

Niloofar Miresghallah

✉ niloofar@cs.washington.edu

RESEARCH INTERESTS

- Privacy-Preserving and Fair ML
- Federated Learning
- Natural Language Processing
- Efficient Methods for ML

EDUCATION

Ph.D. in Computer Science

UC San Diego, USA

2018-2023

CGPA 3.90/4.00

M.S. in Computer Science

UC San Diego, USA

2018-2020

CGPA 3.90/4.00

B.Sc. in Computer Engineering, Computer Architecture

Sharif University of Technology, Iran

2014-2018

CGPA 18.12/20.00

RESEARCH EXPERIENCE

Postdoctoral Scholar

University of Washington

June 2023-Present

Advisors: Yejin Choi, Yulia Tsvetkov

Part-time Researcher

Microsoft Semantic Machines

Research Group

Oct 2022-July 2023

Mentors: Richard Shin, Yu Su, Tatsunori Hashimoto, Jason Eisner

Research Intern

Microsoft Semantic Machines

Research Group

Jun 2022-Sep 2022

Mentors: Richard Shin, Yu Su, Tatsunori Hashimoto, Jason Eisner

Research Intern

Microsoft Research

Algorithms Group

Jan 2022-March 2022

Mentors: Sergey Yekhanin, Arturs Backurs

Research Intern

Microsoft Research AI

Language and Intelligent Assistance - Federated Learning

June 2021-Sep 2021

Mentors: Dimitrios Dimitriadis, Robert Sim

Graduate Research Assistant

Berg Lab, CSE Department, UC San Diego

Sep 2020-Present

Advisor: Taylor Berg-Kirkpatrick

Research Intern

Microsoft Research AI

Knowledge Technologies and Intelligent Experiences (KTX)

June 2020-Sep 2020

Mentor: Robert Sim

Graduate Research Assistant

ACT Lab, CSE Department, UC San Diego

Sep 2018-Sep 2020

Advisor: Hadi Esmeilzadeh

RAMP Next Generation Platform Technologies Intern

Western Digital Co. Research and Development

June 2019-Sep 2019

Mentor: Anand Kulkarni

Undergraduate Research Assistant

CE Department, Sharif University of Technology

Sep 2016 - June 2018

Advisor: Hamid Sarbazi-Azad

AWARDS

- Momental Foundation Mistletoe Research Fellowship (MRF) Finalist (2023)
- Rising Stars in EECS (2022)
- Rising Star in Adversarial Machine Learning (2022)
- *NCWIT* (National Center for Women & IT) Collegiate Award winner (2020)
- Qualcomm Innovation Fellowship Finalist (2021)
- UCSD CSE Excellence in Leadership and Service Award (2022)
- FAccT Doctoral Consortium (2022)
- Ranked 249th among 223K Participants in the Natl. Univ. Entrance Exam in Math, 2014
- Ranked 57th among 119K Participants in the Natl. Univ. Entrance Exam in Foreign Lang., 2014
- Admitted to Natl. Org. for Exceptional Talents (NODET), ~2% Accept. Rate, 2008 & 2012

PUBLICATIONS

Conference

1. M. Zhang, T. He, T. Wang, **N. Mireshghallah**, B. Chen, H. Wang, and Y. Tsvetkov, "Latticegen: A cooperative framework which hides generated text in a lattice for privacy-aware generation on cloud", in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL Findings)*, Aug. 2024.
2. T. Sorensen, J. Moore, J. Fisher, M. Gordon, **N. Mireshghallah**, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, *et al.*, "A roadmap to pluralistic alignment", in *The Fourty-first International Conference on Machine Learning (ICML)*, Jul. 2024.

3. **N. Mireshghallah**, H. Kim, X. Zhou, Y. Tsvetkov, M. Sap, R. Shokri, and Y. 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.
4. **N. Mireshghallah**, J. Mattern, S. Gao, R. Shokri, and T. 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.
5. X. Tang, R. Shin, H. A. Inan, A. Manoel, **N. Mireshghallah**, Z. Lin, S. Gopi, J. Kulkarni, and R. Sim, “Privacy-preserving in-context learning with differentially private few-shot generation”, in *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.
6. J. Forristal, **N. Mireshghallah**, G. Durrett, and T. Berg-kirkpatrick, “A block metropolis-hastings sampler for controllable energy-based text generation”, in *Proceedings of the 2023 The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, Dec. 2023.
7. **N. Mireshghallah**, N. Vogler, J. He, O. Florez, A. El-Kishky, and T. 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)*, Dec. 2023.
8. J. Mattern, **N. Mireshghallah**, Z. Ji, B. Scholkop, M. Sachan, and T. 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.
9. **N. Mireshghallah**, R. Shin, Y. Su, T. Hashimoto, and J. 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.
10. **N. Mireshghallah**, A. Backurs, H. A. Inan, L. Wutschitz, and J. Kulkarni, “Differentially private model compression”, *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2022.
11. **N. Mireshghallah**, K. Goyal, A. Uniyal, T. Berg-Kirkpatrick, and R. 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)*, Dec. 2022.
12. **N. Mireshghallah**, A. Uniyal, T. Wang, D. Evans, and T. Berg-Kirkpatrick, “Memorization in nlp fine-tuning methods”, in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Selected for **Oral Presentation**, Dec. 2022.
13. **N. Mireshghallah**, V. Shrivastava, M. Shokouhi, T. Berg-Kirkpatrick, R. Sim, and D. 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)*, Jul. 2022.
14. H. Brown, K. Lee, **N. Mireshghallah**, R. Shokri, and F. Tramèr, “What does it mean for a language model to preserve privacy?”, in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, Jun. 2022.

15. **N. Mireshghallah**, K. Goyal, and T. 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.
16. **N. Mireshghallah** and T. 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), Selected for Oral Presentation*, Nov. 2021.
17. T. Koker, **N. Mireshghallah**, T. Titcombe, and G. Kaissis, “U-Noise: Learnable noise masks for interpretable image segmentation”, in *2021 IEEE International Conference on Image Processing (ICIP)*, Sep. 2021.
18. **N. Mireshghallah**, H. A. Inan, M. Hasegawa, V. Rühle, T. Berg-Kirkpatrick, and R. 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)*, Jun. 2021.
19. **N. Mireshghallah**, M. Taram, A. Jalali, A. T. Elthakeb, D. Tullsen, and H. Esmailzadeh, “Not all features are equal: Discovering essential features for preserving prediction privacy”, in *Proceedings of The Web Conference 2021 (WWW)*, Apr. 2021.
20. A. T. Elthakeb, P. Pilligundla, **N. Mireshghallah**, A. Cloninger, and H. Esmailzadeh, “Divide and conquer: Leveraging intermediate feature representations for quantized training of neural networks”, in *The Thirty-seventh International Conference on Machine Learning (ICML)*, Jul. 2020.
21. **N. Mireshghallah**, M. Taram, A. Jalali, D. Tullsen, and H. Esmailzadeh, “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)*, Mar. 2020.

Journal

1. A. T. Elthakeb, P. Pilligundla, **N. Mireshghallah**, A. Yazdanbakhsh, and H. Esmailzadeh, “Releq: A reinforcement learning approach for automatic deep quantization of neural networks”, in *IEEE Micro*, Sep. 2020.
2. **N. Mireshghallah**, M. Bakhshalipour, M. Sadrosadati, and H. Sarbazi-Azad, “Energy-efficient permanent fault tolerance in hard real-time systems”, in *IEEE Transactions on Computers*, Apr. 2019.

Workshop

1. P. Basu, T. Singha Roy, R. Naidu, Z. Muftuoglu, S. Singh, and **N. Mireshghallah**, “Benchmarking differential privacy and federated learning for BERT models”, in *Machine Learning for Data Workshop at ICML 2021*, Jun. 2021.
2. R. Naidu, A. Priyanshu, A. Kumar, S. Kotti, H. Wang, and **N. Mireshghallah**, “When differential privacy meets interpretability: A case study”, in *Responsible Computer Vision Workshop at CVPR 2021*, Jun. 2021.
3. A. Uniyal, R. Naidu, S. Kotti, S. Singh, P. J. Kenfack, **N. Mireshghallah**, and A. Trask, “DP-SGD vs. PATE: Which has less disparate impact on model accuracy?”, in *Machine Learning for Data Workshop at ICML 2021*, Jun. 2021.

4. T. Farrand, **N. Mireshghallah**, S. Singh, and A. Trask, “Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy”, in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS 2020)*, *Privacy-Preserving Machine Learning in Practice workshop (PPMLP)*, Nov. 2020.

Preprint

1. **N. Mireshghallah**, A. M. Kassem, O. Mahmoud, H. Kim, Y. Tsvetkov, Y. Choi, S. Saad, and S. Rana, “Alpaca against vicuna: Using llms to uncover memorization of llms”, *ArXiv preprint arXiv:2403.04801*, Mar. 2024.
2. M. Duan, A. Suri, **N. Mireshghallah**, S. Min, W. Shi, L. Zettlemoyer, Y. Tsvetkov, Y. Choi, D. Evans, and H. Hajishirzi, “Do membership inference attacks work on large language models?”, *ArXiv preprint arXiv:2402.07841*, 2024.
3. X. Fu, Z. Wang, S. Li, R. K. Gupta, **N. Mireshghallah**, T. Berg-Kirkpatrick, and E. Fernandes, “Misusing tools in large language models with visual adversarial examples”, *ArXiv preprint arXiv:2310.03185*, 2023.
4. A. Priyanshu, S. Vijay, A. Kumar, R. Naidu, and **N. Mireshghallah**, “Are chatbots ready for privacy-sensitive applications? an investigation into input regurgitation and prompt-induced sanitization”, *ArXiv preprint arXiv:2305.15008*, 2023.
5. M. H. Garcia, A. Manoel, D. M. Diaz, **N. Mireshghallah**, R. Sim, and D. Dimitriadis, “Flute: A scalable, extensible framework for high-performance federated learning simulations”, *ArXiv preprint arXiv:2203.13789*, 2022.
6. **N. Mireshghallah**, M. Taram, P. Vepakomma, A. Singh, R. Raskar, and H. Esmailzadeh, “Privacy in deep learning: A survey”, *ArXiv preprint arXiv:2004.12254*, Apr. 2020.

Patent

1. **N. Mireshghallah** and H. Esmailzadeh, “Methods of providing data privacy for neural network based inference”, US Patent 009062-8427, Mar. 2020.
2. **N. Mireshghallah**, H. Esmailzadeh, and M. Taram, “Method and system of learning noise on information from inferences by deep neural network”, US Patent 009062-8413, Oct. 2019.

INVITED TALKS

- Jun. 2024: *Alpaca against Vicuna: Using LLMs to Uncover Memorization of LLMs*, CMU
- Apr. 2024: *What is differential privacy? And what is it not?*, Generative AI and Law workshop, DC
- Apr. 2024: *Membership Inference Attacks and Contextual Integrity for Language*, Meta AI Research
- Apr. 2024: *Safety in LLMs: Privacy and Memorization*, Guest lecture at Georgia Tech
- Mar. 2024: *Can LLMs Keep a Secret? Testing Privacy Implications of LLMs*, UCSB
- Mar. 2024: *Can LLMs Keep a Secret? Testing Privacy Implications of LLMs*, UCLA
- Mar. 2024: *Safety in LLMs: Privacy and Memorization*, Guest lecture for CMU 11-830 Ethics Course
- Feb. 2024: *Can LLMs Keep a Secret? Testing Privacy Implications of LLMs*, UT Austin
- Feb. 2024: *Can LLMs Keep a Secret? Testing Privacy Implications of LLMs*, Google Brain
- Jan. 2024: *Can LLMs Keep a Secret? Testing Privacy Implications of LLMs*, UW CSE Colloquium
- Nov. 2023: *Privacy Auditing and Protection in Large Language Model*, UW eScience Institute
- Sep. 2023: *What does privacy-preserving NLP entail?*, CISPA Helmholtz Center for Security
- Sep. 2023: *Auditing and Mitigating Safety Risks in LLMs*, Max Planck Institute - Next 10 in AI
- May 2023: *Privacy Auditing and Protection in Large Language Models*, Mila
- May 2023: *Federated Learning and Privacy Regularization*, Tutorial on Privacy-Preserving NLP, EACL 2023
- May 2023: *Auditing and Mitigating Safety Risks in Large Language Models*, Cohere for AI
- Apr. 2023: *Learning-free Controllable Text Generation*, LLM Interfaces Workshop and Hackathon
- Apr. 2023: *Auditing and Mitigating Safety Risks in Large Language Models*, University of Washington
- Feb. 2023: *How much can we trust large language models?*, Ethics Workshop at NDSS 2023
- Feb. 2023: *Privacy Auditing and Protection in Large Language Models*, Google's FL Seminar
- Oct. 2022: *How much can we trust large language models?*, UT Austin NLP
- Sep. 2022: *Mix and Match: Learning-free Controllable Text Generation*, Johns Hopkins University
- Aug. 2022: *How much can we trust large language models?*, Adversarial ML workshop at KDD 2022
- Mar. 2022: *What Does it Mean for a Language Model to Preserve Privacy?*, MSR Cambridge
- Feb 2022: *What Does it Mean for a Language Model to Preserve Privacy?*, PriSec ML Interest Group
- Jan. 2022: *Style Pooling: Automatic Text Style Obfuscation for Fairness*, UCSD AI lunch
- Dec. 2021: *Improving Attribute Privacy and Fairness for Natural Language Processing*, University of Maine
- Nov. 2021: *Style Pooling: Automatic Text Style Obfuscation for Fairness*, National University of Singapore
- Oct. 2021: *Privacy-Preserving Natural Language Processing*, Big Science for Large Language Models Panel
- Jul. 2021: *Privacy and Interpretability of DNN Inference*, Research Society MIT Manipal
- Jun. 2021: *Low-overhead Techniques for Privacy and Fairness of DNNs*, Alan Turing Institute
- Apr. 2021: *Not All Features are Equal*, UCSD CSE Research Open House
- Mar. 2021: *Shredder: Learning Noise Distributions to Protect Inference Privacy*, Split Learning Workshop
- Feb. 2021: *Introduction to NLP and Career Prospects*, University Institute Of Engineering and Technology
- Oct. 2020: *Privacy and Fairness in DNN Inference*, Machine Learning and Friends Lunch at UMass Amherst
- Sep. 2020: *Privacy-Preserving Natural Language Processing*, OpenMined Privacy Conference
- Sep. 2020: *Private Text Generation through Regularization*, Microsoft Research AI Breakthroughs Workshop
- Jul. 2020: *Shredder: Learning Noise Distributions to Protect Inference Privacy*, NCWIT

TA EXPERIENCE

CSE Department of UC San Diego

CSE 151A (Undergraduate Machine Learning)

Fall 2021

CSE 251A (Graduate Machine Learning)

Winter 2021

CSE 276C (Graduate Mathematics for Robotics)

Fall 2020

CSE 141 (Undergraduate Computer Architecture)

Spring 2020

CSE 240D (Graduate Accelerator Design for Deep Learning)

Winter & Fall 2019

CE Department of Sharif University

Digital Electronics, Computer Architecture, Signals and Systems, Probability and Statistics,

Numerical Methods

2016-2018

DIVERSITY, INCLUSION & MENTORSHIP

- Widening NLP (WiNLP) co-chair
- NAACL 2022 D&I co-chair
- 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
- Course instructor for the OpenMined Privacy Course
- Mentor for the UC San Diego Women Organization for Research Mentoring (WORM) in STEM
- Mentor for the USENIX Security 2020 Undergraduate Mentorship Program
- Volunteer at the Women in Machine Learning Workshop Held at NeurIPS 2019
- Invited Speaker at the Women in Machine Learning and Data Science (WiMLDS) NeurIPS 2019 Meetup
- Mentor for the UCSD CSE Early Research Scholars Program (CSE-ERSP) in 2018

ORGANIZED EVENTS

- Co-organizer of the Generative AI and Law (Gen Law) workshop at ICML 2024
- Co-organizer of the Privacy Regulation and Protection in Machine Learning workshop at ICLR 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 (Gen Law) workshop at ICML 2023
- Co-organizer of the Widening NLP (WiNLP) workshop at EMNLP 2023
- Co-organizer of the Generative AI + Law (GenLaw) workshop at ICML 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
- Privacy & Fairness Roundtable lead at AFCP workshop at NeurIPS 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 Widening NLP (WiNLP) workshop at EMNLP 2021
- Co-leader for the “Machine Learning for Privacy: An Information Theoretic Perspective” Break-out session at the Women in Machine Learning (WiML) Un-workshop Held at ICML 2021
- Co-organizer of the Privacy-Preserving Machine Learning (PPML) Workshop at MICCAI 2021
- Co-organizer of the Distributed Private Machine Learning (DPML) Workshop at ICLR 2021
- Co-organizer of the SoCal joint Machine Learning and Natural Language Processing 2021 Symposium
- Co-leader for the “Feminist Perspectives for Machine Learning & Computer Vision” Break-out session at the Women in Machine Learning (WiML) Un-workshop Held at ICML 2020

PROFESSIONAL SERVICES

- PC member for ACML CCS 2024
- Reviewer for EACL 2024
- Reviewer for CHI 2024
- Reviewer for NeurIPS 2023 Workshop Proposals
- Reviewer for INLG 2023
- Reviewer for TMLR Journal
- Reviewer for FAccT 2023, 2024
- PC member for the AFCP workshop at NeurIPS 2022
- PC member for the TSRML Workshop at NeurIPS 2022
- Reviewer for AAAI 2023, 2024
- Ethics PC member for EMNLP 2022
- Reviewer for NAACL 2022 Student Research Workshop (SRW)
- Reviewer for NeurIPS 2020, 2021, 2022, 2023
- Reviewer for ICML 2020, 2021, 2022, 2023, 2024
- Reviewer for ICLR 2021 (Outstanding Reviewer Award), 2022, 2023, 2024
- Reviewer for IEEE S&P magazine
- Reviewer for CCS 2021 Poster Sessions
- Shadow PC member for IEEE Security and Privacy Conference Winter 2021
- Artifact Evaluation Program Committee Member for USENIX Security 2021
- Security & Privacy Committee Member and Session Chair for Grace Hopper Celebration (GHC) 2020
- Reviewer for ACM TACO Journal
- Reviewer for IEEE TC Journal
- Program Committee Member for Latinx in AI Research Workshop at ICML 2020
- Program Committee Member for the Workshop on Human Interpretability in ML at ICML 2020
- Program Committee Member for the ML for Computer Architecture and Systems Workshop at ISCA 2020
- Artifact Evaluation Program Committee Member for ASPLOS 2020

SKILLS

- | | | |
|-----------|--------------|----------------|
| ○ Python | ○ TVM | ○ Verilog |
| ○ PyTorch | ○ MATLAB | ○ x86 Assembly |
| ○ C/C++ | ○ TensorFlow | ○ Azure ML |