Risk or Chance?

Large Language Models and Reproducibility in HCI Research

Thomas Kosch, Humboldt University of Berlin
Sebastian Feger, TH Rosenheim

Insights

- → The rapid adoption of LLMs signifies their potential to become commonplace tools for data analysis and substitutes for human study participants in HCI research.
- → Anticipated reproducibility challenges surrounding LLM adoption in HCI, including bias, data-analysis support, documentation requirements, and publication pressure, highlight the need for proactive discussions and the development of best practices.

Large language models (LLMs) affect and transform most areas of daily life, from education and gaming to creativity and work. Exemplary LLMs are Llama, Alpaca, and GPT-4, with the latter being made accessible to the general public through OpenAI's ChatGPT. Consequently, ChatGPT reached 1 million users within five days after its release and currently has more than 180 million users (https://explodingtopics.com/blog/chatgpt-users). In light of the fast-paced developments over the past few years and today's adoption of LLMs across large parts of society, these models have also attracted a great deal of research interest. With the rapid development in natural language processing and the great accessibility of LLMs to the public, they are anticipated to become a commonplace tool for data analysis in human-computer interaction research. For example, LLMs have been explored to accelerate or support the analysis of textual data in HCI [1], support the user-centered design (UCD) process [2], and simulate human samples or replicate user studies [3], with the latter practices being scrutinized by recent research [4].

As with most significant scientific developments, today's high research transparency and validity standards demand a systematic understanding of how using LLMs affects reproducibility. Reproducibility is a

IMAGE BY KROT_STUDIO / SHUTTERSTOCK.COM

major concern across scientific fields. HCI, in particular, is subject to diverse reproducibility challenges due to the wide range of research methodologies employed.

The broader scientific community has initiated complex discussions on the potential impact of machine learning and artificial intelligence on reproducibility [5,6]. We aim to contribute to this discourse by focusing specifically on the research and adoption practices of LLMs in the HCI community. This aligns with previous work that advocates for the unique role and responsibility of HCI and human subject research in promoting reproducible practices. Our goal is to design processes and tools that foster scientific reproducibility, thereby enhancing the credibility and validity of HCI research.

In this article, we explore how the increasing adoption of LLMs across all user experience (UX) design and research activities affects reproducibility in HCI. In particular, we review upcoming reproducibility challenges through the lens of analogies—from past to future (mis) practices, including p-hacking and prompt hacking, general bias, support in data analysis, documentation, education requirements, and potential accelerated publication pressure on the community. We discuss the risks and chances for each of these lenses with the expectation that a broader discussion will help shape best practices and contribute to valid and reproducible practices around using LLMs in HCI research.

LARGE LANGUAGE MODELS AS A RESEARCH TOOL

LLMs are built using neural networks, which are computational models inspired by the structure and function of the human brain. A key component of these models is the transformer architecture. LLMs are trained on vast amounts of text data, including books, articles, websites, and other written sources. This data teaches the model about language patterns, syntax, semantics, and other linguistic features. The text data is tokenized, breaking it down into smaller units such as words, subwords, or characters, which are then represented numerically to capture semantic relationships between tokens. During the pretraining phase, the transformer analyzes the tokens. It adjusts its internal parameters to minimize a loss function, which measures the difference between the model's predictions and the actual outcomes. After pretraining, the model can be fine-tuned on specific tasks or domains to improve performance. Once trained and tuned, the model can be used for various tasks such as text generation, classification, or language translation.

Yet the inherent architecture of LLMs challenges the reproducibility of their outputs. Most LLMs, including GPT-3.5 and GPT-4, are autoregressive models, generating each subsequent token based on preceding tokens within the same sequence. Thus, LLMs cannot holistically self-adjust or validate their outputs. Due to this architectural limitation, LLMs show reduced reasoning and cannot utilize human resources during generation to improve the precision or validation of the output. Although users can apply prompt engineering techniques to elicit more-thoughtful responses by either integrating specific phrasings within a prompt or conversing over multiple messages with LLMs, these techniques increase the likeliness of unequal outputs when repeating the process. Here we explain the implications of these limitations.

Value lock-in. These characteristics complicate the reproducibility of outputs. LLMs are prone to so-called value lock-ins, meaning that LLMs construct their understanding of

HCI, in particular, is subject to diverse reproducibility challenges due to the wide range of research methodologies employed.

human behavior and decision making by analyzing the norms and attitudes found in human-written texts, such as those found on the Internet and in books. LLMs usually undergo training only once, however, missing changes in user standing and opinions that can happen in the future. As a result, the responses they produce may continue to reflect attitudes and beliefs from the time of their initial training. The implications generate research results that appear meaningful, although they may not reflect the facts and beliefs of participants (for example, text-based analyses that comprise interviews or think-aloud data are prone to this).

Paradoxically, updates to the used LLM may change the research results, making reproducibility more challenging. While specific parameters, such as the temperature function as magnitude for output "randomness," can be recorded, the exchange of whole models may unpredictably affect the research results. Providing access to previous LLMs, for example, through a repository outlining the older LLM version, used parameters, prompts, and outputs, may circumvent some of the described issues.

Training bias. LLMs are limited to knowledge that is represented in their dataset. For example, various LLMs were trained on data available on the Internet and then fine-tuned using reinforcement learning with human feedback. Consequently, LLMs are implicitly designed by a fraction of users who share similar properties: people with access to participate, design, and publish on the Internet. LLMs may reflect views from Western, educated, industrialized, affluent, and democratic backgrounds, reinforcing cultural biases and generating stereotypical representations of marginalized populations when used to analyze research data. Human reviewers can act as secondary observers to investigate the outputs for biases. However, this approach is prone to confirmation biases, where humans seek out, understand, prefer, and remember information that aligns with their existing beliefs or values. Depending on a reviewer's background, this may amplify the bias in LLMs through further confirmation.

Hallucination. Hallucinations have challenged LLMs since their adoption by the public. LLM-based text generation is susceptible to producing unintended content, leading to a decline in the quality of the results. If the results are not factually cross-evaluated with human experimenters, hallucinated results may not be noticed and may be published. This weighs heavily when using LLMs to support or simulate human participation processes [4], such as when a lawyer tasked a language model with serving as a legal assistant, resulting in a court filing filled with fictitious legal references. This occurred partly due to the lawyer's trust in the authenticity of the referenced cases, which could have been revealed through factual testing. How would qualitative and quantitative analysis results be tested in the context of HCI research? The HCI research community must validate methods that test for hallucinations in HCI research results.

IMPLICATIONS FOR REPRODUCIBILITY IN HCI RESEARCH

HCI research is characterized by a methodological diversity in designing and evaluating systems that pose various reproducibility challenges already today [7,8]. The expected widespread adoption of LLMs as part of ideation support [9], the substitution of human participants for design requirement mapping and system evaluations [4,2], and support for data analysis [1] open up an entirely new spectrum of reproducibility challenges in HCI research. Following are risks and chances for HCI reproducibility that we have mapped and our initial recommendations across these different phases of UX research.

Learning from today's reproducibility challenges. Numerous factors contribute to the reproducibility challenges we face today. In quantitative research, one major issue relates to p-hacking. P-hacking refers to the selective reporting of statistical tests to achieve statistically significant results, leading to inflated false positives and compromised reproducibility. This practice not only undermines the integrity of scientific research but also perpetuates a cycle of erroneous findings.

While moving toward the increasing adoption of LLMs in HCI research, we must consider how to address existing reproducibility challenges and carefully navigate new pitfalls that might arise from LLMs. Analogous to p-hacking, using LLMs during UX research is vulnerable to the so-called prompt hacking. Prompt hacking of LLMs mirrors p-hacking in research by manipulating inputs to influence outputs. As p-hacking selectively reports statistical tests to achieve desired results, prompt hacking skews LLM responses by adjusting input prompts. Both practices compromise integrity: p-hacking distorts scientific findings, while prompt hacking biases language model outputs. Recognizing these parallels highlights the importance of transparency and integrity in research and AI development, urging us to prioritize robust methodologies to uphold credibility and reliability.

We propose the following to avoid repeating mistakes:

- The adoption of LLMs as part of UX research must consider reproducibility challenges and specifically identify analogous issues introduced by LLMs.
- Current best practices for using LLMs should be adopted. Regarding the examples of p-hacking and prompt hacking, the applicability of established tools like preregistration and transparent, prompt protocols as part of manuscript submissions and paper publications should be evaluated.

Bias across user experience research. UX research and system evaluation fundamentally require knowledge about human perception and experience. In this regard, HCI research faces the issue of bias by sampling too few experiences or through a biased set of human samples. In this context, Albrecht Schmidt et al. [2] stress that the "basic idea is that LLMs encode human experiences, which may be drawn upon in design." Expanding on this, we perceive a substantial opportunity for LLMs to increase information's robustness and consequent reproducibility throughout the UX research process, from requirements mapping to system evaluation.

This approach, however, also represents reproducibility risks. Today, HCI research is confronted with the criticism that findings often reflect the perspectives of young, educated, and

Western communities. LLMs can further aggravate this problem if they are sampled on reports and experiences from a distinct part of society. Many experiences will reflect technology-savvy individuals with a tendency for a younger population. LLMs are also likely to favor specific languages and cultures. Any transparency regarding these training biases makes further assessment of risks difficult.

Concerning bias and its impact on research reproducibility, we map multiple requirements:

- HCI should support the development of LLMs to make the selection of training data and subsequent biases transparent.
- The HCI community should use multiple LLMs across UX research to reflect broad human perspectives wherever necessary (for example, combining LLMs that lean toward different regions or cultures).
- Research and reviewers must critically examine the interconnections between LLM training transparency, combinations of different LLMs, interplay with additional direct human subject reporting, and the resulting deviation for the specific activity and use case.

LLMs for cross-validation and analysis support. HCI is subject to a rich diversity of research methods unmatched in most other fields. While we appreciate this richness, it comes with various reproducibility issues subject to specific methods, which generally affect HCI research reproducibility. Many HCI researchers are involved in activities across multiple methodologies. As it is difficult to gain the same expertise across all methods, individual researchers might find it more challenging to evaluate and counteract reproducibility issues for some of the methods they employ.

We see an opportunity for LLMs to help educate researchers about reproducibility pitfalls across methods and support the validation of research findings. For example, regarding qualitative methods, which have been subject to claims of reproducibility, LLMs might provide additional verification complementing manual data analysis. This also holds an opportunity to counter the bias of a single or few interpreters analyzing qualitative research data, as Wilbert Tabone and Joost de Winter [1] demonstrated. Similarly, LLMs can support quantitative data analyses by cross-checking and reasoning about the applicability of statistical tests and the validation of calculations through a second entity. At the same time, we note the risk of overreliance on LLMs across HCI research methods.

LLMs can help improve research reproducibility as an assistive tool across various HCI research methods. Besides providing cross-validation, LLMs can help fill individual gaps in best practices. LLMs should carefully be used as a supportive tool, however, rather than a single source of analysis, bearing the risk of overreliance, bias, and subsequent irreproducibility.

Defining new reporting requirements and educating the community. More documentation of data and metadata must be required to represent a key issue for research reproducibility. Various initiatives across research fields attempt to encourage or require scientists to provide the most accurate documentation of recording conditions, hardware setups, and software specifications, among others. The introduction of LLMs into research practices requires developing new best practices regarding the reporting of LLM usage. Diverse fields like HCI must specifically investigate and pose requirements that apply across research methods.

We suggest the following:

- Establish precise documentation

The structure, training methodologies, and frequent updates of LLMs are not designed to yield consistent results, which can affect reproducibility.

requirements as part of publication venues that demand detailed information regarding the scope of LLM use, prompts entered, and corresponding metadata like concrete LLMs used and their specific versions.

- Provide accessible educational resources for the wider HCI research community that explain those reporting requirements and general LLM use challenges, such as bias, hallucination, and value lock-in.
- Contribute to the development and incentivize the use of transparent and accessible LLMs that provide detailed information on the type of data used for training and remain accessible to the community for reproduction. This targets the primary reproducibility concern that older or specific versions of commercial and proprietary LLMs become unavailable to the public.

The risk of increased research pressure on HCI reproducibility. Researchers face the challenge of producing high-quality research while ideally generating sufficient output to advance their careers. The use of LLMs provides an opportunity to increase efficiency. For example, Orit Shaer et al. [9] demonstrated that LLMs can support creative ideation processes. Along those lines, Schmidt et al. [2] envisioned the partial substitution of human participants through LLMs, and Tabone and de Winter [1] discussed using LLMs in data analysis and reporting. These opportunities might create pressure, as Schmidt et al. [2] stated: "By using LLMs, we might make UCD cheaper and hence more widely applicable; at the same time, though, we put pressure on the field to move this way to stay competitive. Hence, the transparency about how UCD is conducted and to what extent models are used is critical."

Considering the concern of research reproducibility, we must also address the potential risks associated with the pressure to adopt LLMs. The premature introduction of LLMs into general HCI research could result in bad scientific practices before best practices across the diverse HCI methodologies can be established and communicated. Moreover, the anticipated increase in paper submissions could strain an already busy peer review system, potentially affecting the time reviewers can dedicate to assessing the reproducibility implications of LLM use.

We suggest the following:

- Manage and communicate expectations and address concerns regarding LLM use and potentially perceived pressure within and across HCI laboratories.
- Develop, communicate, and demand best practices as quickly as possible as part of paper submission requirements. This might happen through dedicated workshops and panels at HCI publication venues.
- Educate peer reviewers about best practices and evaluate whether and how LLMs might be used as a support tool in increasingly complex and rich peer review processes, considering potential pitfalls like bias, hallucination, and value lock-in.

RISK OR CHANCE FOR REPRODUCIBILITY IN HCI?

LLMs present a significant opportunity for HCI research to accelerate data analysis and disseminate results. While acknowledging the benefits of employing LLMs to make HCI data analysis procedures more accessible to the community, it is essential to approach their increasing use cautiously. The structure, training methodologies, and frequent updates of LLMs are not designed to yield consistent results, which can affect reproducibility, manifest biases, and increase pressure on publication processes. In this article, we suggest a discourse on the influence of LLMs within the HCI community and their impact on research reproducibility. We aim to generate interest in a series of focused discussions that we plan to organize soon. These conversations will explore the establishment of specialized scientific platforms for disseminating research conducted in this domain, with the goal of maintaining reproducibility when employing AI tools for HCI data analysis.

Endnotes

1. Tabone, W. and De Winter, J. Using ChatGPT for human-computer interaction research: A primer. *Royal Society Open Science 10*, 9 (2023).
2. Schmidt, A., Elagroudy, P., Draxler, F., Kreuter, F., and Welsch, R. Simulating the human in HCD with ChatGPT: Redesigning interaction design with AI. *Interactions 31*, 1 (2024), 24–31.
3. Aher, G., Arriaga, R.I., and Kalai, A.T. Using large language models to simulate multiple humans and replicate human subject studies. *Proc. of the 40th International Conference on Machine Learning*. ACM, New York, 2023.
4. Agnew, W. et al. The illusion of artificial inclusion. *Proc. of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, 2024, Article 286.
5. Ball, P. Is AI leading to a reproducibility crisis in science? *Nature 624*, 7990 (2023), 22–25.
6. Gibney, E. Could machine learning fuel a reproducibility crisis in science. *Nature 608*, 7922 (2022), 250–251.
7. Feger, S.S., Dallmeier-Tiessen, S., Woźniak, P.W., and Schmidt, A. The role of HCI in reproducible science: Understanding, supporting and motivating core practices. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, 2019.
8. Wacharamanotham, C., Eisenring, L., Haroz, S., and Echtler, F. Transparency of CHI research artifacts: Results of a self-reported survey. *Proc. of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, 2020.
9. Shaer, O., Cooper, A., Mokryn, O., Kun, A.L., and Shoshan, H.B. AI-augmented brainwriting: Investigating the use of LLMs in group ideation. *Proc. of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, 2024, Article 1050.

Thomas Kosch is a professor at Humboldt University of Berlin studying how artificial intelligence can facilitate an understanding between users and interfaces using computational interaction design. He analyzes user behavior, context, and physiological metrics to develop interfaces that foster shared understanding, and refines HCI research methodologies to ensure scientific applicability, validity, replicability, and transparency.
→ thomas.kosch@hu-berlin.de

Sebastian Feger is a professor at TH Rosenheim, Germany, where he researches various domains, from smart tangibles to smart home privacy. His Ph.D. focused on designing for reproducible science and HCI's role in reproducible practices. He applies these experiences to integrating LLMs across UX research responsibly and reproducibly.
→ sebastian.feger@th-rosenheim.de

DOI: 10.1145/3695765