

# Taeyoun Kim

✉ danielkty96@gmail.com    🌐 Website    📄 Google Scholar

## Education

---

- MS**    **Carnegie Mellon University** Machine Learning    Pittsburgh, PA  
Aug. 2023 - Current
- Expected Dec. 2024
  - Funded by the Kwanjeong Educational Foundation
- BSE**    **Yonsei University** Electrical & Electronic Engineering    Seoul, South Korea  
March 2016 - Feb. 2022
- GPA: 3.92/4.0 (4.10/4.3)
  - 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

---

### Testing the Limits of Jailbreaking Defenses with the Purple Problem

Taeyoun Kim\*, Suhas Kotha\*, Aditi Raghunathan

NeurIPS 2024 Safe GenAI (ICLR 2025 under review)

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**, Masters Research    Pittsburgh, PA  
Aditi Raghunathan    June 2023 - Current
- **RAG/Bias/NLP** Mitigated bias in the entire RAG system by reverse-biasing the embedder. Revealed that light fine-tuning a small 109M embedder can overcome bias in Llama 3.1 405B Instruct. Empirically estimated the sensitivity of LLMs (Llama / Gemma / Mistral) to change in bias and found that LLMs have different sensitivity to political bias. (collab. w/ Maarten Sap)
  - **Jailbreaking/Adversarial Robustness/ML** Constructed the Purple Problem to test if jailbreaking defenses can prevent the generation of a single word: **purple**. Broke all defenses under the Purple Problem through GCG adaptive attacks by better initialization and more compute. 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

Byungjoo Lee

Seoul, South Korea  
July 2022 - June 2023

- **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**, Undergraduate Researcher  
Dong-Hyun Kim

Seoul, South Korea  
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).

**Yonsei Multi-dimensional Insight Lab**, Undergraduate Researcher  
Sanghoon Lee

Seoul, South Korea  
Spring 2021

- **AR/Vision** Reconstructed hand meshes with features extracted from HoloLens 2 through PointNet. Optimized training for supervised learning by analyzing hand features.

## 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 - Current

National Science and Technology Scholarship

2021, 2018

Yonsei Veritas Scholarship

2017, 2016

High Honors

2021, 2018, 2017, 2016

Honors

2021, 2018, 2016

1st Place, Yonsei EE Autonomous Race Competition

2017

## Involvement/Service

---

**President, CMU KGSA Soccer**

Pittsburgh, PA

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

Oct. 2023 - Current

**Member, Yonsei Tea**

- Educated, learned, and shared knowledge of Eastern tea drinking culture.

Seoul, South Korea

March 2023 - July 2023

**Sergeant, Reconnaissance, Republic of Korea Army**

Gang-wondo, South Korea

- Mandatory military service near the Demilitarized Zone (DMZ).

July 2019 - Jan. 2021

**President/Member, Yonsei AFKN Listener's Club (ALC)**

Seoul, South Korea

- 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.

March 2016 - June 2019

## 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) Probabilistic Graphical Models, (PhD) Convex Optimization, (PhD) Advanced Introduction to Machine Learning, (PhD) Machine Learning in Practice, (PhD) Deep Reinforcement Learning and Control, Intelligent Control