

KYUYOUNG KIM

kykim144@gmail.com

kykim0.github.io

EDUCATION

KAIST AI, Seoul, South Korea

Ph.D., Artificial Intelligence

Aug 2023 - Present

- Advisor: Prof. Jinwoo Shin
- Research areas: language models, post-training, reasoning, agentic systems, AI safety
- Expected graduation: June 2027

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

Algorithmic Intelligence Lab

May 2023 - Present

- Advisor: Prof. Jinwoo Shin
- Researching post-training methods for developing reliable and efficient language models.
- Investigated reasoning, personalization, and agentic applications of language models.

University of Washington, Seattle, WA, USA

Visiting Researcher

Aug 2025 - May 2026

- Advisor: Prof. Sewoong Oh
- Developed an RL method with process supervision for training language models with robust reasoning.
- Proposed evaluation methods for assessing the faithfulness of RL-induced reasoning.

Stanford University, Stanford, CA, USA

Research Assistant

Stanford Intelligent Systems Lab

Sep 2021 - Dec 2022

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

WORK EXPERIENCE

Meta, Menlo Park, CA, USA

Research Scientist Intern

Meta Superintelligence Labs

Jun 2026 - Oct 2026

- Developed an RL method for training models to generate executable retrieval programs for agentic search.
- Introduced execution-verified program mutation to facilitate exploration of retrieval strategies.

Google, Mountain View, CA, USA

Senior Software Engineer

Google Assistant

Oct 2016 - Apr 2021

- Developed data-to-text generation models and built large-scale pipelines for training and evaluation.
- Led localization efforts to extend support for low-resource languages using ML models.
- 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, launching it across the United States and Israel.
- Designed logging infrastructure and improved matching algorithms and user modeling.

Display Ads

Jun 2013 - Nov 2015

- Developed the serving infrastructure for Gmail monetization and enhanced core auction algorithms.

PUBLICATIONS

Conference proceedings

- [1] **K. Kim**, Z. Xiao, Z. Wang, A. Atreya, K. Yang, J. Shin, K. Hao. Learning to Program Retrievals: Composing Search Primitives via Reinforcement Learning. Under review, 2026.
- [2] **K. Kim**, P. Sheng, A. Hebbar, P. Xu, Y. Xie, K. Wang, R. Xin, C. Wei, Z. Wang, J. Shin, P. Viswanath, S. Oh. Representation-Aligned Auxiliary Supervision for Language Model Adaptation. Under review, 2026.
- [3] **K. Kim**, K. Wang, Y. Xie, P. Xu, P. Sheng, C. Wei, Z. Wang, J. Shin, P. Viswanath, S. Oh. Correct Answers from Sound Reasoning: Verifiable Process Supervision for Language Models. In *COLM*, 2026.
- [4] H. Jeon, **K. Kim**, J. Shin. RedacBench: Can AI Erase Your Secrets? In *ICLR*, 2026.
- [5] **K. Kim**^{*}, H. Jeon^{*}, J. Shin. Self-Refining Language Model Anonymizers via Adversarial Distillation. In *NeurIPS*, 2025.
- [6] **K. Kim**, J. Shin, J. Kim. Personalized Language Models via Privacy-Preserving Evolutionary Model Merging. In *EMNLP (oral)*, 2025.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] **K. Kim**^{*}, A. Seo^{*}, H. Liu, J. Shin, K. Lee. Margin Matching Preference Optimization: Enhanced Model Alignment with Granular Feedback. In *EMNLP Findings*, 2024.
- [11] **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.
- [12] 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.
- [13] 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:2509.25973, 2025.
- [2] A. Corso, **K. Kim**, S. Gupta, G. Gao, M. Kochenderfer. A Deep Reinforcement Learning Approach to Rare Event Estimation. 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:2107.04512, 2021.

Technical reports

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

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

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

HONORS & SERVICE

Honors

- Oral presentations: EMNLP 2025, CoRL 2022
- Qualcomm Innovation Fellowship Korea 2025 Finalist

Reviewer

- Conferences: AAAI 2026-2027, ACL 2026, COLM 2026, ICLR 2025-2027, ICML 2025-2026, NeurIPS 2025-2026
- Journals: TPAMI 2025