faster than crowdsourcing, which was the previous shortcut *du jour*.

As an anecdote attesting to this, I attended a recent talk where a researcher showed tables of numbers 'proving,' with statistics and p-values, significant results for their method compared to another method. The audience was glued to their seats—until the oracle was revealed to be an LLM judge whose prompt was not even reported, at which point the talk devolved into a flurry of questions from the audience. (The author then countered that they are looking into recruiting human experts to do the annotation instead.) Prompting an LLM for numbers and then running statistics on them is farce, not science. The question is not how to report this kind of methodology—how to document it, standardize it, and so on—but why otherwise-intelligent people feel compelled to default to LLM simulations and to over-trust the AI outputs produced to the extent of running statistical tests on them as ground truths.

Thinking more positively, what should be done? From my perspective, *the question is how to improve standards for quality user research,* not how to circumvent it. Instead of asking how we might standardize a farce, perhaps we in HCI should rather ask **how can we uphold better quality user research?** How can we better conduct user research at scale when we face "society-scale" AI systems like social media, ChatGPT, etc.? Crowdsourcing platforms, which were never perfect arbiters of truth and always subject to gaming the system and unethical practices, now face additional scrutiny as participants can now easily use AI features to simulate responses. As HCI doesn't have additional methods for conducting higher-quality, at-scale user research, this now entices people into resorting to simulations.

Along these lines, for academia, we might consider the narrower question: **how can we prevent the proliferation of low-quality user research data in publications**? Some studies may deserve more thoughtful re-evaluation, while others may not require user evaluation at all if their primary contribution is an innovative system. This is a question that organizers of venues like UIST are returning to—how to foster a culture of peer review that alleviates the pressure to conduct performative studies [2, 3]. And, as "vibe coding" continues, and systems become easier to build, our burden of execution should shift to the *burden of evaluation*: showing *hard proof* that our systems and methods actually work for real users, because that's where the real value lies. Not in shortcutting the evaluation burden, too—that just churns out more slop.

3 LLMs as simulated users require conducting user research with human subjects anyway, and hence aren't worth the trouble

Having contextualized why people might feel pressured into using LLMs to replace users in user research, I now turn to considering the situation from the perspective of the method itself.

The central counter-point to using LLMs to simulate users is this: **it is impossible for peer reviewers to validate 'findings' from synthetic user research, at least without resorting to conducting some real user research anyway.** LLMs are not and never will be 'grounded' empirical evidence of user behavior, and thus, LLM outputs cannot serve as a solid foundation for decision-making and theory-building that depends upon empirical evidence, such as forming design goals or reaching conclusions about user behavior and biases.

For example, I have encountered authors using LLMs as "judges" in recent paper submission cycles. In these cases, if reviewers don't reject the paper outright, they end up asking for details on how aligned these judges' are with human graders (asking, "who validates the validators?" [16]). This then caused authors to do the work to recruit human domain experts to annotate data, and then report how aligned they are with LLM grading using previous techniques like inter-rater reliability scores. All statistics based on the LLM judges, however, remain suspect—for if the inter-rater reliability score is "60-80% aligned with human judgment," how should that factor into statistical calculations?

The problem goes beyond statistical hygiene, however. If authors will be asked to conduct user research anyway (because LLM outputs alone cannot serve as empirical evidence of user behavior), *why not just run human subject studies to begin with?* What was the added value of the LLM simulation? The situation shifts from "how to use LLMs appropriately" to "are the LLM-simulated users even worth the trouble, if they are only ever additional to conducting user research with humans anyway?" Probably not.[1]

Another concern, however, is procedural, and arguably more thorny. In user studies, researchers always have some leeway in recruitment and sampling, procedures which can bias results and rig the dice in favor of results that confirm pre-defined theories. However, though setups can be rigged to bias users in studies, outside of blatant false reporting, reviewers can get a sense of how fair the study methodology was to address the research questions. LLM for user simulations to do not offer the same transparency. LLM simulations are subject to "prompt hacking": if you don't like outcome, just grab a different model or perturb prompts and rinse and repeat until 'results' reach confirmatory significance [6]. A believer might counter this argument with a many-worlds method: to require the authors to report their experiments across many different prompt perturbations and models. Yet, this increases the complexity of analysis anyway, offsetting the presumed gains of speed and low cost; what is more, each model is not of the same quality, offering unknown alignment with human responses and likely biases [7, 15]. Again we are back to square one: are LLMs as simulated users worth the trouble for academic research? The answer, everywhere we turn, seems to be *no*.

4 Where do LLMs fit into user research, if at all?

At this point, a reader might wonder: doesn't this author find any value for using LLMs in user research?

I believe LLMs can be useful as brainstorming tools and to prepare material as part of the practical process of *preparation* for user research: e.g., refine interview transcripts, identify research goals, strengthen experiment design through feedback on study protocols. Such usage might be reported (although this itself is a bit strange—do I always report doing web searches or asking the advice of psychology colleagues over coffee?). Indeed, in the longer forms of this workshop's call for papers, this type of preparatory support is also up for discussion.

Where we must draw the line is LLMs as *simulated research participants* (a phrase which contains an oxymoron, for a simulation cannot "participate"). At best, an LLM can speculate on a study design's potential drawbacks or outcomes; such results are always informal and cursory, much like a human collaborator might speculate on how participants may behave, run an informal pilot study, or conduct a heuristic evaluation prior to a user study. But in the same way that this human speculation cannot serve as empirical evidence to ground later decisions and scientific theories, so too it goes for LLMs. As we would not dream of claiming informal chats with the lab-down-the-hall as solid evidence for scientific theory, so too must we treat LLM simulations: at best as speculative tools that can raise questions and help us improve the process of UX research, not as a replacement for human subjects themselves.

With that perspective in mind, then, I return to the title of the workshop—to develop standards and documentation for LLMs as simulated research participants. The standard, as I see it, should be thus—*LLMs are not acceptable replacements for human subjects studies.* At best, they can be used in early-stage design and to support the preparation of user studies, but never serve as substitutes for interacting with real human beings. Where documentation is requested, it would be in the more informal ways HCI scholars typically report similarly informal methodologies in the early stages of research—e.g., in systems research, with a brief paragraph acknowledging how early, informal feedback impacted their design. In this regard, LLMs are more like web search engines, rules of thumb, and chats with colleagues—always used in the process of research, rarely rigorously reported—and thus, not somehow unique technologies that deserve special treatment. At best, authors might be required to reveal how they used LLMs as part of their preparation for user research and early-stage design iteration—but a simple paragraph, rather than a rigorous standard, would suffice.

5 Conclusion

It is not lost on this author that this is an unpopular opinion in this day and age. Yet, I wanted to raise it in the spirit of academic debate. It would be a very sad day if, at CHI 2030 (or 2040), 95% of systems are "validated" through simulated user studies—this merely inflates the replicability crisis, further entrenches academia's deterioration into a game of quantity over quality, and further divorces scientists from the people who, need I remind them, they serve. It strikes me as convenient that synthetic user research arrives at the very time that, everyday on social media, evangelists of the AI hype machine announce the end-to-end automation of all research, lauding the benefits of publishing hundreds or thousands of papers a month. The rise of dreams to "replace" users with simulations in user research is not a coincidence, but merely another part of that same machine.

After reflecting on this for a while, however, I believe that the *central* issue lies in people's tendency to anthropomorphize [8] LLMs—eager to call them "synthetic users" and "research participants," rather than framing these tools as another analytic or discount

[1] At best, one might resort to LLM judges in a situation where the corpus (of tasks, outputs, etc.) to analyze is extremely large, but they must report alignment with humans anyway in order to ground their claims (say, with a random sample of said corpus). Thus, the question is what scientific value the LLM judges provide, for the cost and trouble to run the analysis. One could counter that perhaps the goal of the research is to produce an LLM judge that reasonably aligns with humans, yet this argument too falls apart—what humans, where, and when? The validity of the instrument falls apart unless it is rigorously and continuously tested and re-tested against human subjects as societies and cultural contexts change. That would require millions of dollars of upkeep, a large company, and rigorous knowledge of psychometric statistics, which most UX researchers don't have and can't assume.

method along the lines of heuristic evaluations. The workshop organizers have, unfortunately, also fallen prey to this tendency. LLMs *can*, certainly, help us prepare for user research, foresee problems with our design, or roughly estimate how emergent social behavior might change under different policy decisions, but the problem is then loosely equating this analytic quality with a real user. LLM simulations cannot and should not replace the necessary step of conducting real user research on real human beings. Those who champion LLMs in user research need to get past this anthropomorphic tendency and develop a more tactful framing.

To conclude, regardless of whether I like it or not, people remain excited by the prospect of using LLM simulations to replace human subjects in user research. To that end, the present workshop serves as a critical space for much-needed analysis and debate. Yet, I found the title of this workshop odd, as it risks framing LLMs-as-simulated-research participants as a foregone conclusion, rather than a matter yet to be decided. To combat this perception, I took a strong counterposition to remind HCI researchers that they still have the power to decide whether LLMs as "simulated participants" have any place at all in user research, especially in our conferences and journals. While industry might proceed with discount approximations, the role of academic research is to uphold high standards of quality and, quite often, to serve and reflect the needs and values of real people. I hope this piece sparks productive debate and look forward to future conversations with colleagues around this topic.

References

[1] William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–12.

[2] Andy Crabtree. 2025. H is for human and how (not) to evaluate qualitative research in HCI. *Human–Computer Interaction* (2025), 1–24.

[3] Saul Greenberg and Bill Buxton. 2008. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 111–120.

[4] Rex Hartson and Pardha S Pyla. 2012. *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier.

[5] Aaron Hertzmann. 2023. The Curse of Performative User Studies. *IEEE Computer Graphics and Applications* 43, 6 (2023), 112–116.

[6] Thomas Kosch and Sebastian Feger. 2025. Prompt-Hacking: The New p-Hacking? *arXiv preprint arXiv:2504.14571* (2025).

[7] Eduard Kuric, Peter Demcak, and Matus Krajcovic. 2026. Synthetic Participants Generated by Large Language Models: A Systematic Literature Review. *Research Square* (10 03 2026). doi:10.21203/rs.3.rs-9057643/v1 Preprint.

[8] Takuya Maeda and Anabel Quan-Haase. 2024. When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1068–1077.

[9] Jennifer McGinn and Christopher LaRoche. 2014. Fast, cheap, and powerful user research. *Interactions* 21, 3 (2014), 62–65.

[10] Brian McInnis and Gilly Leshed. 2016. Running user studies with crowd workers. *Interactions* 23, 5 (2016), 50–53.

[11] Karla Felix Navarro, Eugene Syriani, and Ian Arawjo. 2026. Reporting and Reviewing LLM-Integrated Systems in HCI: Challenges and Considerations. *arXiv preprint arXiv:2602.05128* (2026). Conditionally accepted to CHI 2026.

[12] Jakob Nielsen. 1989. Usability engineering at a discount. In *Proceedings of the third international conference on human-computer interaction on Designing and using human-computer interfaces and knowledge based systems (2nd ed.)*. 394–401.

[13] Jakob Nielsen. 1994. Guerrilla HCI: Using discount usability engineering to penetrate the intimidation barrier. *Cost-justifying usability* (1994), 245–272.

[14] Dan R Olsen Jr. 2007. Evaluating user interface systems research. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*. 251–258.

[15] Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh, and Seraphina Goldfarb-Tarrant. 2026. Lost in Simulation: LLM-Simulated Users are Unreliable Proxies for Human Users in Agentic Evaluations. *arXiv preprint arXiv:2601.17087* (2026).

[16] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[17] Robert Soden, Austin Toombs, and Michaelanne Thomas. 2024. Evaluating interpretive research in HCI. *Interactions* 31, 1 (2024), 38–42.