

Enyi (Olivia) Jiang

Contact Information

enyij2@illinois.edu · enyij@stanford.edu · 217-200-2424
[enyijiang.github.io](https://github.com/enyijiang) · [GitHub](#) · [LinkedIn](#)

Research Interests

My research focuses on AI safety and alignment at the intersection of safety and interpretability research. I am particularly interested in understanding and improving model safety through internal representations—studying when behavioral evaluations fail to capture latent risk, how safety-relevant signals emerge during reasoning, and how these signals can support more reliable monitoring and intervention. More broadly, I am interested in building trustworthy AI systems that remain robust under distribution shift and adversarial conditions, as well as applying machine learning to scientific and societal domains, including climate science and healthcare.

Education

- 2023–2027 **University of Illinois Urbana-Champaign**
(expected) Ph.D. in Computer Science. Advisors: Sanmi Koyejo and Nancy Amato.
- 2020–2023 **University of Illinois Urbana-Champaign**
M.S. in Computer Science (thesis track). Advisor: Sanmi Koyejo.
Thesis: *Federated Domain Adaptation for Healthcare*.
- 2016–2020 **University of Illinois Urbana-Champaign**
B.S. in Computer Engineering. Graduated with Highest Honors.
- 2016–2020 **Zhejiang University, ZJU-UIUC Institute**
B.Eng. in Electronics and Computer Engineering. Outstanding Graduate.

Research Appointment

- 2026–present **Stanford University**, Department of Computer Science
Visiting Ph.D. researcher, Stanford Trustworthy AI Research (STAIR) Lab.

Teaching

University of Illinois Urbana-Champaign

- Spring & **CS 441: Applied Machine Learning** — Teaching Assistant.
Fall 2026 Instructor: [Prof. Marco Morales Aguirre](#).
- Fall 2025 **CS 521: Trustworthy AI Systems** — Teaching Assistant.
Instructor: [Prof. Gagandeep Singh](#).
- Fall 2021 **CS 447: Natural Language Processing** — Teaching Assistant.
Instructor: [Prof. Julia Hockenmaier](#).
- Fall 2020 **CS 361: Probability & Statistics for Computer Science**
Teaching Assistant. Instructor: [Prof. Hongye Liu](#).
- Spring 2020 **CS 225: Data Structures** — Teaching Assistant.
Instructor: [Prof. Klaus-Dieter Schewe](#).
- Fall 2019 **MATH 213: Discrete Structures** — Teaching Assistant.
Instructor: [Prof. Klaus-Dieter Schewe](#).

Mentoring

- **Shardul Marathe** — Undergraduate, Computer Science, Stanford University.
- **Hanwen Zhang** — M.C.S., Computer Science, UIUC.
- **David Shu Cheung** — Undergraduate → M.S., Computer Science, UIUC.
- **Sunwoo Baek** — M.C.S., Computer Science, UIUC.
- **Nischay Singh** — Undergraduate, Computer Science, UIUC.
- **Skye Qiu** — Undergraduate, Computer Science and Statistics → M.C.S., Computer Science, UIUC.

AI Alignment @ UIUC — Mentor

Red-teaming/Adversarial and Interpretability streams.

Professional Service

Reviewer: NeurIPS, ICLR, ICML, TMLR, CVPR, ECCV, AAAI, AISTATS, COLM, and TPAMI.

Peer-Reviewed Publications

* denotes equal contribution.

- [1] **Enyi Jiang**, Changming Xu, Nischay Singh, and Gagandeep Singh.
Robust Answers, Fragile Logic: Probing the Decoupling Hypothesis in LLM Reasoning.
Transactions on Machine Learning Research (TMLR), 2026. [Paper](#).
- [2] **Enyi Jiang**, David S. Cheung, and Gagandeep Singh.
Towards Generalized Certified Robustness with Multi-Norm Training.
Transactions on Machine Learning Research (TMLR), 2026. [Paper](#).
- [3] **Enyi Jiang** and Gagandeep Singh.
RAMP: Boosting Adversarial Robustness Against Multiple ℓ_p Perturbations for Universal Robustness.
Advances in Neural Information Processing Systems (NeurIPS), 2024. [Paper](#).
- [4] **Enyi Jiang***, Yibo Jacky Zhang*, and Oluwasanmi O. Koyejo.
Principled Federated Domain Adaptation: Gradient Projection and Auto-Weighted Scheme.
International Conference on Learning Representations (ICLR), 2024. [Paper](#).
- [5] Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, **Enyi Jiang**, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping, and Jiawei Han.
EVIDENCEMINER: Textual Evidence Discovery for Life Sciences.
Annual Meeting of the Association for Computational Linguistics (ACL), System Demonstrations, 2020. [Paper](#); [demo](#).

Preprints and Manuscripts Under Review

- M1. Chengxiao Wang*, **Enyi Jiang***, Xiaojing Liao, and Sanmi Koyejo.
CLEAR: Continuous Latent Adapter Routing for Better Utility-Preserving LLM Safety Alignment.
Submitted to AAAI 2027.
- M2. **Enyi Jiang***, Anders Gjørbye*, Yibo Jacky Zhang, and Sanmi Koyejo.
When Behavioral Safety Evaluation Fails: A Representation-Level Perspective.
Submitted to NeurIPS 2026; accepted by Stanford Causal Science Conference, Frontiers in AI Evaluation.
[Preprint](#); [poster](#).
- M3. **Enyi Jiang** and Wu Sun.
Environment-Adaptive Preference Optimization for Wildfire Prediction.
Preprint. [Paper](#).
- M4. **Enyi Jiang**, Yibo Jacky Zhang, Yinglun Xu, Andreas Haupt, and Sanmi Koyejo.
Latent Adversarial Regularization for Offline Preference Optimization.
Submitted to NeurIPS 2026. [Preprint](#).

Other Scholarly Work

- 2023 **Enyi Jiang**. *Federated Domain Adaptation for Healthcare*. M.S. thesis, University of Illinois Urbana-Champaign. [Thesis](#).
- 2022 **Enyi Jiang** and Haisong Zhang. *A New Coarse and Fine-grained Chinese Word Segmentation Benchmark*. Chinese patent application no. CN202210157392.7.
- 2019 Shirui Luo, Jiahuan Cui, Madhu Vellakal, Jian Liu, **Enyi Jiang**, Seid Koric, and Volodymyr Kindratenko. *Review and Examination of Input Feature Preparation Methods and Machine Learning Models for Turbulence Modeling*. [Paper](#).

Research Internships

- Summers 2025, 2026 **Meta**, Menlo Park, CA — Research Intern
Mentors: Dake Chen (2025), Gary Jiang (2026). Investigated diffusion distillation and acceleration methods for image and video generation.
- Jun–Aug 2023 **Argonne National Laboratory**, remote — Research Intern
Host: Ravi K. Madduri. Integrated federated domain adaptation algorithms into [APPFL](#); collaborated on custom losses and metrics and surveyed federated learning frameworks.
- Feb–May 2023 **TensorOpera AI (formerly FedML)**, remote — Research Intern
Developed federated learning algorithms for healthcare across four hospital silos; investigated small-data settings using generative models and self-supervised learning.
- Dec 2020–Apr 2021 **Tencent AI Lab**, NLP Center, Guangdong, China — NLP Intern
Proposed coarse- and fine-grained Chinese word segmentation labeling standards, a benchmark covering difficult cases and varied domains, and cleaned word embeddings; first author of the associated patent application.

Honors and Awards

- 2022 Second place, UIUC Ashby Prize in Computational Science Hackathon.
- 2020 Outstanding Graduate, Zhejiang University.
- 2019 Dean's List, University of Illinois Urbana-Champaign.
- 2018 Meritorious Winner, Mathematical Contest in Modeling.
- 2017, 2018 Academic First-Class Scholarship, ZJU-UIUC Institute.
- 2017 First-Class Student Award, Zhejiang University.

References

Sanmi Koyejo

Assistant Professor, Department of Computer Science
Stanford University
sanmi@cs.stanford.edu

Nancy M. Amato

Abel Bliss Professor of Engineering; Director, Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
namato@illinois.edu

Volodymyr Kindratenko

Research Associate Professor, National Center for Supercomputing Applications
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
kindrtnk@illinois.edu