

Enyi (Olivia) Jiang

AI Safety & Alignment

enyij@stanford.edu | 217-200-2424

enyijiang.github.io | github.com/enyijiang | LinkedIn

Profile

AI safety and alignment researcher and final-year UIUC CS Ph.D. candidate (expected 2027), visiting Stanford STAIR. Develop representation-level safety evaluation, preference optimization, and selective intervention methods that preserve model utility. First/co-first-authored work at NeurIPS, ICLR, and TMLR, with foundations in adversarial robustness and domain adaptation. At Meta, built image/video generation methods for production pipelines and ran multi-node, multi-GPU training with FSDP.

Technical Expertise

LLM Safety	Representation-level safety audits, token/embedding-level adversarial testing, reasoning robustness evaluation, selective safety LoRA adapters.
Alignment Training	Offline preference optimization (DPO, SimPO), latent adversarial regularization, policy/reference representation alignment.
RL & Agents	Familiarity with GRPO, RLHF, agent rollouts, tool-use, and multi-turn trajectory training.
Distributed Training	PyTorch, FSDP, multi-node/multi-GPU model training.
Multimodal Models	Image/video generation, diffusion models, distillation, generation acceleration, production pipeline integration.
Robust Learning	Adversarial and certified training, multi-norm threat models, distribution shift, federated learning, domain adaptation.
Engineering	Python, NumPy, pandas, scikit-learn, C/C++, Java, Linux, Bash, Git.

Education

University of Illinois Urbana-Champaign Ph.D., Computer Science (M.S. 2023).	Ph.D. expected 2027
UIUC / Zhejiang University, ZJU-UIUC Institute B.S., Computer Engineering (UIUC, Highest Honors); B.Eng., Electronics and Computer Engineering (ZJU).	2016 – 2020

Experience

Stanford University Visiting Student Researcher, Stanford Trustworthy AI Research (STAIR) Lab	2026 – Present Stanford, CA
- Co-led CLEAR, using a hidden-state gate to control a safety LoRA adapter. Reduced HarmBench attack success from 32.3% to 0.5% on Llama-3-8B-Instruct, with up to 7.1 percentage points higher GSM8K accuracy than globally applied safety SFT or LoRA. - Developed representation-level safety auditing and a Latent Vulnerability Score. Across three model families, constructed models showed 2.5–3.1× the vulnerability of aligned bases at the targeted middle layer despite safe-looking outputs.	
Meta Research Intern Image & Video Generation	Summers 2025 & 2026 Menlo Park, CA
- Developed diffusion distillation and acceleration methods for image/video generation; work was integrated into Meta's production pipeline. - Ran multi-node, multi-GPU training with PyTorch FSDP for image/video generation models. - Received an internship return offer.	
Argonne National Laboratory Research Intern	Jun – Aug 2023 Remote
Integrated federated domain adaptation algorithms into open-source APPFL; reproduced results and implemented custom losses and evaluation metrics with the team.	
TensorOpera AI (formerly FedML) Research Intern	Feb – May 2023 Remote
Developed federated learning algorithms across four hospital silos; explored diffusion models and self-supervised learning with limited data.	

- Designed Chinese word segmentation labeling standards and a benchmark spanning difficult cases and diverse domains; cleaned word embeddings and first-authored the associated patent application.

Research: AI Safety & Alignment

- Selective safety intervention | CLEAR:** Co-led continuous hidden-state routing of a safety adapter, intervening selectively while preserving benign capabilities; evaluated the safety-utility tradeoff against global safety SFT and LoRA. [Paper]
- Representation-level evaluation | Audit gap:** Constructed models that appear safe under static behavioral audits yet remain vulnerable to bounded internal perturbations. Developed intervention-based auditing and a Latent Vulnerability Score to quantify this gap. [Paper]
- Preference optimization | GANPO:** Designed a plug-in latent adversarial regularizer that aligns policy/reference representation distributions, complementing token-level objectives in DPO and SimPO. Improved robustness under distribution shift and high-entropy sampling with minor computational overhead and comparable downstream performance. [Code]
- Reasoning safety | MATCHA:** Built answer-conditioned token- and embedding-level adversarial probes and an automated evaluation framework. Showed that correct answers can coexist with inconsistent reasoning, identified greater fragility in multi-step and commonsense tasks, and studied transfer to black-box LLMs. [Code]
- Robustness foundations | RAMP/CURE:** Developed logit pairing and gradient projection to balance accuracy and competing multi-norm adversarial objectives; extended robustness to certified training through bound alignment and training strategies.

Research: Applied AI

- Climate AI | PyroAdapt:** Built a pretrain-retrieve-rank framework for wildfire prediction under spatial heterogeneity and temporal shift. Across California, raised daily average precision from **21.62% to 24.57%** and Top-5% recall from **18.70% to 22.79%**, capturing **344 additional positive cell-days** under a fixed detection budget.
- Healthcare AI | Federated domain adaptation:** Designed gradient projection and auto-weighted aggregation to combine source and target model updates under domain shift. Improved adaptation with limited target data across healthcare and general datasets; studied demographic information as auxiliary signals for COVID-19 imaging in MIDRC.

Publications

* denotes equal contribution.

- Enyi Jiang**, Changming Xu, Nischay Singh, Tian Qiu, Gagandeep Singh. *Robust Answers, Fragile Logic: Probing the Decoupling Hypothesis in LLM Reasoning*. (TMLR 2026)
- Enyi Jiang**, David S. Cheung, Gagandeep Singh. *Towards Generalized Certified Robustness with Multi-Norm Training*. (TMLR 2026)
- Enyi Jiang**, Gagandeep Singh. *RAMP: Boosting Adversarial Robustness Against Multiple ℓ_p Perturbations for Universal Robustness*. (NeurIPS 2024)
- Enyi Jiang***, Yibo Jacky Zhang*, Oluwasanmi O. Koyejo. *Principled Federated Domain Adaptation: Gradient Projection and Auto-Weighting*. (ICLR 2024)
- Chengxiao Wang*, **Enyi Jiang***, Xiaojing Liao, Sanmi Koyejo. *CLEAR: Continuous Latent Adapter Routing for Utility-Preserving LLM Safety Alignment*. Trustworthy AI for Good Workshop @ NeurIPS 2026 (accepted). (Paper)
- Enyi Jiang***, Anders Gjølbbye*, Yibo Jacky Zhang, Sanmi Koyejo. *When Behavioral Safety Evaluation Fails: A Representation-Level Perspective*. Preprint, 2026. Stanford Causal Science Conference, Frontiers in AI Evaluation (poster). (Paper) (Poster)
- Enyi Jiang**, Yibo Jacky Zhang, Yinglun Xu, Andreas Haupt, Nancy Amato, Sanmi Koyejo. *Latent Adversarial Regularization for Offline Preference Optimization*. Preprint, 2026. (Paper)
- Enyi Jiang**, Wu Sun. *PyroAdapt: Adapting Wildfire Prediction under Spatial Heterogeneity and Temporal Shift*. Preprint, 2026. (Paper)
- Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, **Enyi Jiang**, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping, Jiawei Han. *EVIDENCEMINER: Textual Evidence Discovery for Life Sciences*. ACL 2020, System Demonstrations. (Paper) (System demo)
- Enyi Jiang**. *Federated Domain Adaptation for Healthcare*. University of Illinois Urbana-Champaign, 2023. (M.S. Thesis)
- Enyi Jiang**, Haisong Zhang. *A New Coarse and Fine-grained Chinese Word Segmentation Benchmark*. Chinese patent application no. CN202210157392.7, 2022.
- Shirui Luo, Jiahuan Cui, Madhu Vellakal, Jian Liu, **Enyi Jiang**, Seid Koric, Volodymyr Kindratenko. *Review and Examination of Input Feature Preparation Methods and Machine Learning Models for Turbulence Modeling*. 2019. (Paper)

Service

Founder, Inside AI Safety, Inc. — independent nonprofit AI safety research organization.

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

Mentor, AI Alignment @ UIUC: Red-teaming/Adversarial and Interpretability streams.