Taeyoun Kim

danielkty96@gmail.com https://danielkty.github.io Google Scholar

Education

MS **Carnegie Mellon University** Machine Learning — Pittsburgh, PA
Aug. 2023 - Dec. 2024

- Funded by the Kwanjeong Educational Foundation

BSE **Yonsei University** Electrical & Electronic Engineering — Seoul, South Korea
March 2016 - Feb. 2022

- High Honors: Top 3%
- Thesis: Impact of Different Joints on Creating a 3D Hand Mesh
- Funded through the National Science and Technology Scholarship (South Korean Government)

Publications

Reasoning as an Adaptive Defense for Safety
Taeyoun Kim, Fahim Tajwar, Aditi Raghunathan, Aviral Kumar
Preprint (arXiv)

Mitigating Bias in RAG: Controlling the Embedder
Taeyoun Kim, Jacob Springer, Aditi Raghunathan, Maarten Sap
ACL Findings 2025 (arXiv)

Testing the Limits of Jailbreaking Defenses with the Purple Problem
Taeyoun Kim*, Suhas Kotha*, Aditi Raghunathan
NeurIPS 2024 Safe GenAI (arXiv)

Predicting the Performance of Foundation Models via Agreement-on-the-Line
Rahul Saxena*, **Taeyoun Kim***, Aman Mehra*, Christina Baek, Zico Kolter, Aditi Raghunathan
NeurIPS 2024 (arXiv)

The Application of Local Sub-voxel Shifting on Multi-echo GRE-based Myelin Water Imaging
Taeyoun Kim, Muyul Park, Jaeuk Yi, Dong-Hyun Kim
ICMRI 2021 (Oral)

Research Experience

CMU, Research Assistant/Masters Research — Pittsburgh, PA
Aviral Kumar, Aditi Raghunathan — June 2023 - Current

- **Reasoning/Safety/ML** Created an online RL training recipe for LLM safety that trains models to adaptively reason by thinking shorter on clear-cut harmful prompts and longer on ambiguous prompts. Beat larger open-weight models such as Llama and SOTA defenses such as circuit breakers.
- **RAG/Bias/ML** (collab. Maarten Sap) Found that representational bias in RAG is best mitigated by reverse-biasing the embedder. Revealed that light fine-tuning a small 109M embedder through PEFT or WiSE-FT can overcome bias in Llama 3.1 405B Instruct. Empirically estimated the bias sensitivity of LLMs (Llama, Gemma, Mistral) and found that LLMs have different sensitivity to political bias.
- **Jailbreaking/Adversarial Robustness/ML** Constructed the Purple Problem to test if jailbreaking defenses can prevent the generation of a single word: **purple**. Broke existing defenses such as DPP on Advbench harmful behaviors to 1.7% Defense Success Rate.
- **Out-of-distribution Robustness/ML** Used randomly initialized heads during fine-tuning in Foundation models to exhibit Agreement-on-the-Line to out-of-distribution

shifts. Estimated out-of-distribution performance for different model families (i.e., BERT, GPT, OPT, Llama, Alpaca, Vicuna) using Agreement-on-the-Line and reduced the mean absolute percentage error to 1.64% on SQuAD-Shifts (SOTA: 2.8%).

Yonsei Esports (YES) Lab, Research Assistant — Seoul, South Korea
Byungjoo Lee — July 2022 - June 2023

- **Human Modeling/RL/HCI** Modeled human point-and-click behavior through N-step TD SAC to understand human motor and visual control with the BUMP model. Implemented human foveal vision as inputs to vision models.

Yonsei MILAB, Research Assistant — Seoul, South Korea
Dong-Hyun Kim — June 2021 - Sept. 2021

- **MRI** Reduced Gibbs-artifacts in mGRE-based MWF mapping via local sub-voxel shifting by creating an exponential filter. Reduced the problem of blurring during artifact removable compared to Tukey filtering (SOTA).

Teaching Experience

TA (PhD) Advanced Introduction to Machine Learning (10-715) — Fall 2024

- Led recitation on convex optimization (bounds for GD and SGD, duality, Slater's conditions, KKT conditions), held office hours, helped make exams.

Awards/Fellowships

Kwanjeong Educational Foundation Fellowship	2023, 2024
National Science and Technology Scholarship	2018, 2021
Yonsei Veritas Scholarship	2016, 2017
High Honors	2016, 2017, 2018, 2021
Honors	2016, 2018, 2021
1st Place, Yonsei EE Autonomous Race Competition	2017

Involvement/Service

President, CMU KGSA Soccer — Pittsburgh, PA
Oct. 2023 - Dec. 2024

- Organized soccer games as part of the Korean Graduate Students Association.

Member, Yonsei Tea — Seoul, South Korea
March 2023 - July 2023

- Educated, learned, and spread Eastern tea drinking culture.

Sergeant, Reconnaissance, Republic of Korea Army — Gang-wondo, South Korea
July 2019 - Jan. 2021

- Military service near the De-Militarized Zone (DMZ).

President/Member, Yonsei AFKN Listener's Club (ALC) — Seoul, South Korea
March 2016 - June 2019

- President in 2018. Taught exchange students Korean language and culture. Created study sessions and games for language exchange. Gave tours around Seoul. Took part in school festivals and joint events with other ALC organizations.

Language

GRE 165/170/5.0	July 2, 2022
TOEFL 117 (30/29/28/30)	July 11, 2021

Selected Courses

(PhD) Theoretical and Empirical Foundations of Modern Machine Learning, (PhD) Deep Reinforcement Learning and Control, (PhD) Probabilistic Graphical Models, (PhD) Convex Optimization, (PhD) Advanced Introduction to Machine Learning, (PhD) Machine Learning in Practice, Intelligent Control