

# Nilofar Miresghallah

🌐 [github.io/~miresghallah](https://github.io/~miresghallah)    ✉ [nilofar@cmu.edu](mailto:nilofar@cmu.edu)    ☎ +1 (619) 888-9954  
🔍 Google Scholar    🌐 GitHub    🐦 @nilofar\_mire    in LinkedIn

## RESEARCH FOCUS

---

Privacy-preserving AI systems and contextual integrity in LLMs; LLM policy, ethics, and societal implications of ML; benchmarking and evaluation of language models; memorization, membership inference, and data integrity; LLM reasoning and post-training; AI for science and health.

## POSITIONS

---

**Founding Member of Technical Staff**, *humans&* Nov 2025–Present

**Assistant Professor**, *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 Alignment: Privacy & Security; collab. w/ FAIR Chemistry)* May–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 / Part-time Researcher**, *Microsoft (Semantic Machines; Microsoft Research)* Jun 2021–Jul 2023  
Privacy-preserving synthetic user data generation, differentially private model compression, semantic parsing. Mentors: Richard Shin, Yu Su, Tatsunori Hashimoto, Jason Eisner, Sergey Yekhanin, Arturs Backurs, Dimitrios Dimitriadis, Robert Sim.

**Research Intern**, *Western Digital Co. Research and Development* Summer 2019  
Mentor: Anand Kulkarni.

## EDUCATION

---

**Ph.D. & M.S., 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., Computer Engineering**, *Sharif University of Technology, Iran* Sep 2014–Jun 2018  
CGPA: 18.12/20.00.

## AWARDS, GRANTS & HONORS

---

Delta Institute Scientific Fellow	2026
OpenAI Mental Health Research Grant	2026
Prime Intellect Academic Research Compute Grant	2026
Greenwall Foundation Faculty Scholars Program in Bioethics, Finalist	2026
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 <sup>th</sup> of 223,000)	2014
National University Entrance Exam in Foreign Languages (Ranked 57 <sup>th</sup> of 119,000)	2014
National Organization for Exceptional Talents (NODET), Admitted (~2% Acceptance Rate)	2008

## PUBLICATIONS

---

\* denotes equal contribution. For the full list, see Google Scholar.

## Preprints & Under Review

- A. Hughes, A. Goldberg, P. Jha, A. Perer, N. Aletras, **N. Mireshghallah**, “Boundary-targeted Membership Inference Attacks on Safety Classifiers,” *arXiv:2605.22373*, 2026.
- K. Han, R. Zhang, K. Wei, H. Mahdavi, **N. Mireshghallah**, A. Farimani, “SMDD-Bench: Can LLMs Solve Real-World Small Molecule Drug Design Tasks?” *arXiv:2605.21740*, 2026.
- K. Monteiro, M. Park, A. Ioffrida, A. Sanna, **N. Mireshghallah**, Y. Wang, S. Das, “When Are LLM Inferences Acceptable? User Reactions and Control Preferences for Inferred Personal Information,” *arXiv:2605.10013*, 2026.
- X. Liu, **N. Mireshghallah**, J. C. Ginsburg, T. Chakrabarty, “Alignment Whack-a-Mole: Finetuning Activates Verbatim Recall of Copyrighted Books in Large Language Models,” *arXiv:2603.20957*, 2026.
- **N. Mireshghallah** and T. Li, “Privacy Is Not Just Memorization,” Technical Report, 2025.

## Conference Papers

- H. Kim\*, **N. Mireshghallah\***, M. Duan, R. Xin, S. S. Li, J. Jung, D. Acuna, Q. Pang, H. Xiao, G. E. Suh, S. Oh, Y. Tsvetkov, P. W. Koh, Y. Choi, “Privasis: Synthesizing the Largest ‘Public’ Private Dataset from Scratch,” *ICML 2026*.
- **N. Mireshghallah**, N. Mangaokar, N. Kokhlikyan, A. Zharmagambetov, M. Zaheer, S. Mahloujifar, K. Chaudhuri, “CIMemories: A Compositional Benchmark for Contextual Integrity of Persistent Memory in LLMs,” *ICLR 2026*.
- J. Zhou, **N. Mireshghallah**, T. Li, “Operationalizing Data Minimization for Privacy-Preserving LLM Prompting,” *ICLR 2026*.
- T. Sorensen, B. Newman, J. Moore, C. Y. Park, J. Fisher, **N. Mireshghallah**, L. Jiang, Y. Choi, “Spectrum Tuning: Post-Training for Distributional Coverage and In-Context Steerability,” *ICLR 2026*.
- R. Xin\*, **N. Mireshghallah\***, S. S. Li, M. Duan, H. Kim, Y. Choi, Y. Tsvetkov, S. Oh, P. W. Koh, “A False Sense of Privacy: Evaluating Textual Data Sanitization Beyond Surface-Level Privacy Leakage,” *SaTML 2026*.
- A. F. Cooper, ..., **N. Mireshghallah**, ..., K. Lee, “Machine Unlearning Doesn’t Do What You Think: Lessons for Generative AI Policy, Research, and Practice,” *NeurIPS 2025 (Oral)*.
- J. Hayes, ..., **N. Mireshghallah**, ..., A. F. Cooper, “Strong Membership Inference Attacks on Massive Datasets and (Moderately) Large Language Models,” *NeurIPS 2025*.
- S. Hallinan, J. Jung, M. Sclar, X. Lu, A. Ravichander, S. Ramnath, Y. Choi, S. P. Karimireddy, **N. Mireshghallah**, X. Ren, “The Surprising Effectiveness of Membership Inference with Simple N-Gram Coverage,” *COLM 2025*.
- X. Zhou, ..., **N. Mireshghallah**, ..., M. Sap, “HAICOSYSTEM: An Ecosystem for Sandboxing Safety Risks in Human–AI Interactions,” *COLM 2025*.
- X. Lu, M. Sclar, S. Hallinan, **N. Mireshghallah**, J. Liu, S. Han, A. Ettinger, L. Jiang, K. Chandu, N. Dziri, Y. Choi, “AI as Humanity’s Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text,” *ICLR 2025 (Oral)*.
- I. C. Ngong, J. P. Near, **N. Mireshghallah**, “Differentially Private Learning Needs Better Model Initialization and Self-Distillation,” *NAACL 2025 (Oral)*.
- A. Ravichander, J. Fisher, T. Sorensen, X. Lu, M. Antoniak, B. Y. Lin, **N. Mireshghallah**, C. Bhagavatula, Y. Choi, “Information-Guided Identification of Training Data Imprint in (Proprietary) Large Language Models,” *NAACL 2025 (Honorable Mention Candidate, Oral)*.
- A. Kassem\*, O. Mahmoud\*, **N. Mireshghallah\***, H. Kim, Y. Tsvetkov, Y. Choi, S. Saad, S. Rana, “Alpaca against Vicuna: Using LLMs to Uncover Memorization of LLMs,” *NAACL 2025*.
- L. Jiang, K. Rao, S. Han, A. Ettinger, F. Brahman, S. Kumar, **N. Mireshghallah**, X. Lu, M. Sap, Y. Choi, N. Dziri, “WildTeaming at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer Language Models,” *NeurIPS 2024*.
- T. Chen, **N. Mireshghallah\***, A. Asai\*, S. Min, J. Grimmelmann, Y. Choi, H. Hajishirzi, L. Zettlemoyer, P. W. Koh, “CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation,” *EMNLP 2024*.
- M. Duan, A. Suri, **N. Mireshghallah**, S. Min, W. Shi, L. Zettlemoyer, Y. Tsvetkov, Y. Choi, D. Evans, H. Hajishirzi, “Do Membership Inference Attacks Work on Large Language Models?” *COLM 2024*.
- **N. Mireshghallah\***, M. Antoniak\*, Y. More\*, Y. Choi, G. Farnadi, “Trust No Bot: Discovering Personal Disclosures in Human–LLM Conversations in the Wild,” *COLM 2024*.
- T. Sorensen, J. Moore, J. Fisher, M. Gordon, **N. Mireshghallah**, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, T. Althoff, Y. Choi, “A Roadmap to Pluralistic Alignment,” *ICML 2024*.
- **N. Mireshghallah\***, H. Kim\*, X. Zhou, Y. Tsvetkov, M. Sap, R. Shokri, Y. Choi, “Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory,” *ICLR 2024 (Spotlight)*.
- X. Tang, R. Shin, H. A. Inan, A. Manoel, **N. Mireshghallah**, Z. Lin, S. Gopi, J. Kulkarni, R. Sim, “Privacy-Preserving In-Context Learning with Differentially Private Few-Shot Generation,” *ICLR 2024*.
- M. Zhang, T. He, T. Wang, **N. Mireshghallah**, B. Chen, H. Wang, Y. Tsvetkov, “LatticeGen: A Cooperative Framework which Hides Generated Text in a Lattice for Privacy-Aware Generation on Cloud,” *NAACL 2024 (Findings)*.
- **N. Mireshghallah**, J. Mattern, S. Gao, R. Shokri, T. Berg-Kirkpatrick, “Smaller Language Models are Better Black-box Machine-Generated Text Detectors,” *EACL 2024*.
- **N. Mireshghallah**, R. Shin, Y. Su, T. Hashimoto, J. Eisner, “Privacy-Preserving Domain Adaptation of Semantic Parsers,” *ACL 2023*.

- J. Mattern, **N. Mireshghallah**, Z. Jin, B. Schölkopf, M. Sachan, T. Berg-Kirkpatrick, “Membership Inference Attacks against Language Models via Neighbourhood Comparison,” *ACL 2023 (Findings)*.
- **N. Mireshghallah\***, N. Vogler\*, J. He, O. Florez, A. El-Kishky, T. Berg-Kirkpatrick, “Non-Parametric Temporal Adaptation for Social Media Topic Classification,” *EMNLP 2023*.
- J. Forristal, **N. Mireshghallah**, G. Durrett, T. Berg-Kirkpatrick, “A Block Metropolis-Hastings Sampler for Controllable Energy-Based Text Generation,” *CoNLL 2023*.
- **N. Mireshghallah**, A. Backurs, H. A. Inan, L. Wutschitz, J. Kulkarni, “Differentially Private Model Compression,” *NeurIPS 2022*.
- **N. Mireshghallah**, A. Uniyal, T. Wang, D. Evans, T. Berg-Kirkpatrick, “Memorization in NLP Fine-tuning Methods,” *EMNLP 2022 (Oral)*.
- **N. Mireshghallah**, K. Goyal, A. Uniyal, T. Berg-Kirkpatrick, R. Shokri, “Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks,” *EMNLP 2022*.
- **N. Mireshghallah**, V. Shrivastava, M. Shokouhi, T. Berg-Kirkpatrick, R. Sim, D. Dimitriadis, “UserIdentifier: Implicit User Representations for Simple and Effective Personalized Sentiment Analysis,” *NAACL 2022*.
- H. Brown, K. Lee, **N. Mireshghallah**, R. Shokri, F. Tramèr, “What Does it Mean for a Language Model to Preserve Privacy?” *FAccT 2022*.
- **N. Mireshghallah**, K. Goyal, T. Berg-Kirkpatrick, “Mix and Match: Learning-Free Controllable Text Generation using Energy Language Models,” *ACL 2022*.
- **N. Mireshghallah**, T. Berg-Kirkpatrick, “Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness,” *EMNLP 2021 (Oral)*.
- **N. Mireshghallah**, H. A. Inan, M. Hasegawa, V. Rühle, T. Berg-Kirkpatrick, R. Sim, “Privacy Regularization: Joint Privacy–Utility Optimization in Language Models,” *NAACL 2021*.
- **N. Mireshghallah**, M. Taram, A. Jalali, A. T. Elthakeb, D. Tullsen, H. Esmaeilzadeh, “Not All Features Are Equal: Discovering Essential Features for Preserving Prediction Privacy,” *WWW 2021*.
- T. Koker, **N. Mireshghallah**, T. Titcombe, G. Kaissis, “U-Noise: Learnable Noise Masks for Interpretable Image Segmentation,” *ICIP 2021*.
- A. T. Elthakeb, P. Pilligundla, **N. Mireshghallah**, A. Cloninger, H. Esmaeilzadeh, “Divide and Conquer: Leveraging Intermediate Feature Representations for Quantized Training of Neural Networks,” *ICML 2020*.
- **N. Mireshghallah**, M. Taram, A. Jalali, D. Tullsen, H. Esmaeilzadeh, “Shredder: Learning Noise Distributions to Protect Inference Privacy,” *ASPLOS 2020*.

#### Journal Papers

- A. T. Elthakeb, P. Pilligundla, **N. Mireshghallah**, A. Yazdanbakhsh, H. Esmaeilzadeh, “ReLeQ: A Reinforcement Learning Approach for Automatic Deep Quantization of Neural Networks,” *IEEE Micro*, 2020.
- **N. Mireshghallah**, M. Bakhshalipour, M. Sadrosadati, H. Sarbazi-Azad, “Energy-Efficient Permanent Fault Tolerance in Hard Real-Time Systems,” *IEEE Transactions on Computers*, 2019.

#### Workshop Papers

- H. Mahdavi, **N. Mireshghallah**, ..., V. Honavar, “RefGrader: Automated Grading of Mathematical Competition Proofs using Agentic Workflows,” *NeurIPS 2025 Workshop on MATH-AI*.
- R. Zhang, M. Kaniselman, **N. Mireshghallah**, “Reinforcement Learning Improves Traversal of Hierarchical Knowledge in LLMs,” *NeurIPS 2025 Workshop on Foundations of Reasoning in Language Models (FoRLM)*.
- Y. Bae\*, ..., **N. Mireshghallah**, “PPMI: Privacy-Preserving LLM Interaction with Socratic Chain-of-Thought Reasoning and Homomorphically Encrypted Vector Databases,” *Workshop on Privacy-Preserving ML at CRYPTO 2025*.
- N. G. Brigham, C. Gao, T. Kohno, F. Roesner, **N. Mireshghallah**, “Breaking News: Case Studies of Generative AI’s Use in Journalism,” *Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024*.
- K. Lee, A. F. Cooper, C. A. Choquette-Choo, K. Liu, M. Jagielski, **N. Mireshghallah**, L. Ahmed, J. Grimmelmann, D. Bau, C. De Sa, ..., “Machine Unlearning Doesn’t Do What You Think,” *GenLaw Workshop at ICML 2024 (Extended Abstract)*.
- R. Naidu, A. Priyanshu, A. Kumar, S. Kotti, H. Wang, **N. Mireshghallah**, “When Differential Privacy Meets Interpretability: A Case Study,” *Responsible Computer Vision Workshop at CVPR 2021*.
- P. Basu, T. Singha Roy, R. Naidu, Z. Muftuoglu, S. Singh, **N. Mireshghallah**, “Benchmarking Differential Privacy and Federated Learning for BERT Models,” *Machine Learning for Data Workshop at ICML 2021*.
- A. Uniyal, R. Naidu, S. Kotti, S. Singh, P. J. Kenfack, **N. Mireshghallah**, A. Trask, “DP-SGD vs. PATE: Which Has Less Disparate Impact on Model Accuracy?,” *Machine Learning for Data Workshop at ICML 2021*.
- T. Farrand, **N. Mireshghallah**, S. Singh, A. Trask, “Neither Private Nor Fair: Impact of Data Imbalance on Utility and Fairness in Differential Privacy,” *PPMLP Workshop at CCS 2020*.
- **N. Mireshghallah**, M. Taram, P. Vepakomma, A. Singh, R. Raskar, H. Esmaeilzadeh, “Privacy in Deep Learning: A Survey,” *arXiv:2004.12254*, 2020.
- M. H. Garcia, A. Manoel, D. M. Diaz, **N. Mireshghallah**, R. Sim, D. Dimitriadis, “FLUTE: A Scalable, Extensible Framework for High-Performance Federated Learning Simulations,” *arXiv:2203.13789*, 2022.

#### Patents

- J. M. Eisner, E. C. Shin, **N. Mireshghallah**, T. B. Hashimoto, Y. Su, “Privacy-Preserving Generation of Synthesized

Training Data,” US Patent Application 18/321,460 (2024).

- **N. Mireshghallah**, H. Esmailzadeh, “Methods of Providing Data Privacy for Neural Network-based Inference,” US Patent 11,487,884.
- **N. Mireshghallah**, H. Esmailzadeh, M. Taram, “Method and System of Learning Noise on Information from Inferences by Deep Neural Network,” US Patent 009062-8413 (2019).

#### Policy Write-ups

- **N. Mireshghallah**, T. Li, “Privacy Is Not Just Memorization,” Technical Report, 2025.
- A. F. Cooper, K. Lee, J. Grimmelmann, D. Ippolito, C. Callison-Burch, C. A. Choquette-Choo, **N. Mireshghallah**, ..., “Report of the 1st Workshop on Generative AI and Law,” *Yale Law & Economics Research Paper*, SSRN-4634513, 2023.

## SELECTED INVITED TALKS

---

### Recent Keynotes & Invited Talks

- Simons Institute (UC Berkeley), Workshop on Trust in Decentralized Systems — “2026 Is the New 2016, but Make It Privacy: On Federated Memory, Contextual Privacy, and Personalized Agents” *Mar. 2026*
- 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*
- Meta AI / FAIR Alignment Group — “What You Should \*Really\* Worry About When It Comes to Generative AI and Privacy” *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 Non-verbatim Memorization” *Aug. 2025*
- Workshop on Collaborative and Federated Agentic Workflows (CFAgentic) at ICML 2025 — Invited Talk *Jul. 2025*
- Fifth Workshop on Trustworthy NLP at NAACL 2025 (TrustNLP) — Invited Talk *May 2025*
- UC Berkeley School of Information — “Privacy, Copyright, and Data Integrity: The Cascading Implications of Generative AI” *Feb. 2025*
- Stanford University NLP Seminar — “Privacy, Copyright and Data Integrity” *Jan. 2025*
- NeurIPS 2024 Red Teaming GenAI Workshop — **Keynote**: “A False Sense of Privacy: Semantic Leakage and Non-literal Copying in LLMs” *Dec. 2024*
- NeurIPS 2024 PrivacyML Tutorial — Panelist *Dec. 2024*
- Future of Privacy Forum — Technologist Roundtable for Policymakers *Nov. 2024*
- National Academies (NASEM), Forum on Cyber Resilience — “Oversharing with LLMs is Underrated” *Aug. 2024*
- Johns Hopkins University, CS Department Seminar *Dec. 2024*
- UMass Amherst NLP Seminar; Northeastern Khoury Security Seminar; UC Santa Barbara, UCLA NLP Seminars; Georgia Tech School of Interactive Computing *2024*
- Google Brain Tech Talk; Meta AI Research; LinkedIn Research; SRI International C3E Workshop; Simons Collaboration TOC4Fairness Seminar; UW Allen School Colloquium *2024*

### Academic 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, EPFL, ETH Zurich, UT Austin, UVA, UNC Chapel Hill.

### Earlier Talks (selected)

- CISA Helmholtz Center for Information Security (2023); Max Planck Institute for Software Systems (2023); NDSS EthiCS Workshop — Keynote (2023); Google Federated Learning Seminar (2023); Mila / McGill (2023); EACL 2023 Tutorial Co-instructor.
- KDD AdvML Workshop (2022); UT Austin (2022); Microsoft Research Cambridge (2022); Johns Hopkins (Guest Lecture, 2022).
- Alan Turing Institute (2021); National University of Singapore (2021); Big Science Panel (2021); Split Learning Workshop (2021).
- UMass Amherst Machine Learning and Friends Lunch (2020); OpenMined Privacy Conference (2020); Microsoft Research AI Breakthroughs Workshop (2020).

## ORGANIZED EVENTS & WORKSHOPS

---

- Co-organizer, Memorization and Trustworthy Foundation Models Workshop at ICML 2025

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

## PROFESSIONAL SERVICE

---

- Area Chair: COLM 2025 & 2026
- D&I Co-chair: NAACL 2025, NAACL 2022
- Widening NLP (WiNLP) Co-chair: 2022–2024
- Shadow PC Member: IEEE Security and Privacy Conference, Winter 2021
- Artifact Evaluation PC Member: USENIX Security 2021, ASPLOS 2020
- PC Member: LatinX in AI Research Workshop at ICML 2020 (LXAI); MLArchSys Workshop at ISCA 2020; WHI Workshop at ICML 2020
- Reviewer: ICLR, NeurIPS, ICML, ACL, EMNLP, IEEE TC, ACM TACO, and others

## DIVERSITY & INCLUSION

---

- Mentor, Women in Machine Learning (WiML) Workshop at NeurIPS 2025, 2024, 2020
- Panelist, CMU SCS Panel: Navigating the Academic Job Market, 2025
- Mentor, ACL Mentorship: “How to Broadcast Your Research to a Wider Audience?” 2025
- Mentor, Graduate Women in Computing (GradWIC) at UCSD, 2020–2023
- Course Instructor, OpenMined Privacy Course, 2020
- Mentor, USENIX Security Undergraduate Mentorship Program, 2020
- Mentor, ICLR 2021
- Co-leader, “Feminist Perspectives for ML & Computer Vision” Break-out at WiML Un-workshop, ICML 2020
- Volunteer / Invited Speaker, Women in Machine Learning Workshop & Meetup, NeurIPS 2019
- Mentor, UCSD CSE Early Research Scholars Program (CSE-ERSP), 2018

## FEATURED PRESS & MEDIA

---

- The Information Bottleneck podcast (Jan. 2026) — on the future of generative AI
- Science News Explores (Nov. 2025) — “5 things to remember when talking to a chatbot”
- Help Net Security (Oct. 2025) — “Most AI privacy research looks the wrong way”
- Washington Post (Aug. 2025) — on AI hype, evaluation metrics, and how people judge AI capabilities
- Science (2025) — “AI writing is improving, but it still can’t match human creativity”
- Jay Shah Podcast (Feb. 2025) — “Differential Privacy, Creativity & Future of AI Research in the LLM Era”
- Thesis Review podcast with Sean Welleck — on Auditing and Mitigating Safety Risks in LLMs
- UW News (2024) — on advances in math and reasoning in new ChatGPT versions
- Washington Post (2024) — on Google AI overview errors, chatbot usage patterns, and opting out of AI training
- WIRED (2024) — “How to Stop Your Data From Being Used to Train AI”