RESEARCH STATEMENT

Understanding and Improving AI Safety through Internal Representations

Enyi (Olivia) Jiang

enyij@stanford.edu · enyij2@illinois.edu · enyijiang.github.io

A model can produce the correct answer while offering an unstable explanation, or refuse a harmful request while remaining vulnerable to a small internal perturbation. These discrepancies motivate the central question of my research: **what evidence is needed to establish that an AI system's safety is robust, rather than contingent on the conditions under which it is evaluated?**

My research focuses on AI safety and alignment at the intersection of safety and interpretability. I study how internal representations can reveal risks that behavioral evaluations miss, guide alignment, and support monitoring and targeted intervention. My long-term goal is to connect a mechanistic understanding of safety in open-weight models with effective monitoring and steering of agentic AI systems under distribution shift and adversarial conditions. I approach this goal through three connected activities: diagnosing hidden vulnerabilities, shaping representations during learning, and developing monitoring and control methods for deployment. My work in adversarial robustness provides the methodological foundation, while healthcare and climate science motivate settings in which reliability must extend beyond a benchmark.

1. Diagnosing the gap between observed behavior and robustness

Correct answers can conceal fragile explanations. In *MATCHA* [1], I developed methods to test the stability of generated reasoning while holding the predicted answer fixed. Token- and embedding-level perturbations can destabilize a model's rationale even when its final answer remains correct. This separates two properties often conflated in evaluation: answer accuracy and explanation robustness. The finding motivates closer examination of the processes associated with a prediction. It does not, by itself, establish that a generated chain of thought faithfully describes internal computation; that distinction is central to my interpretability agenda.

Safe outputs can conceal internal vulnerability. In *When Behavioral Safety Evaluation Fails* [2], my collaborators and I examine a complementary discrepancy in safety evaluation. We construct models whose observed refusals resemble those of safety-aligned base models, yet whose harmful behavior is more accessible through bounded internal perturbations. Static output tests, and even a fixed probe on clean activations, can miss this difference. Our intervention-based analysis measures how readily a model's safety deteriorates when its internal state is perturbed, providing evidence that observational audits alone do not capture.

Together, these projects motivate an evaluation agenda that connects model behavior with controlled tests of internal vulnerability. I want to determine which signals predict failures, which are causally relevant to those failures, and when either relationship breaks under adaptation. This requires pairing representation analysis with interventions and held-out failure conditions. A useful safety signal should remain informative beyond the examples on which it was discovered, and its limitations should be measurable.

2. Improving alignment and robustness through learning

Using representations as a training signal. Diagnosis becomes more useful when it informs how models are trained. In *GANPO* [3], I introduce latent adversarial regularization for offline preference optimization. The method penalizes divergence between policy and reference representations using an adversarial objective, complementing token-level constraints. It can be integrated into existing preference-optimization objectives and provides structural feedback under distribution shift and noise with modest computational overhead. This work establishes a practical route for incorporating internal representations into alignment. It also raises a question for my future work: which aspects of a reference representation should be preserved, and which should change as capabilities and tasks evolve?

My ongoing work on conditional runtime intervention, including *CLEAR* [4], extends this interest to the safety–utility tradeoff. I am investigating how safety mechanisms can be applied selectively while retaining useful behavior. This direction connects training-time representation shaping with deployment-time control: the first seeks to reduce vulnerability, while the second responds when evidence suggests that a particular computation requires additional scrutiny.

Learning to withstand multiple failure conditions. My earlier work provides tools for reasoning about competing reliability objectives. In *RAMP* [5], I develop multi-norm adversarial training that addresses tradeoffs among perturbation models and between accuracy and robustness. In *CURE* [6], I extend this line of work to certified training across multiple norm-bounded threat models. These projects combine analysis of robustness tradeoffs with training methods designed to improve protection across conditions. Their guarantees remain tied to specified threat models; empirical generalization to additional perturbations is a separate claim.

This distinction shapes how I approach language-model safety. Norm-bounded robustness in vision does not automatically imply semantic safety in language. Instead, I bring forward the discipline of specifying an intervention set, identifying competing objectives, and evaluating the conditions under which protection transfers. My goal is to develop similarly explicit foundations for representation-level safety: what is being protected, against which changes, and with what evidence?

Grounding reliability in scientific and societal settings. My work on federated domain adaptation [7] develops gradient-projection aggregation and auto-weighting to transfer information when a target client has limited data and differs from source clients. Motivated in part by healthcare, this work addresses the challenge of learning across heterogeneous environments. In complementary work on environment-adaptive preference optimization for wildfire prediction [8], I investigate learning under environmental shift and long-tailed data. These applications motivate a broader research question: how should a learning system adapt when familiar statistical patterns no longer support reliable decisions?

I plan to use healthcare and climate collaborations to test the scope of my methods, with evaluation criteria developed around the scientific task. Reliability may require different evidence in a clinical prediction system, a wildfire model, and a language-model assistant. Connecting these settings through explicit assumptions about shift and failure will help identify which principles transfer and which mechanisms must remain domain-specific.

3. Future agenda: Mechanistic safety and agentic control

My future research will connect two questions: **how do open-weight models implement safety, and how can that understanding support the monitoring and steering of AI agents?** I aim to build a research program in which mechanistic investigation informs actionable interventions, and failures in agentic settings expose the limits of our understanding. These directions build on my work in representation-level evaluation, latent alignment, and robustness.

Direction I: Monitoring and steering agentic AI for safety. I will develop methods to monitor agents as they reason, use tools, and interact with an environment over multiple steps. The central challenge is to detect an emerging unsafe trajectory early enough to change its course. I will investigate how signals from internal representations can be combined with observable actions, tool outputs, and interaction history to distinguish acceptable task progress from developing risk. This extends my work on behavioral evaluation gaps from individual responses to sequences of decisions.

An initial project will examine matched agent trajectories that achieve similar task progress but differ in their safety outcomes. I will study when warning signals first emerge, how they evolve across reasoning and action steps, and whether they remain informative on unfamiliar tasks. Comparisons with action-based monitors and static probes will test the added value of internal access. Evaluation will measure detection lead time, missed unsafe actions, false alarms, and robustness to adversarial changes in the environment.

I will then connect monitoring to steering: targeted interventions that redirect an agent while preserving its ability to complete legitimate tasks. Candidate mechanisms include modifying safety-relevant activations, conditioning subsequent planning, and triggering additional verification before consequential actions. Controlled comparisons will test which interventions change downstream behavior and whether their effects persist over the rest of a trajectory. I will evaluate safety alongside task completion and intervention cost, using independent outcome checks to distinguish reduced risk from merely reduced monitor activation. Longer-term work will examine how monitoring and steering remain effective as agents adapt and operate over longer horizons.

Direction II: Mechanistic understanding of open-weight model safety. I will investigate which internal features and computational pathways contribute to safety-relevant behavior, how they change during alignment, and why they fail under perturbation or adaptation. Access to model weights and activations enables controlled comparisons across checkpoints and interventions on candidate mechanisms. Building on my representation-level safety evaluation, I will distinguish features that predict refusal from mechanisms that causally influence whether the model behaves safely.

An initial project will compare models before and after safety alignment, using matched benign and harmful requests to identify candidate changes. Activation patching, ablation, and targeted steering will test their causal roles. I will ask whether a mechanism generalizes across request types, whether it affects benign capabilities, and whether its influence survives fine-tuning or adversarial pressure. Comparing multiple open-weight model families will help establish the scope of each finding. The objective is to develop falsifiable explanations of safety behavior, including evidence about where those explanations cease to apply.

The two directions will inform each other. Mechanistic findings will supply hypotheses for agent monitors and precise targets for steering; agentic failure cases will test whether mechanisms identified in isolated responses remain relevant during extended interaction. Together, they define my long-term goal: to make AI safety more understandable at the level of model computation and more actionable at the level of autonomous behavior.

Selected References

[1] Enyi Jiang, Changming Xu, Nischay Singh, Tian Qiu, and Gagandeep Singh. Robust Answers, Fragile Logic: Probing the Decoupling Hypothesis in LLM Reasoning. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2505.17406.

[2] Enyi Jiang, Anders Gjølbye, Yibo Jacky Zhang, and Sanmi Koyejo. When Behavioral Safety Evaluation Fails: A Representation-Level Perspective. *arXiv preprint arXiv:2606.08044*, 2026. URL https://arxiv.org/abs/2606.08044.

[3] Enyi Jiang, Yibo Jacky Zhang, Yinglun Xu, Andreas Haupt, Nancy Amato, and Sanmi Koyejo. Latent Adversarial Regularization for Offline Preference Optimization. *arXiv preprint arXiv:2601.22083*, 2026. URL https://arxiv.org/abs/2601.22083.

[4] Chengxiao Wang, Enyi Jiang, Xiaojing Liao, and Sanmi Koyejo. CLEAR: Continuous Latent Adapter Routing for Better Utility-Preserving LLM Safety Alignment. Manuscript submitted to AAAI 2027. Chengxiao Wang and Enyi Jiang contributed equally, 2026.

[5] Enyi Jiang and Gagandeep Singh. RAMP: Boosting Adversarial Robustness Against Multiple ℓ_p Perturbations for Universal Robustness. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2402.06827.

[6] Enyi Jiang, David S. Cheung, and Gagandeep Singh. Towards Generalized Certified Robustness with Multi-Norm Training. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2410.03000.

[7] Enyi Jiang, Yibo Jacky Zhang, and Sanmi Koyejo. Principled Federated Domain Adaptation: Gradient Projection and Auto-Weighting. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2302.05049.

[8] Enyi Jiang and Wu Sun. Environment-Adaptive Preference Optimization for Wildfire Prediction. *arXiv preprint arXiv:2605.12435*, 2026. URL https://arxiv.org/abs/2605.12435.