Privacy, Copyright and Data Integrity: The Cascading Implications of Generative AI

My research **identifies and addresses emerging challenges in data use for AI**. I explore the **interplay between data, its influence on models, and the expectations of the people** who regulate and use these models. Questions such as 'What if these systems reproduce someone's address, copyrighted text, or harmful content?' are now central in regulatory discussions [1]. I address these challenges by limiting undesirable exposure of data through novel algorithms and grounding these algorithms in legal and social frameworks. I posit that **data protection is not clear-cut**, especially for generative AI: though a model should not produce an individual's address, it should be capable of producing the address to the nearest hospital [2]. Specifically, my research explores:

- ***Uncovering mechanisms of data memorization and exposure.*** My work is among the first to successfully mount novel membership inference attacks (MIAs, see below) on large language models (LLMs) [3, 4], sparking further research [5] and uncovering the underlying mechanisms of data memorization [6]. I demonstrate how training content can be memorized and reproduced in forms and contexts different from its original appearance [7, 8], revealing new potential attack vectors.
- ***Mitigating data exposure algorithmically.*** I propose algorithmic approaches to mitigate privacy risks throughout the LLM supply chain[9–13]. Notably, I introduced a differentially private data synthesis method that generates user-like sequences [11] along with empirical, information -theoretic methods [14–16] that protect the privacy of queries at inference time. These innovations have had direct industry impact[1] and won the 2020 NCWIT award.
- ***Grounding algorithms in legal and social frameworks.*** I introduce modeling paradigms and evaluation benchmarks tailored to existing legal frameworks and human needs [8, 18–21]. Through this work, I developed the first privacy benchmark based on contextual integrity [19], designed the first copyright risk assessment for non-literal copying [8], and analyzed large-scale, real-world human chatbot interactions to reveal concerning patterns of personal information disclosure [22].

Emerging technology provides tools to help build the 'best-performing' models; however, if we cannot convey or even evaluate their capabilities and limitations in a digestible way, they will not be useful to individuals and society. With the experiences I have gained in academia, industry, and working with policymakers, I am well-equipped to bridge this gap.

1 Uncovering Mechanisms of Data Memorization and Exposure

The question 'Did you use my data to train your model?' is dominating generative AI and policy discourse as artists, writers, and institutions try to determine if their data has contributed to a model for copyright and attribution purposes. However, this question has been prevalent in the computer privacy domain for many years. Membership inference attacks (MIAs) were introduced as tools to address this issue. These attacks work by solving a binary classification problem for each target data point, determining whether it was a member of the training dataset or not. My work pioneered this research direction for LLMs. I have (1) adapted existing MIAs to LLMs [3], (2) introduced novel membership inference and extraction attacks specific to LLMs [4, 7], and (3) evaluated and analyzed the limits and capabilities of MIAs for LLMs [5, 6]. To advance this research direction, I presented my work by invitation at numerous venues across disciplines, including the C3E workshop[2] (by the SRI and the NSA), the NDSS 2023 EthiCS workshop and the NeurIPS 2024 red-teaming workshop. My research has also been covered by *WIRED* and *The Washington Post* articles.[3]

Membership inference attacks for LLMs. Most membership inference attacks threshold a score $f(x; M)$ for the target point x and a target model M to determine membership, as shown in Figure 1.

[1] A startup was founded four years ago based on the patent [17] relating to this direction:`https://protopia.ai`.
[2] Computational Cybersecurity in Compromised Environments (C3E) Workshop `https://cps-vo.org/group/c3e`
[3] WIRED – How to Stop Your Data From Being Used to Train AI and Washington Post – How to opt out

At its simplest, f can be the loss of x under M. I demonstrated that this simple heuristic yields high false negatives in LLMs since it ignores sample complexity. I proposed a stronger attack that uses likelihood ratios with a reference model to calibrate loss and account for sample complexity [3], as shown by $f_{Reference}$ in the Figure. Follow-up work eliminated the need for a reference model by proposing a new $f_{Neighborhood}$ based on loss function local optimality around x [4]; our intuition was that models have higher loss curvature around training data. We showed that this new 'neighborhood-based' attack outperforms the reference-based one, placing it among state-of-the-art MIAs for LLMs.

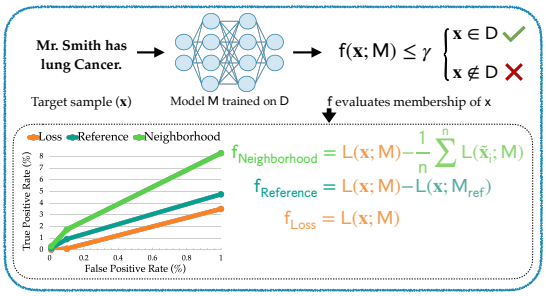

Figure 1: Overview of how membership inference attacks work on LLMs. I propose new heuristics for membership score function f based on a reference model's loss [3] or local optimality of loss [4].

Evaluating and analyzing membership inference attacks. In 2023, with open-data models emerging, my collaborators and I conducted a comprehensive study on MIAs for pre-training data in language models [5, 7]. Surprisingly, we found that **all MIAs consistently demonstrate near-random performance** on these models. We attributed this to (1) single-epoch exposure of training data and (2) inherent n-gram overlap between members and non-members. Our findings prompted several studies from other labs, revealing other limitations in existing benchmarks [23, 24]. We released our code and dataset as a package, and it is now the most starred LLM-MIA repository on GitHub,[4] with over 38K dataset downloads as of September 2024.

2 Mitigating Data Exposure Algorithmically

My research identifies risks that occur during different stages of generative AI deployment and addresses them, either through differentially private methods that assume worst-case scenarios [9–11] or through task-specific, information-theoretic techniques [12, 14, 15]. I have been invited to talk about differential privacy (DP) at the GenLaw workshop in DC and have co-instructed a tutorial on privacy mitigations at EACL 2023. I have also co-organized numerous workshops on this topic since 2021.[5]

Protecting training data via training-time interventions. I developed regularization terms that alter training to reduce authorship attribution from a model's hidden states (using adversarial classifiers and triplet loss), decreasing memorization and improving the disparate impact that privacy mitigations have been shown to have [12]. For scenarios with unknown downstream tasks or threats, differential privacy can provide worst-case guarantees. I proposed the first differentially private compression algorithms for LLMs [10], demonstrating that for private data, pruning outperforms distillation in model performance under the same privacy budget, contrary to some non-private settings.

Protecting training data via private data synthesis. Training-time interventions are effective only when we know user needs, which is not always the case. How can we improve models based on user interactions without compromising privacy? To address this, I led a project at Microsoft that introduced a new data synthesis algorithm that generates data practitioners can directly examine, while still providing DP guarantees [11]. Our key innovation was a two-stage approach using semantic parse trees as latent variables, which preserves the distribution more accurately. In follow-up work, we proposed synthesizing differentially private data for in-context learning examples using differentially private decoding instead of fine-tuning [9], addressing challenges in creating small, task-specific datasets. Our approach, implemented by LlamaIndex,[6] enables developers to create private datasets for use in various LLM applications.

[4] `https://github.com/iamgroot42/mimir`
[5] At ICLR 2021, ACL 2022, NAACL 2022, AAAI 2024, and ICLR 2024.
[6] `https://llamahub.ai/l/llama-packs/llama-index-packs-diff-private-simple-dataset`

Protecting user queries at inference time. As cloud-based execution of generative AI becomes prevalent, protecting the privacy of inference queries has emerged as a critical challenge. My research pioneered solutions to address these concerns [14–16]. My first approach, Shredder, optimally splits the neural network between edge and cloud, then adds carefully crafted noise to intermediate activations sent to the cloud; this approach reduces information leakage while maintaining accuracy. It won the NCWIT award in 2020 and is one of the earliest works in the field of split learning. Building on Shredder, I developed Cloak, which improves on its predecessor by learning the noise distribution's standard deviation directly through backpropagation. This technique preserves only the most relevant features in user inputs (Figure 2). These works have been patented and form the foundation for an ongoing startup [17].

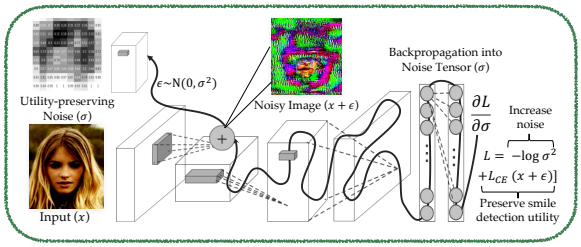

Figure 2: Learning utility-aware noise distributions to protect queries at inference time [14, 15].

3 Grounding Algorithms in Legal and Social Frameworks

I maintain that LLM researchers should ground algorithms and benchmarks in existing policy, legal, and social frameworks to facilitate their adoption. In this vein, I have led multiple projects [9, 18, 19], collaborated with lawyers [21] and co-organized workshops on the intersection of Generative AI and Law [25].[7]

Evaluating privacy in LLMs based on contextual integrity theory. To evaluate emergent privacy risks as language models become integral to user-facing technologies, I proposed ConfAIde [19], a novel benchmark based on Helen Nissenbaum's theory of contextual integrity. This multi-tiered framework assesses models' ability to make context-appropriate decisions about information flow, considering factors such as sender, receiver, and intended use (see Figure 3). Our evaluation revealed that even advanced commercial models like ChatGPT (GPT-turbo-3.5) fail to discern appropriate information disclosure 92% of the time. Of these failures, 30% were attributed to the model's inability to distinguish the user's intentions or perspective, often referred to as "theory of mind", or to misguided attempts at helpfulness. As the first adaptation of contextual integrity to LLMs, this work has already sparked multiple follow-ups from industry and academia [26, 27].

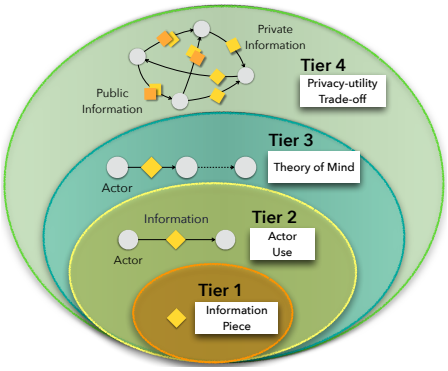

Figure 3: Different tiers of our LLM privacy benchmark [19].

Evaluating non-literal memorization based on copyright law. In response to the growing concern over AI models reproducing training data that results in copyright lawsuits,[8] I proposed a study on non-literal copying – the reproduction of event sequences or characteristics without exact text matching – drawing inspiration from non-AI copyright litigation [28, 29]. My collaborators and I showed that while instruction-tuned, aligned models generally exhibit less literal regurgitation compared to their base counterparts, they occasionally demonstrate higher non-literal regurgitation [8]. This insight highlights potential risks in these models and underscores the complexity of copyright issues in AI, paving the way for deeper studies of semantic memorization.

4 Future Directions

Technology continues to advance at a much faster pace than the tools and guidelines that safeguard it. Building 'safe' technology is not instantaneous; it is gradual. To make progress, we must proactively

[7] GenLaw@ICML 2023 and GenLaw@ICML 2024

[8] https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

uncover and control the risks of emergent architectures while deepening our formal understanding of how these systems learn. I aim to expand my research agenda and bridge disciplinary gaps by focusing on the areas listed below.

Uncovering emergent risks. With emergent capabilities of language models come emergent risks. For instance, while in-context learning helps models generalize to new tasks, it opens vulnerabilities like leakage of private examples, as our work shows [9, 19]. As paradigms like inference-as-a-service, retrieval-augmented generation, and agentic AI proliferate, we face potential new attack vectors. These execution modes require systematic evaluation through dynamic, evolving benchmarks. My experience in red-teaming models [30] and building robust benchmarks [5, 8, 19] positions me to lead this work. I plan to expand on my prior work using one agent to uncover vulnerabilities in another [7], drawing from fuzzing methods in computer security to proactively identify new risks.

Comprehensively modeling memorization beyond text and transformers. I aim to develop robust, semantic-based notions of memorization, formalizing data leakage beyond rigid verbatim definitions. As AI systems become multimodal and multilingual, we must understand how different modalities (text, images, code, speech) interact and potentially mask or amplify each other, how concepts and meanings transfer across forms, and how new training paradigms like reinforcement learning with human feedback affect cross-modal data dynamics. Our work has shown that instruction tuning can cause unexpected data reconstruction and leakage across languages [7] while decreasing output diversity [31]. By studying these phenomena, I aim to build a comprehensive model of learning to predict and mitigate undesired content generation while deepening our understanding of model internals.

Controlling LLMs for societal impact. I aim to develop both training-time guarantees and inference-time control methods for LLM safeguards. At training-time, I want to develop differentially private synthesis methods [3, 25] to enable secure data sharing in domains like healthcare and unlock scientific discovery. At inference time, I plan to create dynamic, easily updatable fine-grained controls that give users local control over their data [14]. Building on my work in contextual integrity [19] and energy-based controllable decoding [32, 33], I aim to develop symbolic decoding methods based on belief tracking for nuanced, context-sensitive content generation. I plan to expand LLMs' impact in policy and professional domains through controlled generation with source attribution for copyright and algorithmic disgorgement cases [34], while enabling private data personalization, building on my earlier work [18, 35]. These defenses, as part of a comprehensive defense-in-depth approach, aim to mitigate potential harms, acknowledging that while not 100% effective in isolation, they contribute significantly to overall system security.

References

[1] Haleluya Hadero and David Bauder. New York Times sues Microsoft, Open AI over use of content. *Globe & Mail (Toronto, Canada)*, pages B1–B1, 2023.

[2] Hannah Brown, Katherine Lee, Niloofar Mireshghallah, R. Shokri, and Florian Tram'er. What Does it Mean for a Language Model to Preserve Privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, June 2022.

[3] Niloofar Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, December 2022.

[4] Justus Mattern, Niloofar Mireshghallah, Zhijing Jin, Bernhard Scholkop, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership Inference Attacks against Language Models via Neighbourhood Comparison. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL Findings)*, Jul 2023.

[5] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do Membership Inference

Attacks Work on Large Language Models? In *The First Conference on Language Modeling (COLM)*, October 2024.

[6] Niloofar Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. Memorization in NLP Fine-tuning Methods. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Oral Presentation*, December 2022.

[7] Niloofar Mireshghallah, Aly M Kassem, Omar Mahmoud, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi, Sherif Saad, and Santu Rana. Alpaca against Vicuna: Using LLMs to Uncover Memorization of LLMs. *arXiv preprint arXiv:2403.04801*, march 2024.

[8] Tong Chen, Niloofar Mireshghallah, et al. CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation. *arXiv:2407.07087*, July 2024.

[9] Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Niloofar Mireshghallah, et al. Privacy-Preserving In-Context Learning with Differentially Private Few-Shot Generation. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[10] Niloofar Mireshghallah, Arturs Backurs, Huseyin A Inan, Lukas Wutschitz, and Janardhan Kulkarni. Differentially Private Model Compression. *Advances in Neural Information Processing Systems (NeurIPS)*, Dec 2022.

[11] Niloofar Mireshghallah, Richard Shin, Yu Su, Tatsunori Hashimoto, and Jason Eisner. Privacy-Preserving Domain Adaptation of Semantic Parsers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL, Volume 1: Long Papers)*, Jul 2023.

[12] Niloofar Mireshghallah, Huseyin A. Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. Privacy Regularization: Joint Privacy-Utility Optimization in Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, June 2021.

[13] Niloofar Mireshghallah and Taylor Berg-Kirkpatrick. Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP),* ***Oral Presentation***, November 2021.

[14] Niloofar Mireshghallah, Mohammadkazem Taram, Ali Jalali, Dean Tullsen, and Hadi Esmaeilzadeh. Shredder: Learning Noise Distributions to Protect Inference Privacy. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, March 2020.

[15] Niloofar Mireshghallah, Mohammadkazem Taram, Ali Jalali, Ahmed Taha Elthakeb, Dean Tullsen, and Hadi Esmaeilzadeh. Not All Features Are Equal: Discovering Essential Features for Preserving Prediction Privacy. In *Proceedings of The Web Conference 2021 (WWW)*, April 2021.

[16] Teddy Koker, Niloofar Mireshghallah, Tom Titcombe, and Georgios Kaissis. U-Noise: Learnable Noise Masks for Interpretable Image Segmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, September 2021.

[17] Niloofar Mireshghallah and Hadi Esmaeilzadeh. METHODS OF PROVIDING DATA PRIVACY FOR NEURAL NETWORK BASED INFERENCE. March 2020. US Patent 11,288,379.

[18] Niloofar Mireshghallah, Nikolai Vogler, Junxian He, Omar Florez, Ahmed El-Kishky, and Taylor Berg-Kirkpatrick. Non-Parametric Temporal Adaptation for Social Media Topic Classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, December 2023.

[19] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR* ***Spotlight****)*, 2024.

[20] Niloofar Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. Smaller Language Models are Better Black-box Machine-Generated Text Detectors. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2024.

[21] Katherine Lee, A Feder Cooper, Christopher A Choquette-Choo, Ken Liu, Matthew Jagielski, Niloofar Mireshghallah, Lama Ahmed, James Grimmelmann, David Bau, Christopher De Sa, et al. Machine Unlearning Doesn't Do What You Think. July 2024.

[22] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. In *The First Conference on Language Modeling (COLM)*, October 2024.

[23] Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. LLM Dataset Inference: Did you train on my dataset? *arXiv preprint arXiv:2406.06443*, 2024.

[24] Debeshee Das, Jie Zhang, and Florian Tram'er. Blind baselines beat membership inference attacks for foundation models. *arXiv preprint arXiv:2406.16201*, 2024.

[25] A Feder Cooper, Katherine Lee, James Grimmelmann, Daphne Ippolito, Christopher Callison-Burch, Christopher A Choquette-Choo, Niloofar Mireshghallah, et al. Report of the 1st Workshop on Generative AI and Law. *arXiv:2311.06477*, 2023.

[26] Sahra Ghalebikesabi, Eugene Bagdasaryan, Ren Yi, Itay Yona, Ilia Shumailov, Aneesh Pappu, Chongyang Shi, Laura Weidinger, Robert Stanforth, Leonard Berrada, et al. Operationalizing Contextual Integrity in Privacy-Conscious Assistants. *arXiv preprint arXiv:2408.02373*, 2024.

[27] Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action. *Advances in Neural Information Processing Systems (NeurIPS)*, Dec 2024.

[28] Stephen Rebikoff. Restructuring the test for copyright infringement in relation to literary and dramatic plots. *Melbourne University Law Review*, 25(2):340–373, 2001.

[29] Pamela Samuelson. A fresh look at tests for nonliteral copyright infringement. *Nw. UL Rev.*, 107:1821, 2012.

[30] Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, et al. WildTeaming at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer Language Models. *arXiv:2406.18510*, preprint.

[31] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A Roadmap to Pluralistic Alignment. In *The Fourty-first International Conference on Machine Learning (ICML)*, July 2024.

[32] Niloofar Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. Mix and Match: Learning-free Controllable Text Generation using Energy Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (ACL, Volume 1: Long Papers)*, May 2022.

[33] Jarad Forristal, Niloofar Mireshghallah, Greg Durrett, and Taylor Berg-Kirkpatrick. A Block Metropolis-Hastings Sampler for Controllable Energy-based Text Generation. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 403–413, 2023.

[34] Alessandro Achille, Michael Kearns, Carson Klingenberg, and Stefano Soatto. AI model disgorgement: Methods and choices. *Proceedings of the National Academy of Sciences*, 121(18):e2307304121, 2024.

[35] Niloofar Mireshghallah, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim, and Dimitrios Dimitriadis. UserIdentifier: Implicit User Representations for Simple and Effective Personalized Sentiment Analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, July 2022.