Human-Centered Privacy Research in the Age of Large Language Models

Tianshi Li
tia.li@northeastern.edu
Northeastern University
Boston, MA, USA

Sauvik Das
sauvik@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Hao-Ping (Hank) Lee
haopingl@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Dakuo Wang
d.wang@neu.edu
Northeastern University
Boston, MA, USA

Bingsheng Yao
arthuryao33@gmail.com
Rensselaer Polytechnic Institute
Troy, NY, USA

Zhiping Zhang
zhip.zhang@northeastern.edu
Northeastern University
Boston, MA, USA

ABSTRACT

The emergence of large language models (LLMs), and their increased use in user-facing systems, has led to substantial privacy concerns. To date, research on these privacy concerns has been model-centered: exploring how LLMs lead to privacy risks like memorization, or can be used to infer personal characteristics about people from their content. We argue that there is a need for more research focusing on the human aspect of these privacy issues: e.g., research on how design paradigms for LLMs affect users' disclosure behaviors, users' mental models and preferences for privacy controls, and the design of tools, systems, and artifacts that empower end-users to reclaim ownership over their personal data. To build usable, efficient, and privacy-friendly systems powered by these models with imperfect privacy properties, our goal is to initiate discussions to outline an agenda for conducting human-centered research on privacy issues in LLM-powered systems. This Special Interest Group (SIG) aims to bring together researchers with backgrounds in usable security and privacy, human-AI collaboration, NLP, or any other related domains to share their perspectives and experiences on this problem, to help our community establish a collective understanding of the challenges, research opportunities, research methods, and strategies to collaborate with researchers outside of HCI.

CCS CONCEPTS

• **Security and privacy** → **Human and societal aspects of security and privacy**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Large language models (LLMs), Generative AI, Privacy, Human-Computer Interaction

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

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3643983

ACM Reference Format:
Tianshi Li, Sauvik Das, Hao-Ping (Hank) Lee, Dakuo Wang, Bingsheng Yao, and Zhiping Zhang. 2024. Human-Centered Privacy Research in the Age of Large Language Models. 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, 4 pages. https://doi.org/10.1145/3613905.3643983

1 BACKGROUND

Large language models (LLMs) are transforming people's lives in many ways, but also present numerous risks — and chief among these risks is privacy. The NLP and system security communities have initiated extensive research into these models, focusing on the new privacy challenges they present and their capabilities for preserving user privacy. One major problem is that these models can memorize and output training data [2, 3, 20]. As the models are trained on vast amounts of data, including user data, this has raised new data leak risks. For instance, research has found that prompting the model to continuously output "poem" can trick it into leaking training data verbatim [14]. Beyond memorization, LLMs can be used to extract personal attributes of individuals from seemingly harmless text [17]. For example, given the text *"I always get stuck there waiting for a hook turn"*, LLMs can help malicious actors infer that this person is in Melbourne because a hook turn is a traffic maneuver particularly used there. Research has also shown that LLMs lack the commonsense about social privacy norms, and have trouble keeping a secret [13] and that instruction-tuned models can be easily tricked by third-party adversaries to ignore privacy-protecting instructions [4].

Despite the privacy issues exhibited in these models and the lack of effective defensive methods, we are witnessing a rapidly growing trend of LLMs being integrated into interactive computing systems and placed in users' hands. The most high-profile LLM application — LLM-based conversational agents (CAs), such as ChatGPT — are increasingly being incorporated into high-stakes application domains including healthcare [11], finance [5, 6, 18], and personal counseling [7, 10]. However, Zhang et al. [21] found that the high utility of the tool and the human-like interactions encourage users to share sensitive and personally identifiable information with LLM-based CAs. Despite this, users constantly face challenges in protecting their privacy due to the inherent tension between privacy and utility, their flawed mental models, and dark patterns in the design of

privacy management features [21]. Given these challenges, we believe the HCI community has a responsibility to foster a paradigm shift in LLM-centered privacy research. This shift should move from research that solely investigates the privacy risks entailed by a model, to research that empowers *humans* to act on and specify their privacy preferences, when interacting with LLM-powered interactive systems, in a usable, convenient, efficient, and effective way. Below are example questions that we hope our research community can explore:

- What do users perceive as the privacy risks of LLMs? How do these perceptions align with the known risks?
- How can we promote data sovereignty in users' use of LLMs?
- How do users' (mis)perceptions of the privacy risks of LLMs affect their ability to manage their privacy when using applications/services powered by LLMs?
- How does the training of LLMs on vast web-scraped data affect users' privacy perceptions and behaviors in general (non-LLM) online disclosure?
- How do different methods of designing interactions powered by LLMs affect users' awareness of how LLMs process their data and privacy-related behaviors?
- How can we educate the general public about the emerging privacy risks entailed by LLMs?
- How can we aid users in dealing with the trade-offs between privacy and other design deliberations (e.g., utility, convenience)?
- What are privacy challenges in personalized LLM-based agents, especially when they are integrated into socio-technical systems?
- How can we manage the tension between an individual user's privacy vs. other values of societal importance, such as safety and alignment efforts?
- How can regulatory efforts be designed to promote the development of privacy-respectful LLM-based systems?
- As more people start writing code with the help of LLMs, how do LLMs affect practitioners' abilities to handle privacy?

The primary goal of this Special Interest Group (SIG) is to bring together researchers with backgrounds in usable security and privacy, human-AI interaction, NLP, or any other related domains to collectively outline a research agenda to address the pressing and emergent privacy challenges entailed by large language models. To facilitate concrete progress towards this goal, below we outline four key areas of focus for the SIG: Understanding Privacy Challenges for Users; Designing Privacy-friendly Interfaces of LLM-based Systems; Building Usable Tools for Privacy Management for LLM-based Systems; Challenges and Solutions Beyond Individual Users.

2 UNDERSTANDING PRIVACY CHALLENGES FOR USERS

In LLM-based systems, one end-to-end model is usually expected to serve all the requests of varied use cases. However, privacy concerns are contextualized and subjective, which means that studying privacy risks solely at the model level can yield an incomplete understanding of real-world issues. Therefore, we believe it is important to investigate research questions such as: What is the impact, on user privacy, of interactions with LLMs in different contexts? What do users perceive as the most significant risks? And how do these perceived risks align or differ from our understanding of the actual risks? This situated, human-centered research method can offer complementary insights to model-centered research for designing privacy-respectful LLM-based systems. Taking research on the LLM memorization risks as an example, experiments with different models can reveal that memorization positively correlates with the size of the model and the frequency of text occurrence [2]. Furthermore, user interviews have revealed that users are more aware and concerned about the risks of memorization when they use ChatGPT to revise original writings, such as novels and research papers, due to concerns about idea theft [21]. The former finding assists model developers in estimating general risks, while the latter is instrumental in designing interfaces that alert users to specific privacy risks when handling tasks with significant privacy implications.

LLMs are trained on tremendous amounts of web scraped data. This suggests a inherently surveillant nature of LLMs, which means that the privacy impact is not limited to direct interactions with LLMs, but can also occur in other non-LLM-related online disclosures. The lack of transparency of training data of the proprietary LLMs has become a focal problem in the ML community, while there is relatively less discussion on the user privacy aspect. How does training LLMs on vast web-scraped data impact the risks of users' general online disclosure? How can we assist users in understanding and preventing these risks? Moreover, as people frequently disclose their communications in private (e.g., emails) or semi-private (e.g., Facebook group posts) contexts with others to ChatGPT (discussed as the interdependent privacy issues in Zhang et al. [21]), to what extent can this affect people's perceptions of privacy and the social dynamics of online activities?

3 DESIGNING PRIVACY-FRIENDLY INTERFACES OF LLM-BASED SYSTEMS

In this aspect, we are interested in one main question: How do different types of interactions with LLMs affect users' privacy-related mental models and behaviors? In addition to LLM-based conversational agents, there are other applications built with LLMs that afford other LLM-powered interactions. For example, one less explicit type of LLM-powered interaction is embedding LLM-based autocompletion in a text editor in a web browser or other desktop applications such as GitHub Copilot. Some applications employ less direct interactions between users and LLMs, including real-world products (e.g., Zoom's AI Companion to summarize meetings for attendees) and academic research project that translates natural language commands to programming language using LLMs [19]. In the above examples, LLMs play a role with different levels of explicitness, which may affect users' mental models and privacy concerns. In LLM-based systems, a current privacy threat is due to the status quo of API-based development. Except for big companies like Google that can host their own models, most of these systems incorporate LLMs via web APIs (e.g., OpenAI APIs, open-source LLM endpoints). This means that data sharing with a third party — the company that hosts the LLM APIs — may not be clear to users. In fact, Zhang et al. [21] discussed an intriguing example in which a user perceived GitHub Copilot as safer than ChatGPT due to the misconception that GitHub Copilot operates entirely on the device.

The interaction modality is another dimension that could potentially impact user privacy. Prior research has suggested that the human-like interactions provided by LLMs may prompt users to disclose more sensitive information [8, 9, 21]. As the multimodal LLMs support voice-based or even video-based interactions (e.g., Google Gemini), there is also an open question about how the more streamlined interactions affect users' disclosure behaviors.

4 BUILDING USABLE TOOLS FOR PRIVACY MANAGEMENT FOR LLM-BASED SYSTEMS

NLP and system security researchers have investigated privacy-preserving techniques that can be applied throughout the model training and inference phases [16]. These backstage strategies may not be helpful for users in managing their context-specific privacy requirements. This points us to a much needed research direction: developing user-facing privacy management tools for LLM-based systems. We further categorize it into two problems: 1) helping users sanitize their input; and 2) helping users censor LLM outputs that may contain personal data.

The challenge lies in managing the tradeoff among utility, convenience, and privacy, necessitating interdisciplinary collaborations among HCI, NLP, and system security. Research with LLM-based CA users has shown that users occasionally remove sensitive information from their input, while this manual process is tedious and easily forgotten [21]. Sometimes, the task they want to achieve is inherently privacy-sensitive (e.g., personal counseling). In the absence of support to manage the trade-off between privacy and utility, most of the time, they had to sacrifice privacy completely for utility. Another privacy/utility trade-off example is the choice between local models and server-based models.

There has been a lot of interest in building personalized LLM agents in HAI and NLP research. However, infusing personal information into LLMs raises significant privacy concerns, especially when the LLMs are used for social tasks that involve other people (e.g., email writing, meeting note taking [13]). In addition to improving the models' capacities to adhere to privacy norms, there is also a need to build HCI systems that allow users to specify privacy preferences, exercise control, avoid failure cases, and establish trust and accurate mental models of the systems' capabilities.

5 CHALLENGES AND SOLUTIONS BEYOND INDIVIDUAL USERS

Challenges. We have discussed challenges in protecting individual users' privacy. However, when discussing the human-centered privacy research agenda, we must also consider the backdrop that LLMs have posed challenges to other issues of broad societal importance, such as safety, ethics, and responsible AI use. Solutions to these challenges can conflict with privacy issues. For example, some levels of monitoring of LLM usage by API providers (e.g., OpenAI) or organizations may be necessary to identify and curtail abusive uses of LLMs — e.g., create fake news stories, facilitate cybersecurity attacks — or prevent accidental leakage of proprietary data. As another example, as ChatGPT and similar language model-based conversational agents make powerful AI more accessible to everyone, they also distribute the obligation to use AI responsibly. For example, publishers like ACM and Elsevier require authors to transparently report their use of generative AI in their manuscripts. However, Zhang et al. [21] highlighted a new privacy concern related to the fear of individuals being found out for using ChatGPT. This privacy concern could potentially hinder the responsible disclosure of LLM and other generative AI use. What other considerations need to be taken into account besides safeguarding personal privacy, and how can HCI research assist in understanding and addressing these issues?

Solutions. HCI and usable privacy researchers have a long history of providing recommendations for policy making and setting industry standards based on solid user research. To address privacy challenges in everyday applications powered by LLMs, we want to initiate a discussion on how we can maintain our successful track record in engaging with regulatory efforts. Our aim is to promote the design and development of LLM systems that respect privacy.

Notably, the prevalence of LLMs as a programming assistant tool may also have an impact on the developers' privacy practices. Research has uncovered security vulnerabilities in LLM-generated code [1, 15], which suggests there may also be an issue with privacy. Prior research has shown that developers often lack awareness of privacy issues in their code [12]. Therefore, if privacy issues exist in code generated by LLM, these issues may persist, potentially leading to more widespread privacy violations in general software.

6 CONCLUSION

The advent of large language models (LLMs) has brought about significant privacy challenges. To build usable, efficient, and privacy-friendly systems powered by these models with imperfect privacy properties, we emphasize the importance of human-centered privacy research in LLMs to complement the extensive model-centered research. This Special Interest Group (SIG) aims to provide a much-needed space for researchers who are interested in tackling these challenges to openly discuss the problems, research opportunities, research methods, and strategies to collaborate with researchers outside of HCI, such as NLP, system security, and public policy.

REFERENCES

[1] Owura Asare, Meiyappan Nagappan, and N. Asokan. 2023. Is GitHub's Copilot as bad as humans at introducing vulnerabilities in code? *Empirical Software Engineering* 28, 6 (September 2023). https://doi.org/10.1007/s10664-023-10380-1

[2] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646* (2022).

[3] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting Training Data from Large Language Models.. In *USENIX Security Symposium*, Vol. 6.

[4] Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. 2023. Can Language Models be Instructed to Protect Personal Information? *arXiv preprint arXiv:2310.02224* (2023).

[5] Sheryl Estrada. 2023. A startup CFO used ChatGPT to build an FP&A tool—here's how it went. https://fortune.com/2023/03/01/startup-cfo-chatgpt-finance-tool/ Accessed: 09/11/2023.

[6] Pedro Ferreira. 2023. Can ChatGPT Improve Technical Analysis and Trading Techniques? https://www.financemagnates.com/trending/can-chatgpt-improve-technical-analysis-and-trading-techniques/ Accessed: 09/11/2023.

[7] Thomas Germain. 2023. A Mental Health App Tested ChatGPT on Its Users. The Founder Said Backlash Was Just a Misunderstanding. https://gizmodo.com/mental-health-therapy-app-ai-koko-chatgpt-rob-morris-1849965534/ Accessed: 09/11/2023.

[8] Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda van Noort, and Edith Smit. 2020. *Privacy Concerns in Chatbot Interactions.* Springer International Publishing, 34–48. https://doi.org/10.1007/978-3-030-39540-7_3

[9] Youjeong Kim and S. Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior* 28, 1 (January 2012), 241–250. https://doi.org/10.1016/j.chb.2011.09.006

[10] Daniel Kimmel. 2023. ChatGPT Therapy Is Good, But It Misses What Makes Us Human. https://www.columbiapsychiatry.org/news/chatgpt-therapy-is-good-but-it-misses-what-makes-us-human. Accessed: 09/11/2023.

[11] Andrew Leonard. 2023. 'Dr. Google' meets its match: Dr. ChatGPT. https://www.latimes.com/science/story/2023-09-08/dr-google-meets-its-match-dr-chatgpt Accessed: 09/11/2023.

[12] Tianshi Li, Yuvraj Agarwal, and Jason I. Hong. 2018. Coconut: An IDE Plugin for Developing Privacy-Friendly Apps: An IDE Plugin for Developing Privacy-Friendly Apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (December 2018), 1–35. https://doi.org/10.1145/3287056

[13] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. *arXiv preprint arXiv:2310.17884* (2023).

[14] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035* (2023).

[15] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. 2022. Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE. https://doi.org/10.1109/sp46214.2022.9833571

[16] Charith Peris, Christophe Dupuy, Jimit Majmudar, Rahil Parikh, Sami Smaili, Richard Zemel, and Rahul Gupta. 2023. Privacy in the Time of Language Models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23)*. ACM. https://doi.org/10.1145/3539597.3575792

[17] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298* (2023).

[18] Mikhail Taver. 2023. ChatGPT is Coming to Finance, So Let's Talk About the Risks and Rewards. https://www.unite.ai/chatgpt-is-coming-to-finance-so-lets-talk-about-the-risks-and-rewards/. Accessed: 09/11/2023.

[19] Jackie (Junrui) Yang, Yingtian Shi, Yuhan Zhang, Karina Li, Daniel Wan Rosli, Anisha Jain, Shuning Zhang, Tianshi Li, James A. Landay, and Monica Lam. 2023. ReactGenie: An Object-Oriented State Abstraction for Complex Multimodal Interactions Using Large Language Models. *arXiv preprint arXiv:2306.09649* (2023). https://doi.org/10.48550/arXiv.2306.09649

[20] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2021. Counterfactual memorization in neural language models. *arXiv preprint arXiv:2112.12938* (2021).

[21] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3642385