Kyuyoung Kim

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

Education

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

- Research interests: generative models, RLHF, AI safety, nature-inspired intelligence.
- Advisor: Prof. Jinwoo Shin

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

- Reinforcement learning from human feedback for safe and effective alignment of generative models.
- Large language models and their applications in AutoML and autonomous agents.
- Advised by Prof. Jinwoo Shin.

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

- Reinforcement learning methods for efficient risk estimation of black-box systems.
- Application of RL methods in validating autonomous vehicle policies.
- Advised by Prof. Mykel Kochenderfer.

Stanford Vision and Learning Lab — Mar 2022 - Dec 2022

- Development of a simulation benchmark for designing and evaluating embodied AI solutions.
- Use of the simulation benchmark to synthesize datasets for computer vision research.
- Advised by Prof. Fei-Fei Li and Prof. Jiajun Wu.

Cornell University, Ithaca, NY, USA
Research Assistant
Finding overlapping communities from subspaces — Oct 2010 - May 2012

- Spectral approaches to finding overlapping community structures.
- Received research funding from Cisco.
- Advised by Prof. David Bindel and Prof. John Hopcroft.

Citation recommendation system — Jan 2011 - May 2011

- Bayesian approaches to document classification.
- Advised by Prof. Thorsten Joachims.

Work Experience

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

- Tech lead in the effort to localize Google Assistant and support low-resource languages.
- Researched deep learning approaches to data-to-text problems, developing key model components.

- With the Translation team, built a large-scale training pipeline for training diverse NMT models.
- Led the system integration effort to serve models in Search and Assistant.
- Built a human-in-the-loop dialogue system to collect diverse, task-oriented datasets for NLU research.

Waze Carpool — Nov 2015 - Oct 2016

- Backend engineer in the effort to build an on-demand ride sharing platform.
- Built the logging infrastructure, enhanced matching algorithms, and improved user modeling.

Display Ads — June 2013 - Nov 2015

- Backend engineer in efforts to monetize Google services such as Gmail.
- Proposed and implemented key auction improvements for Gmail ads.
- Built backend components including the ads server, data pipelines and production monitoring.

Facebook, Menlo Park, CA, USA
Software Engineer Intern — Aug 2012 - Nov 2012

- Backend engineer in the Messaging backend team.
- Designed and implemented NoSQL database to reduce log access latency using Apache HBase.

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**, J. Shin, J. Kim. Personalized Language Models via Privacy-Preserving Evolutionary Model Merging. In progress.

[2] 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.

[3] 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*, Dec. 2024.

[4] **K. Kim***, A. Seo*, H. Liu, J. Shin, K. Lee. Margin Matching Preference Optimization: Enhanced Model Alignment with Granular Feedback. In *EMNLP Findings*, Nov. 2024.

[5] **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.

[6] 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)*, Mar. 2022.

[7] 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*, Nov. 2019.

Preprints

[1] 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.

[2] 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

Multithreading in Julia: An Anecdote Feb 2022
Stanford Intelligent Systems Laboratory.

Service

Reviewer: ICML 2025, ICLR 2025

Honors and Awards

Google Display Network Innovation Award Q2 2013
Awarded to an innovative project within the Display Ads org.

Cornell Engineering Research Award 2011
Awarded funding for the research project on finding overlapping communities.

Morgan Stanley Award for Innovation 2011
For the research project on citation recommendation system.

John S. Knight Institute Award 2009
For the essay *Intrinsic and Instrumental Values*.