LLMs as Research Tools: Applications and Evaluations in HCI Data Work

Marianne Aubin Le Quéré
msa258@cornell.edu
Cornell Tech
USA

Hope Schroeder
MIT
USA

Casey Randazzo
Rutgers University, NJ
USA

Jie Gao
Singapore University of Technology and Design
Singapore

Ziv Epstein
Stanford University
Stanford, USA

Simon Perrault
Singapore University of Technology and Design
Singapore

David Mimno
Cornell University
USA

Louise Barkhuus
Rutgers University
USA
IT University of Copenhagen
Copenhagen, Denmark

Hanlin Li
The University of Texas at Austin
USA

ABSTRACT

Large language models (LLMs) stand to reshape traditional methods of working with data. While LLMs unlock new and potentially useful ways of interfacing with data, their use in research processes requires methodological and critical evaluation. In this workshop, we seek to gather a community of HCI researchers interested in navigating the responsible integration of LLMs into data work: data collection, processing, and analysis. We aim to create an understanding of how LLMs are being used to work with data in HCI research, and document the early challenges and concerns that have arisen. Together, we will outline a research agenda on using LLMs as research tools to work with data by defining the open empirical and ethical evaluation questions and thus contribute to setting norms in the community. We believe CHI to be the ideal place to address these questions due to the methodologically diverse researcher attendees, the prevalence of HCI research on human interaction with new computing and data paradigms, and the community's sense of ethics and care. Insights from this forum can contribute to other research communities grappling with related questions.

ACM Reference Format:
Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. 2024. LLMs as Research Tools: Applications and Evaluations in HCI Data Work. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613905.3636301

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05.
https://doi.org/10.1145/3613905.3636301

1 INTRODUCTION

Many scientific fields today are grappling with the changes recent advances in large language models (LLMs) have introduced to research methodology. LLMs are being used as computational tools with new interfaces and abilities for working with data, including for data collection, synthesis, and sensemaking. LLMs are being used to aid with typical research methods in HCI like surveys, user studies, and interviews: these models have been used to simulate data [22, 23, 32], facilitate qualitative coding and perform thematic analysis [14, 18, 47], and even conduct interviews [12]. However, the rate at which LLMs have been adopted as research tools to work with data—both inside and outside our research community—has outpaced our understanding of the empirical and ethical appropriateness of these tools to conduct these research tasks.

We propose a workshop to convene researchers at CHI to set the agenda for the use of LLMs as tools for working with data. We believe the CHI community to be a key venue to facilitate these far-reaching conversations. First, CHI is uniquely poised as a community that equally values qualitative, quantitative, and mixed methods work [8]. With LLMs, the distinction between qualitative and quantitative work may become blurred: for example, LLMs can scale qualitative codebooks, and conversely enable more nuanced labeling of large-scale datasets. We need insights from an interdisciplinary community to bring these perspectives together (e.g. [6]). Second, as the foremost conference that studies the interface of humans and computing, including in research environments (e.g. [31, 37, 51, 52]), CHI is a natural fit to explore how these technologies do and should affect humans doing data work. Third, CHI is a community that deeply considers the societal and ethical implications of technology. This ethical bent is paramount, since the data, representation, and privacy issues of LLMs make their use in methods ethically complex. Insights from this workshop can be useful beyond HCI to every community grappling with issues of data work with LLMs at this moment (e.g. [4, 50]).

We solicit participation in this workshop from academic and industry practitioners in the CHI community who are interested in

conducting data collection and sensemaking work that leverages LLMs as part of their methodology. We will welcome descriptions of research projects that have incorporated LLMs in their data collection and sensemaking workflows, speculative works on applications of LLMs to HCI research methods, and reflections and recommendations on the empirical or ethical evaluations of these tools in data work.

This workshop will be organized around three main objectives:

(1) Bring the community together to discuss, reflect, and share ongoing applications and challenges of using LLMs to work with data in HCI research
(2) Discuss options for establishing methodological validity when using LLMs to work with data in HCI research
(3) Discuss the primary critical and ethical questions regarding the use of LLMs to work with data in HCI research

After the workshop, we will publish a report that synthesizes the discussion and outlines a broad research agenda for the empirically and ethically sound use of LLMs in data collection, processing, and analysis in HCI research.

2 DATA WORK IN HCI

Much of HCI research involves working with some kind of data as part of the research process. Data work has been defined as "any human activity related to creating, collecting, managing, curating, analyzing, interpreting, and communicating data" [7]. In this workshop, we will consider the data work of interest to be data collection, data processing, and data analysis. We describe how these stages of data work intersect with HCI methods and LLM use.

LLMs already play a role in many stages of the data collection process, including synthesis of research data. In HCI research, data collection may involve running interview studies, deploying surveys, or collecting a relevant corpus to answer a research question. Researchers are exploring the use of synthetically generated data as part of research studies [2, 22, 43]. Researchers have also found that even if the judgments of real human crowdworkers are sought, crowdworkers are overwhelmingly using LLMs when answering researcher prompts [44]. These developments necessitate a deeper understanding of how to collect data using LLMs, and understand data produced by LLMs as research artifacts.

The data sensemaking process, which includes data processing and analysis, also has the potential to be transformed through the emerging use of LLMs to conduct research. Studies have already shown that LLMs can be successfully deployed to annotate data [15, 20, 36, 42] and facilitate human collaboration during the qualitative coding process [2]. LLM-generated explanations in open-ended generation tasks like FLUTE sometimes even exceed those produced by humans, which have previously been the gold standard [50]. Past work shows LLMs' ability to summarize dialog [17] and conversations [48], demonstrating their potential use as part of interview data analysis. The interactive capabilities of LLMs could even help researchers interact with their data in new, unstructured ways that lead to insight [19]. These recent developments all point to the promising but understudied emerging use of LLMs in data sensemaking work relevant to HCI research.

In the first part of the workshop, we will seek to characterize LLM use in HCI data work so far:

- How are HCI researchers using LLMs in their data work today?
- What does the use of LLMs in data work enable researchers to do?
- What are the challenges being encountered when trying to use LLMs in data work?
- How do applications of LLMs to data work differ for qualitative and quantitative research? Where do they blur?
- What are some potential future uses of LLMs in qualitative and quantitative HCI data work?

3 EMPIRICAL AND ETHICAL EVALUATIONS OF DATA WORK WITH LLMS

As the usage of LLMs in data work for HCI research proliferates, we need to ensure we have accurate, consistent, and agreed-upon ways to assess the reliability and validity of such methods. The application of specific research techniques often requires abiding by community norms for evaluation and agreement. For example, applying a codebook reliably to data often necessitates some proof of interrater reliability, which can be assessed using accepted metrics such as Cohen's Kappa [28]. Solving a similar problem, the domain of natural language processing has developed specific benchmarks for datasets which allows for the consistent evaluation of new models (e.g. superGLUE [46]; SQUAD [34]). Prior work has synthesized existing approaches for evaluation of qualitative data labels in HCI [27, 49], and others have looked at strategies to integrate crowdworkers labels into HCI data flows [24, 45]. However, it is not a given that these same methods for ensuring validity transfer over when using LLMs to perform a task.

There are specific validity issues that arise when applying LLMs to research tasks. For example, LLMs can be sensitive to even small changes in a prompt, and may not provide deterministic answers. These can pose a challenge to objective, over-time evaluation of the performance or suitability of an LLM to a task [30]. Additionally, LLMs are prone to unpredictable hallucinations, which may have a marked impact on the validity of some methods, but not on others. Finally, the most used LLMs currently are proprietary, which severely limits the ability of researchers to validate the training data, understand their mechanisms, or replicate results.

In the workshop, we will consider issues around methodological evaluations:

- Does using LLMs for annotation produce similar results as humans who do the annotation? Which humans? How would we evaluate this?
- What could be lost for qualitative research and researcher insight if we use LLMs for tasks previously done exclusively by researchers?
- What should we use as a gold standard of accuracy? How can we evaluate the accuracy for the tools built with LLMs for labeling subjective tasks like qualitative data?
- Is explainability important for building trust in LLM judgments about data? How should we build "appropriate" trust and reliance between researchers and LLMs?
- How will we address replicability issues if LLMs are used in data work, but closed-source models may be prone to change, with a potential effect on results?

The second domain for evaluation we will consider in this workshop is ethical evaluation when using LLMs for HCI data work. As such, this workshop will engage the perspective of researchers who have experience working on fairness, bias, privacy, and ethics in HCI and related spaces. Discussing ethical concerns is particularly salient since many university ethics boards have not yet presented a cohesive strategy for ethical use of LLMs in university research.

One major ethical area to consider is the issue of representational bias in language models, but also in the current gold standard of datasets. Machine learning models can inherit biases from their training data [13] or reflect discrimination that exists in societal contexts [9]. They may represent a specific viewpoint, as assessments have found that LLMs are most representative of Western societies and beliefs [3], which may perpetuate bias and discrimination if used in research. On the other hand, while human labelers and researchers are often the gold standard in evaluating accuracy, humans themselves are biased and may entrench an individual's point of view or structural biases [38]. The LLM-human coder tradeoff is anything but clear; for example, Gao et al. show that though AI-powered coding assistance improves inter annotator agreement, it may decrease the diversity of codes.

There are many other potential issues of concern with ethical use of LLMs in HCI data work. For example, do participants have to provide consent for analyzing their data through LLM tools, particularly when these tools are proprietary? This issue is particularly salient for vulnerable or marginalized communities who rely on researchers to responsibly represent their experiences [11, 35, 39]. There is also potential for intellectual property concerns to arise, particularly when LLMs significantly contribute to research findings [26]. Additionally, LLMs such as ChatGPT are also available only through cloud-based interfaces, which may produce unacceptable risks for private and proprietary data.

On the issue of critical and ethical evaluation, we will consider questions including:

- What privacy issues may arise for sensitive participant data if used in an LLM-powered toolkit?
- How can we tackle representation hazards presented by using tools that represent people differentially, including WEIRD data, or are prone to liberal beliefs and stereotypes?
- Are there certain domains where the use of LLMs for data work is more acceptable than others (for example, healthcare or law enforcement)?
- Are there contexts in which the risks outweigh the benefits, or the benefits outweigh the risks of using LLMs in data work for HCI?
- To what extent are ethical issues addressed through the use of open-source LLMs in these tasks? Are there new ethical issues that arise when using open-source LLMs as opposed to closed-source models?
- Can LLMs make analysis accessible to new kinds of researchers in ways that improve diversity in the research community by making some kinds of analysis easier?

4 PREVIOUS WORKSHOPS

There have been many workshops at CHI that have tackled the subject of AI more broadly, most recently [5, 16, 41]. In the last two years, the GenAICHI workshop has gathered a community around questions of how to study generative AI through the lens of HCI research [29]. Another relevant workshop is the 2023 workshop which considered the implications of AI-driven interactive writing assistants [10]. These two workshops focused on how HCI researchers can study new tools enabled by generative AI and LLMs.

In this "LLMs as Research Tools" workshop, we ask not how we can study new systems, but how to make sense of the ways that new systems impact existing methods of data collection and sensemaking in our field. Research methods and HCI data work has been a topic of interest for prior CHI workshops (e.g. [1, 25, 33]), but this workshop would be the first that seeks to map and evaluate how LLMs may be used for data work in HCI. Highlighting the timeliness and need for this space to set norms, a recently-published CSCW 2023 SIG also engages with similar topics [40].

5 ORGANIZERS

The workshop organizing team is a mix of researchers of varying seniority across seven institutions and three countries. The organizers are all active researchers in the areas of HCI, generative AI and LLMs, or AI ethics.

- **Marianne Aubin Le Quéré** is an Information Science PhD Candidate at Cornell Tech. In her work, she uses text-as-data methods to understand how technological innovations impact digital news consumption. She has studied the impact of AI on the news industry, performed studies that leverage LLMs for qualitative coding, and is interested to understand how HCI analysis standards may be shifted through the use of emergent LLMs.
- **Hope Schroeder** (she/her) is a PhD 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. She is interested in how advances in generative models affect methods of investigating communication and interpersonal connection.
- **Casey Randazzo** (she/her) is a Ph.D. candidate at the School of Communication and Information at Rutgers University. Casey focuses on the role of computer-mediated communication in recovery from individual and community-wide trauma. By investigating these areas, Casey promotes the implementation and adoption of trauma-informed HCI principles in platform design and research methodologies. Her efforts aim to prevent retraumatization and ensure that users' voices are accurately and empathetically represented in the data.
- **Gao Jie** (she/her) is a Ph.D. candidate at the Singapore University of Technology and Design in Singapore. Her research during the PhD explores the use of AI and LLMs to enhance HCI research methodologies, particularly qualitative analysis; understanding the trust and reliance challenges in human-AI interactions. Her goal is to improve human-AI collaboration using AI and LLMs and to deepen insights into how humans and AI interact.

- **Ziv Epstein** is a Postdoctoral Fellow at the Stanford Institute for Human-Centered AI. His research focuses on translating insights from design and the social sciences into the development of sociotechnical systems such as generative AI and social media platforms. Ziv has published papers in venues such as the general interest journals Nature, Science and PNAS , as well as top-tier computer science proceedings such as CHI and CSCW.
- **Simon T. Perrault** is an Assistant Professor at the Singapore University of Technology and Design in Singapore. His research lately leverages Large Language Models for self-reflection, detecting misinformation and as a support for researchers performing collaborative qualitative coding. He obtained his PhD degree in computer science from Telecom ParisTech (now Institut Polytechnique, France) in 2013. Prior to joining SUTD, he was a Visiting Professor at the Korean Advanced Institute of Science and Technology (Korea), Assistant Professor at Yale-NUS College (Singapore) and Postdoctoral Fellow at the National University of Singapore.
- **David Mimno** is an Associate Professor in the department of Information Science at Cornell University who studies new methods in natural language processing, computational social science, and digital humanities. He holds a PhD from UMass Amherst and was previously the head programmer at the Perseus Project at Tufts and a researcher at Princeton University. He has been a key organizer for events that center data work and AI tools, including the Text as Data conference 2022 and the ICML workshop on Generative AI and Law 2023.
- **Louise Barkhuus** is a Visiting Professor at Rutgers university and a Professor at the IT University of Copenhagen. She studies how people interact with mobile devices, in particular location based services, and how privacy can be preserved through better design. Her approach to AI data analysis is from a qualitative perspective, where she critically approach new analysis methods. Her future approach is a focus on sensor data as part of AI input.
- **Hanlin Li** is an Assistant Professor at UT Austin. Her research aims to inform policy and design interventions to incentivize responsible data collection and use. She examines the societal and economic impact of data generated by the public, from rating data to social media comments. Her work sits at the intersection of data governance and human-computer interaction.

6 HYBRID WORKSHOP AND ASYNCHRONOUS ENGAGEMENT

We are planning to support a hybrid workshop. Due to difficulties in obtaining visas for travel, funding constraints, and ethical concerns about attending CHI in Hawaii [21], we anticipate that people may wish to participate remotely. Since Hawaii is in a different timezone than many attendees at home, we will have two optional 2-hour events in the week prior to the workshop. These will happen at different times to be as inclusive of diverse timezones as possible. These will cover the same general topics as the main workshop, in a condensed format, and leverage breakout rooms to encourage attendees to get to know each other. Optionally, remote attendees may also choose to participate in the live workshop synchronously to the in-person attendees. We will leverage Zoom and A/V equipment to stream the in-person talks to remote attendees, and call upon remote attendees in the group sharing sessions. Exact details may be adjusted depending on the total number, geographic distribution, and amount of interest from remote participants.

We will use the website as a central hub for asynchronous engagement. With author permission, paper submissions will be published either on ArXiv or on the workshop website. With speaker consent, all talks will be recorded, transcribed, and made available on the website for access after the workshop. For continued engagement after the workshop, we will set up a specific channel in the conference Slack/Discord or set up a fully separate Slack/Discover channel.

7 WORKSHOP ACTIVITIES

7.1 Pre-workshop Plans

Prior to the workshop, the committee will focus on advertising, reviewing proposals, and connecting workshop attendees. The workshop website (https://sites.google.com/view/llmsindatawork) will house key information like the CfP deadlines, schedules, and accepted proposals. We will reach out to potential participants through listservs, social media, personal networks, and events. We feel that the workshop topic is timely and will appeal to many participants.

To field proposals to the workshop, we will have the four organizing PhD students overseeing the review process. Between the organizing PhD students, we hope to get two reviewers per submission, and accept papers such that we can hold a workshop of 30-40 people. Submissions will be discussed with the full organizing committee before participants will be invited to participate. We will select submissions based on how well the submissions can contribute towards the stated aims of the workshop. We will also balance coverage of attendees between qualitative and quantitative researchers, those interested in ethical or empirical evaluations, and junior and more senior researchers.

7.2 Workshop Structure

The workshop will roughly be divided into three key parts. The morning session will be devoted to mapping current and future approaches and challenges to integrating LLMs into HCI data work. The afternoon session will be split between discussing empirical and ethical implications of LLMs being integrated into HCI data work. Overall, we will hold talks when people arrive/come back from lunch to encourage participants to have shared topics of conversations and set the tone for the workshop. Talks will be followed by breakout discussions for workshop participants to share ideas more informally and broadly with each other. A proposed workshop schedule is shown in Table 1.

7.2.1 Morning Session: Applications and Challenges of LLMs in HCI data work. The morning session will be structured around orienting the community to ongoing and future potential efforts to integrate

Title	Description
9am - 9:45am	Introduction to the workshop and keynote talk by a senior researcher on uses of LLMs to conduct HCI data work
9:45am - 10:20am	Three 8-minute talks from invited workshop participants (focused on applications and challenges), with Q&A
10:20am - 10:45am	Morning break
10:45am - 11:30am	Discussions among participants of applications of LLMs to conduct HCI data work and challenges encountered
11:30am - 11:45am	Group sharing & discussion
11:45am - 1:15pm	Break for lunch
1:15pm - 2pm	Four 8-minute talks from invited workshop participants (2 about empirical topics, 2 about ethical topics), with Q&A
2pm - 2:45pm	Discussions among participants of empirical evaluations of LLMs to conduct data work
2:45pm - 3pm	Group sharing & discussion
3pm - 3:30pm	Afternoon break
3:30pm - 4:15pm	Discussions among participants of critical and ethical issues that arise from the use of LLMs to conduct HCI data work
4:15pm - 4:30pm	Group sharing & discussion
4:30pm - 5pm	Idea Synthesis & wrap-up
5pm - evening	Optional Socializing

Table 1: Proposed Workshop Schedule

LLMs in HCI data work. We will conduct a general welcome, followed by a 30-minute keynote talk by a senior researcher in the field on the use of large language models in HCI data work. Keynote speakers will be confirmed closer to the date of the workshop. This will be followed by three, hand-selected invited speakers from the participants. They will each be given 10 minutes to present, and then the workshop organizers will moderate a brief Q&A with the speakers. We will then go into a quick coffee break to allow participants to mingle with each other and start forming connections. For the main discussion session, we will distribute participants to sit at tables loosely organized by methodological area (which we will derive from the workshop submissions). Participants will then discuss their past, current, and future work in the field, and be encouraged to think through some of the challenges they encountered. The wrap-up session will serve to synthesize the discussion to make a main list of the ways that LLMs can be applied to HCI data work.

7.2.2 Afternoon Session: Empirical and ethical evaluations of LLMs in HCI data work. The afternoon session will be geared to help participants think critically about how to evaluate the emergent methods for using LLMs in HCI data work that were discussed in the morning. Since empirical and critical evaluation are two distinct (yet related) concepts, we will devote separate times to both of them to ensure everyone engages with both. We will again invite selected participants who are focused on evaluation-based topics/questions to present their work and briefly moderate Q&A. Following these talks, we will have one breakout session for discussion of empirical evaluations of LLMs in data work, and one breakout session for discussions of ethical evaluations of LLMs in data work. There will be an afternoon break, and informal socializing after the workshop. In the workshop wrap-up, we will recap the discussions that took place throughout the day, and ask participants if they are interested to write up a proposed "research agenda" together, which defines the central open questions in the field.

8 POST-WORKSHOP PLANS

For our post-workshop plan, we will publish a summary of the workshop discussions on the main website. Additionally, we anticipate that there is a space to publish a report to outline the research agenda for data work with LLM-enabled tools in HCI. We will discuss these proposed outcomes at the end of the workshop with participants. If there is sufficient interest, we will seek to publish the agenda either as a public-facing report or as a submission for a relevant journal.

9 CALL FOR PARTICIPATION

Broadly accessible large language models (LLMs) stand to fundamentally reshape the HCI community's suite of methods for working with data. To date, LLM tools have already been used to facilitate qualitative coding, perform thematic analysis, and even mediate interviews or simulate user data. However, we lack a broader understanding of: 1) How LLM-based methods are being used to work with data in HCI, 2) What empirical evaluation strategies are acceptable to the community for establishing validity of data work conducted with LLMs, and 3) How to critically and ethically use LLM methods in HCI research. The goal of this workshop is to gather a community of researchers interested in these topics to map current approaches as a community, documenting the challenges encountered, and norm-set in this rapidly evolving field.

For this hybrid CHI 2024 workshop, we invite junior and senior academics, researchers, and practitioners to submit extended abstracts or short papers. Interested participants should submit a 2-4 page (not including references) proposal using the CHI Extended Abstracts format. We invite submissions including empirical works-in-progress, research or research proposals, and provocations, critical approaches, or position papers. Broadly, paper topics should relate to the use of LLMs to work with data in HCI, epistemic validity and methodological evaluations, and/or critical and ethical perspectives on the use of LLM methods in HCI research. One participant from each submission must register for the workshop and at least one day of the conference. Submissions will be published on the workshop website. We welcome perspectives on applications, as well as methodological and critical evaluation from researchers of different methodological backgrounds, including NLP, qualitative research, user research, and beyond.

REFERENCES

[1] Ferran Altarriba Bertran, Soomin Kim, Minsuk Chang, Ella Dagan, Jared Duval, Katherine Isbister, and Laia Turmo Turmo Vidal. 2021. Social Media as a Design and Research Site in HCI: Mapping Out Opportunities and Envisioning Future Uses. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3411763.3441311

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (July 2023), 337–351. https://doi.org/10.1017/pan.2023.2 Publisher: Cambridge University Press.

[3] Mohammad Atari, Mona J. Xue, Peter S. Park, Damián Blasi, and Joseph Henrich. 2023. Which Humans? https://doi.org/10.31234/osf.io/5b26t

[4] Christopher A. Bail. 2024. Can Generative AI Improve Social Science? (Jan. 2024). https://doi.org/10.31235/osf.io/rwtzs Publisher: OSF.

[5] Gagan Bansal, Alison Marie Smith-Renner, Zana Buçinca, Tongshuang Wu, Kenneth Holstein, Jessica Hullman, and Simone Stumpf. 2022. Workshop on Trust and Reliance in AI-Human Teams (TRAIT). In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3491101.3503704

[6] Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1397–1410. https://doi.org/10.1002/asi.23786 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23786.

[7] Claus Bossen, Kathleen H Pine, Federico Cabitza, Gunnar Ellingsen, and Enrico Maria Piras. 2019. Data work in healthcare: An Introduction. *Health Informatics Journal* 25, 3 (Sept. 2019), 465–474. https://doi.org/10.1177/1460458219864730 Publisher: SAGE Publications Ltd.

[8] Duncan P. Brumby, Ann Blandford, Anna L. Cox, Sandy J. J. Gould, and Paul Marshall. 2017. Understanding People: A Course on Qualitative and Quantitative HCI Research Methods. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 1170–1173. https://doi.org/10.1145/3027063.3027103

[9] Robyn Caplan, Joan Donovan, Lauren Hanson, and Jeanna Matthews. 2018. Algorithmic Accountability: A Primer. https://datasociety.net/library/algorithmic-accountability-a-primer/ Publisher: Data & Society Research Institute.

[10] Minsuk Chang, John Joon Young Chung, Katy Ilonka Gero, Ting-Hao Kenneth Huang, Dongyeop Kang, Mina Lee, Vipul Raheja, and Thiemo Wambsganss. 2023. The Second Workshop on Intelligent and Interactive Writing Assistants. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–5. https://doi.org/10.1145/3544549.3573826

[11] Janet X. Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. 2022. Trauma-Informed Computing: Towards Safer Technology Experiences for All. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3491102.3517475

[12] Felix Chopra and Ingar Haaland. 2023. Conducting Qualitative Interviews with AI. https://doi.org/10.2139/ssrn.4583756

[13] Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press. https://library.oapen.org/handle/20.500.12657/43542 Accepted: 2020-12-15T13:38:22Z.

[14] Stefano De Paoli. 2023. Can Large Language Models emulate an inductive Thematic Analysis of semi-structured interviews? An exploration and provocation on the limits of the approach and the model. https://doi.org/10.48550/arXiv.2305.13014 arXiv:2305.13014 [cs].

[15] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2023. Is GPT-3 a Good Data Annotator? http://arxiv.org/abs/2212.10450 arXiv:2212.10450 [cs].

[16] Upol Ehsan, Philipp Wintersberger, Elizabeth A Watkins, Carina Manger, Gonzalo Ramos, Justin D. Weisz, Hal Daumé Iii, Andreas Riener, and Mark O Riedl. 2023. Human-Centered Explainable AI (HCXAI): Coming of Age. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3544549.3573832

[17] Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization. https://doi.org/10.48550/arXiv.2105.12544 arXiv:2105.12544 [cs].

[18] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2023. CollabCoder: A GPT-Powered Workflow for Collaborative Qualitative Analysis. https://doi.org/10.48550/arXiv.2304.07366 arXiv:2304.07366 [cs].

[19] Adam Hayes. 2023. "Conversing" with Qualitative Data: Enhancing Qualitative Research through Large Language Models (LLMs). https://doi.org/10.31235/osf.io/yms8p

[20] Michael Heseltine and Bernhard Clemm von Hohenberg. 2023. Large Language Models as a Substitute for Human Experts in Annotating Political Text. https://doi.org/10.31219/osf.io/cx752

[21] Josiah Hester. 2023. Why is CHI in Hawaii. https://www.chiinhawaii.info

[22] John J. Horton. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? https://doi.org/10.3386/w31122

[23] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3544548.3580688

[24] Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020. Best Practices for Crowd-based Evaluation of German Summarization: Comparing Crowd, Expert and Automatic Evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy (Eds.). Association for Computational Linguistics, Online, 164–175. https://doi.org/10.18653/v1/2020.eval4nlp-1.16

[25] Marina Kogan, Aaron Halfaker, Shion Guha, Cecilia Aragon, Michael Muller, and Stuart Geiger. 2020. Mapping Out Human-Centered Data Science: Methods, Approaches, and Best Practices. In *Companion Proceedings of the 2020 ACM International Conference on Supporting Group Work (GROUP '20)*. Association for Computing Machinery, New York, NY, USA, 151–156. https://doi.org/10.1145/3323994.3369898

[26] Katherine Lee, A. Feder Cooper, and James Grimmelmann. 2023. Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain. https://doi.org/10.2139/ssrn.4523551

[27] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 72:1–72:23. https://doi.org/10.1145/3359174

[28] Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica* 22, 3 (Oct. 2012), 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

[29] Michael Muller, Lydia B Chilton, Anna Kantosalo, Q. Vera Liao, Mary Lou Maher, Charles Patrick Martin, and Greg Walsh. 2023. GenAICHI 2023: Generative AI and HCI at CHI 2023. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–7. https://doi.org/10.1145/3544549.3573794

[30] Arvind Narayanan. 2023. Evaluating LLMs is a minefield. https://www.aisnakeoil.com/p/evaluating-llms-is-a-minefield

[31] Anna-Marie Ortloff, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. 2023. Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of an Interview and Survey Study on Security Advice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3580766

[32] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. https://doi.org/10.48550/arXiv.2304.03442 arXiv:2304.03442 [cs].

[33] Kathleen Pine, Claus Bossen, Naja Holten Møller, Milagros Miceli, Alex Jiahong Lu, Yunan Chen, Leah Horgan, Zhaoyuan Su, Gina Neff, and Melissa Mazmanian. 2022. Investigating Data Work Across Domains: New Perspectives on the Work of Creating Data. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3491101.3503724

[34] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

[35] Casey Randazzo and Tawfiq Ammari. 2023. "If Someone Downvoted My Posts—That'd Be the End of the World": Designing Safer Online Spaces for Trauma Survivors. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581453

[36] Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire Robertson, and Jay J. Van Bavel. 2023. GPT is an effective tool for multilingual psychological text analysis. https://doi.org/10.31234/osf.io/sekf5

[37] Tim Rietz and Alexander Maedche. 2021. Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445591

[38] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. https://doi.org/10.48550/arXiv.2111.07997 arXiv:2111.07997 [cs].

[39] Carol F Scott, Gabriela Marcu, Riana Elyse Anderson, Mark W Newman, and Sarita Schoenebeck. 2023. Trauma-Informed Social Media: Towards Solutions for Reducing and Healing Online Harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3544548.3581512

[40] Hong Shen, Tianshi Li, Toby Jia-Jun Li, Joon Sung Park, and Diyi Yang. 2023. Shaping the Emerging Norms of Using Large Language Models in Social Computing Research. http://arxiv.org/abs/2307.04280 arXiv:2307.04280 [cs].

[41] Joon Gi Shin, Janin Koch, Andrés Lucero, Peter Dalsgaard, and Wendy E. Mackay. 2023. Integrating AI in Human-Human Collaborative Ideation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3544549.3573802

[42] Petter Törnberg. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. https://doi.org/10.48550/arXiv.2304.06588 arXiv:2304.06588 [cs].

[43] Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science. http://arxiv.org/abs/2305.15041 arXiv:2305.15041 [cs].

[44] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. http://arxiv.org/abs/2306.07899 arXiv:2306.07899 [cs].

[45] Shaun Wallace, Tianyuan Cai, Brendan Le, and Luis A. Leiva. 2022. Debiased Label Aggregation for Subjective Crowdsourcing Tasks. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3491101.3519614

[46] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html

[47] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 75–78. https://doi.org/10.1145/3581754.3584136

[48] Diyi Yang and Chenguang Zhu. 2023. Summarization of Dialogues and Conversations At Scale. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics, Dubrovnik, Croatia, 13–18. https://doi.org/10.18653/v1/2023.eacl-tutorials.3

[49] Himanshu Zade, Margaret Drouhard, Bonnie Chinh, Lu Gan, and Cecilia Aragon. 2018. Conceptualizing Disagreement in Qualitative Coding. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173733

[50] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can Large Language Models Transform Computational Social Science? http://arxiv.org/abs/2305.03514 arXiv:2305.03514 [cs].

[51] Sena Çerçi, Marta E. Cecchinato, and John Vines. 2021. How Design Researchers Interpret Probes: Understanding the Critical Intentions of a Designerly Approach to Research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445328

[52] Asil Çetin, Torsten Moeller, and Thomas Torsney-Weir. 2021. CorpSum: Towards an Enabling Tool-Design for Language Researchers to Explore, Analyze and Visualize Corpora. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3411764.3445145