Synthetic Data Generation for LLM Safeguards

Seanie Lee (KAIST)

SEMINAR ANNOUNCEMENT

- **Date & Time :** ***Apr. 9, 2025, 10:00~11:30***
- **Place :** ***Room 502, ITBT Bldg., Hanyang University***
- **Abstract:** As AI systems become increasingly integrated into critical applications, ensuring their safety and robustness is crucial. This talk focuses on synthetic data generation for red-teaming large language models (LLMs) and knowledge distillation of safety guard models, aiming to improve AI safety with efficiency. In the first half of this talk, I will introduce an automatic red-teaming framework where a small language model is fine-tuned with GFlowNet to generate prompts that elicit harmful responses from target LLMs. This framework enables the discovery of diverse and effective adversarial prompts. In the second half of the talk, I will present an LLM prompting-based data augmentation method for distilling LLM-based safety guard models, which detect and block malicious queries targeted at LLMs, into a smaller model.
- **Bio:** Seanie Lee received his Ph.D. from KAIST (2022–2025), where he also completed his M.S. (2020–2022), under the supervision of Professors Sung Ju Hwang and Juho Lee. He earned his B.A. from Yonsei University (2011–2018). Seanie gained broad research experience through internships at Mila (Jan–Jun 2024), Apple (May–Sep 2023), and the National University of Singapore (Jun–Sep 2022).

Contact : kimtaeuk@hanyang.ac.kr **(컴퓨터소프트웨어학부 김태욱 교수)**