

Sunwoo Kim

Daejeon, South Korea | sunwoo.kim@kaist.ac.kr | +82 010-3293-9692 | [linkedin.com/in/sunwookim1214](https://www.linkedin.com/in/sunwookim1214)

About

Sunwoo Kim is a Master's student majoring in Computer Science at KAIST, advised by professor Alice Oh. His research interests lie in agentic AI safety and reliability. He is especially interested in uncertainty quantification, mechanistic interpretability, and robust evaluation methods.

Education

KAIST, MS in Computer Science Mar 2026 – present

- **Coursework:** Special Topics in Artificial Intelligence <Trustworthy Machine Learning> (A+), AI Ethics (A+, 1st/66)

KAIST, BS in Computer Science Sept 2017 – Feb 2026

- **Coursework:** Artificial Intelligence and Machine Learning (A0), Computer Ethics & Social issues (A+)

American School of Paris, Upper School Sept 2014 – June 2017

- **Activities:** National Honor Society

Experience

Teaching Assistant, CS101 <Introduction to Programming> – KAIST Feb 2026 – June 2026

Individual Study, U&I Lab – KAIST Dec 2024 – Aug 2025

- Researched AI safety and alignment, specifically using mechanistic interpretability methods on LLMs.

Research Intern, RIRO Lab – KAIST June 2024 – Sep 2024

- Participated in robotics research leveraging large multi-modal models for real-time scene-graph generation and environment awareness.

Military Service, Air and Space Combat Development Wing – ROK Air Force Feb 2021 – Nov 2022

- Served as an enlisted person and worked as a translator.

Publications

On the Effect of Uncertainty on Layer-wise Inference Dynamics July 2025

Sunwoo Kim, Haneul Yoo, Alice Oh (ICML, Actionable Interpretability Workshop)

- Investigated how epistemic uncertainty affects LLM inference by leveraging a mechanistic interpretability method.

PapersPlease: A Benchmark for Evaluating Motivational Values of Large Language Models Based on ERG Theory (Oral) July 2025

Junho Myung*, Yeon Su Park*, *Sunwoo Kim**, Shin Yoo, Alice Oh (ACL, Generation, Evaluation & Metrics Workshop)

- Created a benchmark dataset to test LLM alignment with the human values hierarchy based on ERG theory using a role-playing framework.

Skills

Programming Languages: Python, C

Languages: English (native, TOEFL 113), Korean (native), French (novice)