Kyuyoung Kim

kykim@cs.stanford.edu
kykim0.github.io

Education

KAIST AI, Seoul, South Korea
Ph.D., Artificial Intelligence Sep 2023 - Present

- Research focus: Generative models, reinforcement learning, AI safety, nature-inspired intelligence
- Advisor: Prof. Jinwoo Shin
- Expected graduation: June 2027

University of Washington, Seattle, WA, USA
Visiting Researcher, Paul G. Allen School of Computer Science and Engineering Aug 2025 - Present

- Multi-turn reinforcement learning, LLM agents, social intelligence
- Advisor: Prof. Sewoong Oh

Stanford University, Stanford, CA, USA
M.S., Computer Science with Distinction in Research

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

Cornell University, Ithaca, NY, USA
B.S., Computer Science with Distinction

- Graduated *magna cum laude*
- Minor in Applied Mathematics

Research Experience

KAIST AI, Seoul, South Korea
Research Assistant
Algorithmic Intelligence Lab May 2023 - Present

- Researched RLHF methods for safe and effective alignment of generative models
- Researched applications of LLMs in AutoML, privacy, and agentic systems
- Advised by Prof. Jinwoo Shin

Stanford University, Stanford, CA, USA
Research Assistant
Stanford Intelligent Systems Lab Sep 2021 - Dec 2022

- Researched RL methods for efficient risk estimation of black-box systems
- Applied these to validate autonomous vehicle policies
- Advised by Prof. Mykel Kochenderfer

Stanford Vision and Learning Lab Mar 2022 - Dec 2022

- Developed a high-fidelity simulation benchmark for the design and evaluation of embodied AI
- Utilized the simulation environment to generate datasets for computer vision research
- Advised by Prof. Fei-Fei Li and Prof. Jiajun Wu

Work Experience

Google, Mountain View, CA, USA
Senior Software Engineer
Google Assistant Oct 2016 - April 2021

- 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
- Led system integration to deploy models in Search and Assistant
- 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 a backend engineer
- Designed logging infrastructure, enhanced matching algorithms, and improved user modeling

Display Ads June 2013 - Nov 2015

- Contributed to Gmail monetization initiatives as a backend engineer
- Proposed and implemented key improvements to ad auction algorithms
- Built backend systems, including an ads server, data pipelines, and production monitoring tools

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
Prof. Sewoong Oh's lab at the 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