Jimin Yeom | Curriculum Vitae

sion0107@hanyang.ac.kr • jiminyeom.com
• LinkedIn: https://www.linkedin.com/in/jimin-yeom-a5435b33b/

Research Interests

My research goal is to build **trustworthy and interpretable AI by understanding how models learn**. Concretely, I study the optimization problem

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\mathcal{L}\big(f(x+\delta,\ \theta+\eta),\ y\big)\Big]$$

which unifies **adversarial training** (perturbations δ in input space) and the study of **training dynamics** (perturbations η in parameter space). I view this joint objective as a lens through which robustness, generalization, and interpretability can be analyzed together, rather than as separate phenomena.

Trustworthy AI: *Adversarial Robustness* – Robustness–Accuracy Trade-off, Catastrophic Overfitting; *LLM Jailbreaking* – Certifiable Defense.

Interpretable AI: *In-Context Learning*, *Linear Transformer*, *Model Unlearning*, *Reasoning*.

Education

Hanyang University — **Seoul, Republic of Korea**
M.S. in Computer Science — *Sep. 2025 – Aug. 2027 (expected)*
Department of Computer Science & Software
Advisor: *Sungyoon Lee*

Hanyang University — **Seoul, Republic of Korea**
B.S. in Computer Science — *Mar. 2018 – Aug. 2025*
Department of Computer Science & Software

Publications

* denotes equal contribution. All papers listed below are currently **under review**.

[Paper 1] — *Under Review*
Jimin Yeom, Jonghyun Hong, *Sungyoon Lee*.
Keywords: AI Safety, LLM Jailbreaking, Optimization.

[Paper 2] — *Under Review*
Jimin Yeom, *Sungyoon Lee*.
Keywords: AI Safety, Adversarial Robustness, Adversarial Training.

[Paper 3] — *Under Review*
Subin Jang, **Jimin Yeom**, *Sungyoon Lee*.
Keywords: AI Safety, LLM Jailbreaking, Certified Defense.

Ongoing Work

Edge of Stability in Adversarial Training — *2025 – present*
Investigating how the Edge-of-Stability regime manifests under adversarial training and how it interacts with catastrophic overfitting.

Reward Hacking — *2025 – present*
Studying the mechanisms behind reward hacking in alignment-tuned models, with the goal of mitigating it via training-dynamics-aware objectives.