

KYUYOUNG KIM

kykim@cs.stanford.edu

kykim0.github.io

EDUCATION

KAIST AI, Seoul, South Korea

Ph.D., Artificial Intelligence

Sep 2023 - Present

- Advisor: Prof. Jinwoo Shin
- Research areas: RLHF, LLM agents, generative models, AI safety
- Expected graduation: June 2027

University of Washington, Seattle, WA, USA

Visiting Researcher, Paul G. Allen School of Computer Science and Engineering

Aug 2025 - Present

- Advisor: Prof. Sewoong Oh
- Topics: Multi-turn RL, LLM agents, social intelligence

Stanford University, Stanford, CA, USA

M.S., Computer Science with Distinction in Research

- Advisor: Prof. Mykel Kochenderfer
- Thesis: *Adaptive Algorithms for Efficient Risk Estimation of Black-Box Systems*

Cornell University, Ithaca, NY, USA

B.S., Computer Science with Distinction

- Minor in Applied Mathematics

RESEARCH EXPERIENCE

KAIST AI, Seoul, South Korea

Research Assistant

May 2023 - Present

Algorithmic Intelligence Lab

- Advisor: Prof. Jinwoo Shin
- Researching RLHF methods for safe and effective alignment of generative models.
- Working on LLM-based AutoML, personalization, and agentic system design.

University of Washington, Seattle, WA, USA

Visiting Researcher

Aug 2025 - Present

- Advisor: Prof. Sewoong Oh
- Researching multi-turn RL with interactive environments for LLM agents.
- Developing evaluation methods for social intelligence and multi-turn planning in LLMs.

Stanford University, Stanford, CA, USA

Research Assistant

Sep 2021 - Dec 2022

Stanford Intelligent Systems Lab

- Advisor: Prof. Mykel Kochenderfer
- Designed RL-based algorithms for efficient risk estimation of black-box systems.
- Applied risk estimation methods to validate autonomous vehicle policies in simulation.

Stanford Vision and Learning Lab

Mar 2022 - Dec 2022

- Advisors: Prof. Fei-Fei Li and Prof. Jiajun Wu
- Contributed to developing a high-fidelity embodied AI benchmark.
- Generated large-scale synthetic datasets for training embodied agents and vision models.

WORK EXPERIENCE

Google, Mountain View, CA, USA

Senior Software Engineer

Oct 2016 - April 2021

Google Assistant

- Led localization efforts for Google Assistant, extending support to low-resource languages.

- Researched deep learning methods for data-to-text generation, developing key model components.
- Built large-scale pipelines for training translation models in collaboration with the Translation team.
- Developed a human-in-the-loop dialogue system to collect diverse datasets for NLU research.

Waze Carpool

Nov 2015 - Oct 2016

- Developed an on-demand ride-sharing platform as backend engineer.
- Designed logging infrastructure, enhanced matching algorithms, and improved user modeling.

Display Ads

June 2013 - Nov 2015

- Contributed to Gmail monetization initiatives as backend engineer.
- Implemented key improvements to ad auction algorithms and serving infrastructure.

TEACHING

Instructor

Machine Learning Crash Course, Google

May 2019

Teaching Assistant

CS229 Machine Learning, Stanford

Autumn 2022, Summer 2022

CS108 Object-Oriented Systems Design, Stanford

Winter 2022

CS4820 Introduction to Algorithms, Cornell

Spring 2011, Spring 2012

CS3220 Scientific Computation, Cornell

Spring 2012

CS2800 Discrete Structures, Cornell

Spring 2010, Fall 2010

PUBLICATIONS

Conference proceedings

- [1] **K. Kim***, H. Jeon*, J. Shin. Self-Refining Language Model Anonymizers via Adversarial Distillation In *NeurIPS*, 2025.
- [2] **K. Kim**, J. Shin, J. Kim. Personalized Language Models via Privacy-Preserving Evolutionary Model Merging. In *EMNLP (oral)*, 2025.
- [3] D. Choi, S. Oh, S. Dingliwal, J. Tack, **K. Kim**, W. Song, S. Kim, I. Han, J. Shin, A. Galstyan, S. Katiyar, S. B. Bodapati. Mamba Drafters for Speculative Decoding In *EMNLP Findings*, 2025.
- [4] D. Lee*, J. Lee*, **K. Kim**, J. Tack, J. Shin, Y. Teh, K. Lee. Learning to Contextualize Web Pages for Enhanced Decision Making by LLM Agents. In *ICLR*, 2025.
- [5] J. Nam*, **K. Kim***, S. Oh, J. Tack, J. Kim, J. Shin. Optimized Feature Generation for Tabular Data via LLMs with Decision Tree Reasoning. In *NeurIPS*, 2024.
- [6] **K. Kim***, A. Seo*, H. Liu, J. Shin, K. Lee. Margin Matching Preference Optimization: Enhanced Model Alignment with Granular Feedback. In *EMNLP Findings*, 2024.
- [7] **K. Kim**, J. Jeong, M. An, M. Ghavamzadeh, K. Dvijotham, J. Shin, K. Lee. Confidence-aware Reward Optimization for Fine-tuning Text-to-Image Models. In *ICLR*, 2024.
- [8] C. Li, C. Gokmen, G. Levine, R. Martín-Martín, S. Srivastava, C. Wang, J. Wong, R. Zhang, M. Lingelbach, J. Sun, M. Anvari, M. Hwang, M. Sharma, A. Aydin, D. Bansal, S. Hunter, **K. Kim**, A. Lou, C. Matthews, I. Villa-Renteria, J. Tang, C. Tang, F. Xia, S. Savarese, H. Gweon, K. Liu, J. Wu, F.-F. Li. BEHAVIOR-1K: A Benchmark for Embodied AI with 1,000 Everyday Activities and Realistic Simulation. In *CoRL (oral)*, 2022.
- [9] B. Byrne, K. Krishnamoorthi, C. Sankar, A. Neelakantan, D. Duckworth, S. Yavuz, B. Goodrich, A. Dubey, A. Cedilnik, **K. Kim**. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. In *EMNLP-IJCNLP*, 2019.

Preprints

- [1] J. Kim, **K. Kim**, J. Tack, D. Lim, J. Shin. Scalable and Robust LLM Unlearning by Correcting Responses with Retrieved Exclusions. arXiv preprint arXiv:2509.25973, 2025.
- [2] A. Corso, **K. Kim**, S. Gupta, G. Gao, M. Kochenderfer. A Deep Reinforcement Learning Approach to Rare Event Estimation. arXiv preprint arXiv:2211.12470, 2022.

- [3] S. Roy, C. Brunk, **K. Kim**, J. Zhao, M. Freitag, M. Kale, G. Bansal, S. Mudgal, C. Varano. Using Machine Translation to Localize Task Oriented NLG Output. arXiv preprint arXiv:2107.04512, 2021.

Technical reports

- [1] D. Bindel, P. Chew, J. Hopcroft, **K. Kim**, C. Ponce. Finding Overlapping Communities From Subspaces. Technical Report, 2012.

TALKS

Self-Refining Language Model Anonymizers via Adversarial Distillation

Jul 2025

University of Washington

Multithreading in Julia: An Anecdote

Feb 2022

Stanford Intelligent Systems Laboratory

SERVICE

Reviewer

- Conferences: ICLR 2026, AAAI 2026, NeurIPS 2025, ICML 2025, ICLR 2025 (notable reviewer)
- Journals: TPAMI 2025