Taeyoun Kim

Statement of Purpose | EECS PhD Fall 2026

Overview

I am interested in understanding how language models can **self-improve** through **reinforcement learning** in an **aligned** manner on **open-ended tasks**. As we approach harder open domains where training signals are scarce, I believe it is crucial to understand how to extrapolate the capabilities of current models. I started pursuing this direction at CMU with **Dr. Aviral Kumar**, **Dr. Aditi Raghunathan**, and **Dr. Maarten Sap** by examining the robustness of fine-tuning, alignment, and reasoning for safety, leading to first-author publications at **NeurIPS 2024/2025 and ACL Findings**. During my PhD, I aim to (1) extend my work to broader open-ended reasoning and (2) build a multi-agent training framework for non-verifiable tasks while keeping models aligned.

Research Experience

Robustness. My first project on robustness with **Dr. Aditi Raghunathan** addresses a fundamental question in real-world deployment: how do we **estimate out-of-distribution (OOD) performance** of foundation models without labels? Since models lose accuracy on OOD tasks, it is important to anticipate performance loss. I studied Agreement-on-the-Line (AGL) [1], a phenomenon that predicts OOD accuracy from inter-model agreement. My **NeurIPS 2024** paper [2] extends AGL to foundation models, where heavy pre-training could hurt model diversity, making agreement less predictive of accuracy. I trained over 200 models across three fine-tuning strategies (BitFit/IA3/LoRA), three tasks (extractive QA/generative QA/text classification), and six model families and found that randomly initializing the last layer during fine-tuning recovers the diversity lost during pre-training. This helped me reduce the error of model estimation to 1.64% (SOTA: 2.8% [3]), giving practitioners a way to assess model robustness before deployment.

Safety/Alignment. Since fine-tuning could overcome pre-training, I next tried to understand **how well fine-tuning aligns models**. I created a benchmark called **The Purple Problem** (**NeurIPS 2024 Safe GenAI** [4]) to test if preference-tuning can prevent models from generating the simplest output—a single word "*purple*". I stress-tested defenses (e.g., DPO/PPO) by iteratively fine-tuning models to avoid generating "*purple*" and attacking them to produce it, finding that all defenses fail (fail rate: 98.3%). My **ACL Findings 2025** paper [5] with **Dr. Maarten Sap** further shows that fine-tuning to mitigate bias on a system of models (RAG) fails under distribution shift, demonstrating the brittleness of alignment. These work highlight that fine-tuning cannot ensure alignment, even on the simplest task.

Reasoning for Safety. If fine-tuning cannot achieve safety, does reasoning improve safety? While most work focuses on verifiable math [6], I studied **RL for open-ended reasoning in safety** with **Dr. Aviral Kumar**. My **NeurIPS 2025** paper [7] develops an RL recipe to solve over-refusal, where safety-trained models refuse to benign prompts. I identified that large open answer spaces in safety make models reward-hack and lose reasoning capabilities, degenerating into refusals. To fix this, I created an RL training recipe called **TARS** by adding an auxiliary task-completion reward on a general task to encourage reasoning, so that the improved reasoning capabilities *transfer* back to the main safety task. Throughout this project, I became able to run reinforcement learning for over 1000 steps without losing stability. I successfully trained a Qwen2.5-1.5B with TARS to be safer than larger 7B models and SOTA defenses such as deliberative alignment [8], demonstrating that reasoning can improve adaptive safety.

Beyond Safety. Although my work has focused on safety, *I aim to extend my work on language models to broader open-ended domains such as long-form math, scientific discovery, and even apply them to embodied agents*. My experience with reasoning degeneration, a form of reward hacking in large answer spaces, has prepared me to tackle similar problems in other open-ended tasks. Motivated by this, my recent work extends TARS to **creating process rewards**, directly shaping reasoning for better alignment and less reward hacking, a different paradigm than using outcome rewards in verifiable math.

During my PhD (Research Proposal: https://danielkty.github.io/pdfs/research.pdf)

Throughout my work, I realized that reinforcement learning relies on a more capable model to provide rewards or reference solutions. This means our current paradigm of training will fail on difficult open-ended tasks where models lack the capability to *verify* generations or even *generate* correct answers. I would like to build a new training paradigm with **Dr. Dylan Hadfield-Menell**, **Dr. Marzyeh Ghassemi**, **Dr. Jacob Andreas**, and **Dr. Yoon Kim** that focuses on two interrelated research questions when only low-capable models are available without external supervision: (1) how do we induce generation capabilities and (2) how do we induce verification capabilities? They are interrelated because without a verifier, there is no reward, and even with a verifier, training is ineffective if the generation capability is so poor that rewards are near zero.

To answer these questions, I plan to build a multi-agent training framework that bootstraps low-capability models to be both a generator and verifier that mutually improve each other. The main challenge that such self-improvement frameworks [9, 10, 11, 12] fail to address is the **no free lunch (NFL)** problem and **misalignment**; it is unclear how a system of models can improve capabilities and remain aligned without external supervision. My work on TARS [7] provides a starting point. It shows that mixing in a verifiable task can help recover lost capabilities on another task. In a broader sense, this suggests that capabilities can flow from verifiable tasks to non-verifiable tasks because skills or abilities (e.g., sub-proofs, theorems, safety constraints) used for easier problems will *transfer* over to harder ones, addressing NFL and misalignment. I hope to further explore **the dynamics of how capability transfers**, for example, by finding a sample efficient way to interleave easier verifiable math problems when training for difficult non-verifiable math problems.

After my PhD

I plan to stay in academia and make language models improve on open domains by leveraging the capabilities of existing models. A PhD at MIT would help me connect my experience to large-scale multi-agent training. I hope to establish a new training paradigm with MIT's support that relies less on stronger models.

References

[1] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35: 19274–19289, 2022.

[2] Rahul Saxena, Taeyoun Kim, Aman Mehra, Christina Baek, Zico Kolter, and Aditi Raghunathan. Predicting the performance of foundation models via agreement-on-the-line. 2024. URL https://arxiv.org/abs/2404.01542.

[3] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1134–1144, 2021.

[4] Taeyoun Kim, Suhas Kotha, and Aditi Raghunathan. Testing the limits of jailbreaking defenses with the purple problem. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://neurips.cc/workshop/2024/safe-generative-ai.

[5] Taeyoun Kim, Jacob Mitchell Springer, Aditi Raghunathan, and Maarten Sap. Mitigating bias in RAG: Controlling the embedder. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18999–19024, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.974. URL https://aclanthology.org/2025.findings-acl.974/.

[6] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[7] Taeyoun Kim, Fahim Tajwar, Aditi Raghunathan, and Aviral Kumar. Reasoning as an adaptive defense for safety. *arXiv preprint arXiv:2507.00971*, 2025.

[8] Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

[9] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024.

[10] Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11548–11565, 2025.

[11] Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching language model agents how to self-improve. *Advances in Neural Information Processing Systems*, 37:55249–55285, 2024.

[12] Vighnesh Subramaniam, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*, 2025.