

In this section, we feature reports from conferences, symposia, workshops, and similar events, focusing on discussions where the boundaries of HCI and UX are being challenged and where debate is lively and ongoing.



# The State of Large Language Models in HCI Research: Workshop Report

Marianne Aubin Le Quéré, Cornell University, Hope Schroeder, Massachusetts Institute of Technology, Casey Randazzo, Rutgers University, Jie Gao, Singapore-MIT Alliance for Research and Technology

In 2023, four Ph.D. students across different institutions were experimenting with using large language models (LLMs) in HCI research. Aubin Le Quéré was labeling thousands of social media posts according to a qualitative codebook; Schroeder was generating news surveys on the fly; Randazzo was playing with generative agents; and Gao was creating tools for human-AI coding. All these projects began with a key question: *Could LLMs help us accomplish this research task?*

Soon, our assumptions were validated, and our first findings using LLMs were published. Yet, as we conducted this research, we felt we lacked the methodological scaffolding we usually rely on. *Did we document our methods properly? Did we take all the steps we needed to protect user data? Were others using similar methods to validate findings?* We connected over these shared concerns, and it became clear to us that addressing them was important.

These questions motivated the workshop “LLMs as Research Tools: Applications and Evaluations in HCI Data Work,” which gathered more than 40 researchers and practitioners at CHI 2024 [1]. HCI data work refers to the way researchers collect, interpret, and report on qualitative and quantitative data—processes LLMs are already reshaping. For example, models are being used to synthetically simulate data, perform thematic analysis, and even conduct interviews. When any new technology, tool, or method is folded into research processes, the community needs robust ways to evaluate them. LLMs are unique because they can be used for an array of roles and tasks across the research pipeline, giving rise to significant opportunities and challenges [2,3,4]. Unless we reflect on our standards for using LLMs,

we run the imminent risk of weakening the quality of our academic output.

With the rapid adoption of LLMs in both academic and industry research, a space was needed for “LLM curious” researchers to discuss their experiences. For a day, we brought together researchers from across HCI subfields to do the following:

- Discuss, reflect on, and share ongoing applications and challenges of using LLMs to work with data in HCI research.
- Discuss options for establishing methodological validity when using LLMs to work with data in HCI research.
- Discuss the primary critical and ethical questions around the use of LLMs to work with data in HCI research.

Elena Glassman, an assistant professor at Harvard University, opened with the keynote, “Interfaces for Better Characterizing and Leveraging Large Language Models.” Glassman presented user interface cases where LLMs can be deployed to help us make sense of data by, for example, grouping relevant text snippets together or color-coding similar entries. Her talk outlined the allure of using LLMs to work with data and how these models, with appropriate precautions, can be incorporated into existing tools.

Twelve speakers presented their research projects. Discussions were then focused around showcasing the diversity of current LLM applications in HCI data work, methods that evaluate the validity and rigor of work done with LLMs, and frameworks that contend with the ethics of using LLMs in research.

## KEY TAKEAWAYS

**LLMs are permeating the HCI research pipeline.** One thing was clear throughout the day: LLMs are being adopted rapidly

across academic communities, and their influence spans the research pipeline. In discussions about active research projects, researchers came to the table with simulated user environments and annotation approaches for large datasets in more complex ways than previously possible. The main areas of discussion around LLMs were researchers’ interest in working with synthetic data and using LLMs as annotators and to work with qualitative data. While these three use cases entailed different tasks, everyone felt that progress was outpacing evaluation, and that we as a community need to set norms for our work. Such norms could look different for various communities, since they are rooted in different research ontologies and can be evaluated differently.

**We need standards to evaluate the validity and quality of LLMs in HCI research.** While participants were fascinated by potential ways of using LLMs in research, they came from diverse empirical and critical HCI research traditions, leading to a robust discussion about evaluation. The appropriate type of evaluation for an LLM use can vary significantly depending on the research context. Several workshop papers explicitly proposed or evaluated a method, and participants felt that, where possible, we should rely on established frameworks to ensure validity. For example, those that discussed the use of LLMs for annotating data said they could use established standards, such as comparisons between LLMs and human gold standard datasets that use classic accuracy, recall, and precision measures. However, participants also felt an urgent need to tackle transparency and replicability challenges accelerated by LLMs. The nondeterministic nature of LLMs, a lack of standards for reporting prompts and model types, and the absence of

broadly accessible research tooling could all worsen attempts to standardize evaluation. Qualitative researchers in the workshop were particularly concerned about maintaining agency and deep engagement with the data when using LLMs and communicating processes to reviewers, since some traditions of qualitative research in HCI do not use inter-rater reliability as a measure of research rigor. Participants across all groups agreed that, while community norms for using LLMs have not yet been established, researchers must be explicit about their questions, tasks they are using an LLM to achieve, establishing data validity, and effectively communicating the credibility of the results.

*We need contextual norms that allow the ethical use of LLMs in HCI research.* Workshop participants had acute concerns about trust and ethics when using LLMs to work with data. Researchers brought up issues regarding LLMs and participant privacy, how they may bring bias and Western centrality into research, and the appropriateness of private companies influencing research. In particular, since human interviews and studies are key to HCI methods, participants were concerned about how to maintain trust with research participants if they are using LLMs to help with data analysis. Ethical concerns were strongly felt and divisive. For example, some people were curious to study synthetic personas, while others perceived this approach as a fundamental threat to HCI research. Since LLMs are emergent

and frequently changing, participants said that it could be difficult not only to properly validate the appropriateness and replicability of methods that incorporate LLMs but also to understand and keep up with the ethical concerns they pose. Junior researchers were concerned about receiving sufficient guidance to engage with these methods consistently and ethically. In this space, participants felt that more guidance and standards were required to move forward as a field.

## WHY HCI NEEDS TO BE AT THE CENTER OF THIS CONVERSATION

David Mimno, an associate professor at Cornell, wrapped up the workshop with the following observation, “GPT 3.5 has been around since 2020, but it wasn’t until ChatGPT—an easily accessible and free Web UI tool—was introduced that these tools became popular. The underlying technology is not what’s going to cause a revolution; user interaction systems that make them easy to use will.” This point should motivate all of us who study these tools from the perspective of HCI.

HCI researchers are experts on the usability of technology—the factor that catapulted LLMs into popular and scientific consciousness. We now find ourselves at a critical moment where usability, guidelines, and technological advancement must work together to ensure that LLMs are used responsibly in rigorous research. As an interdisciplinary field, it is our responsibility to

bring together the right stakeholders to accomplish this goal. This workshop was a first step toward this future, and we hope there will be many more to come.

## ACKNOWLEDGMENTS

We are grateful to our postdoc and faculty advisors—Ziv Epstein, Simon Perrault, David Mimno, Louise Barkhus, and Hanlin Li—for their guidance throughout this project. We are also grateful to the workshop attendees for their active participation.

## ENDNOTES

1. Aubin Le Quéré, M. et al. LLMs as research tools: Applications and evaluations in HCI data work. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, 2024, Article 479, 1–7.
2. Liao, Q.V. and Vaughan, J.W. AI transparency in the age of LLMs: A human-centered research roadmap. *Harvard Data Science Review*, 5 (2024).
3. Messeri, L. and Crockett, M.J. Artificial intelligence and illusions of understanding in scientific research. *Nature* 627, 8002 (2024), 49–58.
4. Palmer, A., Smith, N.A., and Spirling, A. Using proprietary language models in academic research requires explicit justification. *Nature Computational Science* 4, 1 (2024), 2–3.

**Marianne Aubin Le Quéré** is a Ph.D. candidate in information science at Cornell University. Her research focuses on how AI and other emerging technologies affect news and civic information consumption.  
→ [msa258@cornell.edu](mailto:msa258@cornell.edu)

**Hope Schroeder** is a Ph.D. student at the MIT Center for Constructive Communication and MIT Media Lab. She studies natural language processing and computational social science methods for making sense of the information ecosystem and in small group discourse.  
→ [hopes@mit.edu](mailto:hopes@mit.edu)

**Casey Randazzo** is a Ph.D. candidate in the School of Communication and Information at Rutgers University. She investigates how group interactions with AI agents shape the formation of organizing structures, offering critical insights for organizations looking to deploy AI to facilitate collective action.  
→ [cer124@scarletmail.rutgers.edu](mailto:cer124@scarletmail.rutgers.edu)

**Jie Gao** is a postdoctoral associate in the Mens, Manus, and Machina team at the Singapore-MIT Alliance for Research and Technology. Her research interests include human-AI collaboration, human-AI interaction, and AI for social science.  
→ [jie.gao@smart.mit.edu](mailto:jie.gao@smart.mit.edu)



Workshop participants in Oahu.