Niloofar Mireshghallah

github.io/~mireshghallah niloofar@cmu.edu +1 (619) 888-9954

POSITIONS

Assistant Professor (Incoming)
Carnegie Mellon University *Aug 2025–Present*
Joint appointment: Language Technologies Institute (LTI) & Engineering and Public Policy (EPP), Core member of CyLab

Research Scientist
Meta AI (FAIR) *May 2025–Nov 2025*
Privacy-preserving AI systems and LLM safety

Postdoctoral Scholar
University of Washington *May 2023–May 2025*
Advisors: Yejin Choi, Yulia Tsvetkov

Research Intern
Microsoft (Semantic Machines; Microsoft Research) *Jun 2021–Jul 2023*
Privacy-preserving synthetic user data generation.
Differentially private model compression

EDUCATION

Ph.D. & M.S. in Computer Science
University of California San Diego *Sep 2018–Apr 2023*
Ph.D. Advisor: Taylor Berg-Kirkpatrick
M.S. awarded Jun 2020
CGPA: 3.90/4.00

B.Sc. in Computer Engineering
Sharif University of Technology, Iran *Sep 2014–Jun 2018*
CGPA: 18.12/20.00

RESEARCH FOCUS

Privacy-preserving AI systems, LLM Policy and Ethics, Contextual Integrity in AI, AI for Science and Health.

AWARDS

Tinker Academic Research Compute Grant	*2025*
Modal Academic Research Compute Grant	*2025*
Momental Foundation Mistletoe Research Fellowship (MRF) Finalist	*2023*
Rising Star in Adversarial Machine Learning (AdvML) Award Winner	*2022*
Rising Stars in EECS	*2022*
UCSD CSE Excellence in Leadership and Service Award Winner	*2022*
FAccT Doctoral Consortium	*2022*

Qualcomm Innovation Fellowship Finalist *2021*
NCWIT (National Center for Women & IT) Collegiate Award Winner *2020*
National University Entrance Exam in Math (Ranked 249^{th} of $223,000$) *2014*
National University Entrance Exam in Foreign Languages (Ranked 57^{th} of $119,000$) *2014*
National Organization for Exceptional Talents (NODET) (Admitted, ∼2% Acceptance Rate) *2008*

SELECTED PUBLICATIONS

Policy Write-ups

1. A. F. Cooper, …, **N. Mireshghallah**, …, K. Lee, "**Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice**," *Neural Information Processing Systems (NeurIPS) Position Paper* 2025 **(Oral)**.
2. A. F. Cooper, …, **N. Mireshghallah**, …, E. Zeide, "Report of the 1st Workshop on Generative AI and Law," *Yale Law & Economics Research Paper, Available at SSRN-4634513* (2023).

Patents

1. J. M. Eisner, E. C. Shin, **N. Mireshghallah**, T. B. Hashimoto, and Y. Su, "Privacy-Preserving Generation of Synthesized Training Data," *US Patent Application 18/321,460* (2024).
2. **N. Mireshghallah** and H. Esmaeilzadeh, "Methods of Providing Data Privacy for Neural Network-based Inference," *US Patent 11,487,884*.

Conference Publications

1. **N. Mireshghallah***, …, P. W. Koh, "A False Sense of Privacy: Evaluating Textual Data Sanitization Beyond Surface-Level Privacy Leakage," *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) 2026*.

2. X. Lu, …, **N. Mireshghallah**, …, Y. Choi, "AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text," *International Conference on Learning Representations (ICLR) 2025* **(Oral Presentation)**.

3. X. Zhou, …, **N. Mireshghallah**, …, M. Sap, "HAICOSYSTEM: An Ecosystem for Sandboxing Safety Risks in Human–AI Interactions," *Conference on Language Models (COLM) 2025*.

4. I. C. Ngong, J. P. Near, and **N. Mireshghallah**, "Differentially Private Learning Needs Better Model Initialization and Self-Distillation," *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) 2025* **(Oral Presentation)**.

5. Y. Bae*, …, **N. Mireshghallah**, "PPMI: Privacy-Preserving LLM Interaction with Socratic Chain-of-Thought Reasoning and Homomorphically Encrypted Vector Databases," *Workshop on Privacy-Preserving Machine Learning at the International Cryptology Conference (CRYPTO) 2025*.

6. J. Hayes, …, **N. Mireshghallah**, …, A. F. Cooper, "Strong Membership Inference

Attacks on Massive Datasets and (Moderately) Large Language Models," *Conference on Neural Information Processing Systems (NeurIPS) 2025*.

7. A. Ravichander, …, **N. Mireshghallah**, …, Y. Choi, "Information-Guided Identification of Training Data Imprint in (Proprietary) Large Language Models," *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) 2025* **(Honorable Mention Candidate, Oral Presentation)**.

8. L. Jiang, …, **N. Mireshghallah**, …, N. Dziri, "WildTeaming at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer Language Models," *Conference on Neural Information Processing Systems (NeurIPS) 2024*.

9. M. Duan, …, **N. Mireshghallah**, …, H. Hajishirzi, "Do Membership Inference Attacks Work on Large Language Models?" *Conference on Language Models (COLM) 2024*.

10. **N. Mireshghallah**, M. Antoniak, Y. More, Y. Choi, and G. Farnadi, "Trust No Bot: Discovering Personal Disclosures in Human–LLM Conversations in the Wild," *Conference on Language Models (COLM) 2024*.

11. T. Sorensen, …, **N. Mireshghallah**, …, Y. Choi, "A Roadmap to Pluralistic Alignment," *International Conference on Machine Learning (ICML) 2024*.

12. **N. Mireshghallah**, …, Y. Choi, "Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory," *International Conference on Learning Representations (ICLR) 2024* **(Spotlight Presentation)**.

13. **N. Mireshghallah**, R. Shin, Y. Su, T. Hashimoto, and J. Eisner, "Privacy-Preserving Domain Adaptation of Semantic Parsers," *Annual Meeting of the Association for Computational Linguistics (ACL) 2023*.

14. J. Mattern, …, **N. Mireshghallah**, …, T. Berg-Kirkpatrick, "Membership Inference Attacks against Language Models via Neighbourhood Comparison," *Findings of the Association for Computational Linguistics (ACL) 2023*.

15. **N. Mireshghallah**, A. Uniyal, T. Wang, D. Evans, and T. Berg-Kirkpatrick, "Memorization in NLP Fine-Tuning Methods," *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2022* **(Oral Presentation)**.

16. **N. Mireshghallah**, K. Goyal, A. Uniyal, T. Berg-Kirkpatrick, and R. Shokri, "Quantifying Privacy Risks of Masked Language Models using Membership Inference Attacks," *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2022*.

17. **N. Mireshghallah**, A. Backurs, H. A. Inan, L. Wutschitz, and J. Kulkarni, "Differentially Private Model Compression," *Conference on Neural Information Processing Systems (NeurIPS) 2022*.

18. H. Brown, K. Lee, **N. Mireshghallah**, R. Shokri, and F. Tramèr, "What Does it

Mean for a Language Model to Preserve Privacy?," *ACM Conference on Fairness, Accountability, and Transparency (FAccT) 2022*.

19. **N. Mireshghallah** and T. Berg-Kirkpatrick, "Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness," *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2021* **(Oral Presentation)**.

20. T. Koker, **N. Mireshghallah**, T. Titcombe, and G. Kaissis, "U-Noise: Learnable Noise Masks for Interpretable Image Segmentation," *IEEE International Conference on Image Processing (ICIP) 2021*.

21. **N. Mireshghallah**, ..., R. Sim, "Privacy Regularization: Joint Privacy–Utility Optimization in Language Models," *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) 2021*.

22. **N. Mireshghallah**, ..., H. Esmaeilzadeh, "Not All Features Are Equal: Discovering Essential Features for Preserving Prediction Privacy," *ACM Web Conference (WWW) 2021*.

23. **N. Mireshghallah**, M. Taram, A. Jalali, D. Tullsen, and H. Esmaeilzadeh, "Shredder: Learning Noise Distributions to Protect Inference Privacy," *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2020*.

SELECTED INVITED TALKS

Recent Talks

FAR AI San Diego Alignment Workshop at NeurIPS
What Does It Mean for Agentic AI to Preserve Privacy?, Dec. 2025

Conference on Applied Machine Learning in Information Security (CAMLIS)
Keynote: *What Does It Mean for Agentic AI to Preserve Privacy? Mapping the New Data Sinks and Leaks*, Oct. 2025

Cornell Tech, Digital Life Seminar
Contextual Privacy in LLMs: Benchmarking and Mitigating Inference-Time Risks, Oct. 2025

First Workshop on LLM Security (LLMSec) at ACL 2025
Keynote: *What Does It Mean for Agentic AI to Preserve Privacy?*, Aug. 2025

First Workshop on Large Language Model Memorization (L2M2) at ACL 2025
Keynote:*Emergent Misalignment Through the Lens of Semantic Memorization*, Aug. 2025

Workshop on Collaborative and Federated Agentic Workflows (CFAgentic) at ICML
What Does It Mean for Agentic AI to Preserve Privacy?, July 2025

Stanford University (NLP Seminar)
Privacy, Copyright and Data Integrity: The Cascading Implications of Generative AI, Jan. 2025

Academic Invited Job Talks

Jan.–Mar. 2025

Privacy, Copyright and Data Integrity: The Cascading Implications of Generative AI
Johns Hopkins University, UC Berkeley, Carnegie Mellon University, UW Madison, UPenn, Georgia Tech, UCLA, University of Maryland, University of Michigan, NYU, École Polytechnique Fédérale de Lausanne (EPFL), ETH Zurich, UT Austin, UVA, UNC Chapel Hill

Earlier Talks

Google Brain (2024), Meta AI Research (2024), University of Washington Allen School Colloquium (2024), CISPA Helmholtz Center for Information Security (2023), Max Planck Institute for Software Systems (2023)

DIVERSITY & INCLUSION

Women in Machine Learning (WiML) Workshop at NeurIPS Mentor *2025*
CMU School of Computer Science Panel: Navigating the Academic Job Market *2025*
ACL Mentorship: How to Broadcast Your Research to a Wider Audience? *2025*
NAACL 2025 D&I Co-chair *2025*
Women in ML (WiML) at NeurIPS Mentor *2024*
Widening NLP (WiNLP) Co-chair *2022–2024*
NAACL 2022 D&I Co-chair *2022*
Mentor at ICLR *2021*
Mentor for the Women in Machine Learning (WiML) Workshop at NeurIPS *2020*
Mentor for the Graduate Women in Computing (GradWIC) at UCSD *2020–2023*
Course Instructor for the OpenMined Privacy Course *2020*
Mentor for the USENIX Security Undergraduate Mentorship Program *2020*
Volunteer at the Women in Machine Learning Workshop Held at NeurIPS *2019*

ORGANIZED EVENTS

Co-organizer of The Memorization and Trustworthy Foundation Models Workshop at ICML *2025*
Panelist at Workshop on Collaborative and Federated Agentic Workflows at ICML *2025*
Panelist at Workshop on Technical AI Governance (TAIG) at ICML *2025*
Privacy Session Chair at SAGAI Workshop at IEEE S&P *2025*
Co-organizer of the Generative AI and Law (GenLaw) Workshop at ICML *2024*
Co-organizer of the Privacy Regulation and Protection in Machine Learning Workshop *2024*
Co-organizer of the Private NLP Workshop at ACL *2024*
Co-organizer of the Privacy-Preserving AI (PPAI) Workshop at AAAI *2024*
Co-organizer of the Generative AI and Law (GenLaw) Workshop at ICML *2023*
Co-organizer of the Widening NLP (WiNLP) Workshop at EMNLP *2023*
Co-organizer of the Private NLP Tutorial at EACL *2023*
Co-organizer of the Ethics in NLP Birds of a Feather Session at EMNLP *2022*
Co-organizer of the Broadening Collaborations in ML Workshop at NeurIPS *2022*
Co-organizer of the Widening NLP (WiNLP) Workshop at EMNLP *2022*
Co-organizer of the Private NLP Workshop at NAACL *2022*
Co-organizer of the Federated Learning for NLP Workshop at ACL *2022*
Co-organizer of the Privacy-Preserving Machine Learning (PPML) Workshop at MICCAI *2021*