Hwiyeong Lee

hwi0lee.github.io

Email : hyglee@hanyang.ac.kr
Mobile : +82-10-9037-4778

Research Interest

My research focuses on how LLMs acquire, represent, and update knowledge, and how we can manipulate it post hoc. I treat LLMs as scientific objects of study, using **Mechanistic Interpretability** to uncover their internal structure and address practical challenges in **Knowledge Editing** and **Unlearning**.

Education

Hanyang University *Sep. 2024 – Present*
M.S. in Artificial Intelligence Semiconductor Engineering

Seoul National University of Science and Technology *Mar. 2018 – Aug. 2024*
B.S. in Industrial and Information Systems Engineering

Experience

Graduate Researcher (*Advisor: Taeuk Kim*) *Sep. 2024 – Present*
Hanyang University

- **Causal Validation of Applying Localization to Unlearning**: Conducted research on whether updating only neurons associated with the target knowledge is sufficient for effective unlearning in LLMs, using controlled experiments to test whether localization success causally translates into unlearning success and challenging the assumption that parameter locality is inherently indicative of successful knowledge removal.

Research Intern (*Advisor: Sangheum Hwang*) *Sep. 2023 – Aug. 2024*
Seoul National University of Science and Technology

- **Adversarial Attacks and Defenses for Robust Unlearning**: Conducted research on robust unlearning for image classifiers by constructing adversarial attacks that reveal residual classification ability on target data after unlearning, and showing that pruning the network into a sparse model makes the unlearned classifier more robust to such attacks.

Publications

Does Localization Inform Unlearning? A Rigorous Examination of Local Parameter Attribution for Knowledge Unlearning in Language Models *Nov. 2025*
Hwiyeong Lee, Uiji Hwang, Hyelim Lim, Taeuk Kim
The 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)

Uncovering Hidden Vulnerabilities in Machine Unlearning: Adversarial Attack as a Probe and Pruning as a Solution *Jun. 2024*
Hwiyeong Lee, Jeonghyun Kim, Sangheum Hwang
Korea Computer Congress 2024 (KCC 2024)

Teaching

Teaching Assistant, Object-Oriented Systems Design *Spring 2025*
Department of Computer Science, Hanyang University

Skills

Languages: Python, C, Java, Kotlin, JavaScript

Technologies: PyTorch, HF Transformers, TransformerLens, SAELens, Linux Internals, React, Figma

Language Proficiency

English: Professional Working Proficiency (ETS TOEIC 915, SNU TEPS 425)

Korean: Native Proficiency