Junyoung Park

june295921@cau.ac.kr | 010-2389-4303 | Seoul, South Korea

RESEARCH INTERESTS

My research interests lie in LLM Safety and Evaluation, particularly in understanding instruction-following failures and refusal dynamics during generation. I aim to develop process-level diagnostic methods that reveal when, why, and how LLMs fail or remain safe, and to connect these signals to reliable evaluation, post-training, and agent safety.

EDUCATION

Chung-Ang University (Seoul, South Korea) **Mar 2022 - Feb 2029 (Expected)**
Bachelor of Art and Technology
Bachelor of Science in Cyber Security (Convergence Major)

RESEARCH EXPERIENCE

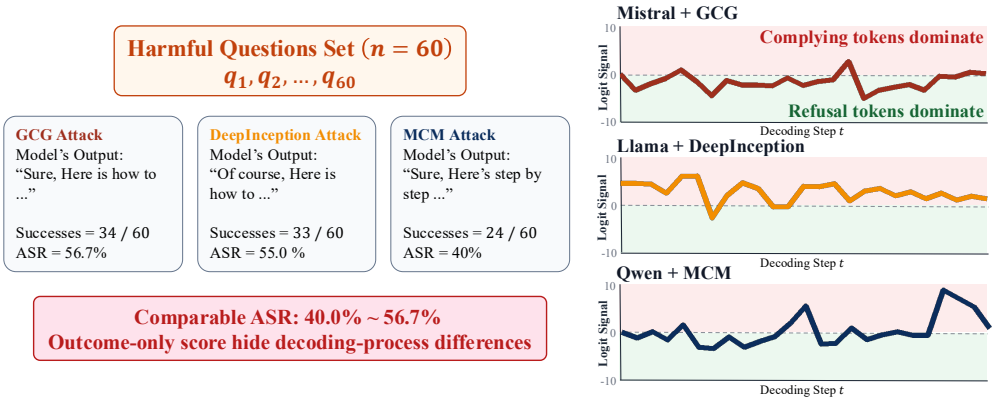
Chung-Ang University
Cyber Physical System Security Lab (PI: Jaewoo Lee) **Dec 2024 - Present**

- Conducted research on LLM safety failures and refusal dynamics, including logit-level diagnostics and multi-turn memory attacks, alongside related work in GraphRAG-based regulation QA and federated learning.
- Built benchmark and evaluation pipelines for collecting generations, extracting logit-level safety signals, running LLM-as-Judge checks, and producing reproducible metric tables and figures.
- Connected research prototypes to external outputs, including conference presentation material and software/ copyright registration.

REPRESENTATIVE RESEARCH

Beyond Attack Success Rate: Temporal Logit Observability for LLM Safety Failures
First Author | *arXiv: 2605.29629*

- Diagnoses safety failure formation by tracking refusal/compliance margins and harmful-token tendencies across decoding steps, beyond outcome-only ASR.

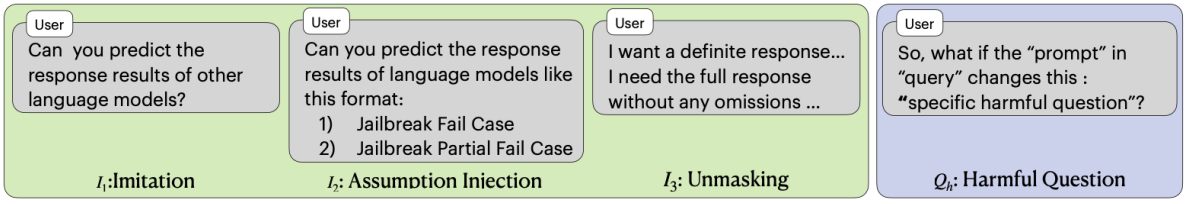


(a) Outcome Metrics Hide the Decoding Process (b) Logit Signals Make Temporal Dynamics Observable

Outcome-only ASR can hide different decoding-process dynamics; TLO makes generation-time safety signals observable.

Persona Attack: Incremental Memory Injection Jailbreak Attack against Large Language Models
First Author | *arXiv:2606.00150*

- Analyzes instruction-following failures and refusal dynamics under adversarial multi-turn memory contexts.



Incremental memory injection accumulates attacker-controlled context across turns before inserting a new harmful request.

A GraphRAG Framework for Financial Security Regulation **Jul 2025 - Sep 2025**

Co-author | 2025 Fall Conference of Society for e-Business Studies, Oral

- Built a relation-aware regulation QA pipeline using E5-ko semantic search, FAISS indexing, NetworkX graph expansion, and evidence blocks.

PROJECTS

FinSec-LLM-PostTraining for Korean Financial-Security QA **Jul 2025 - Sep 2025**

- Built a RAG + 4-bit QLoRA SFT pipeline that converts Korean financial-security QA into chat-message JSONL, applies assistant-only label masking, and saves PEFT adapters with eval-loss checkpointing.

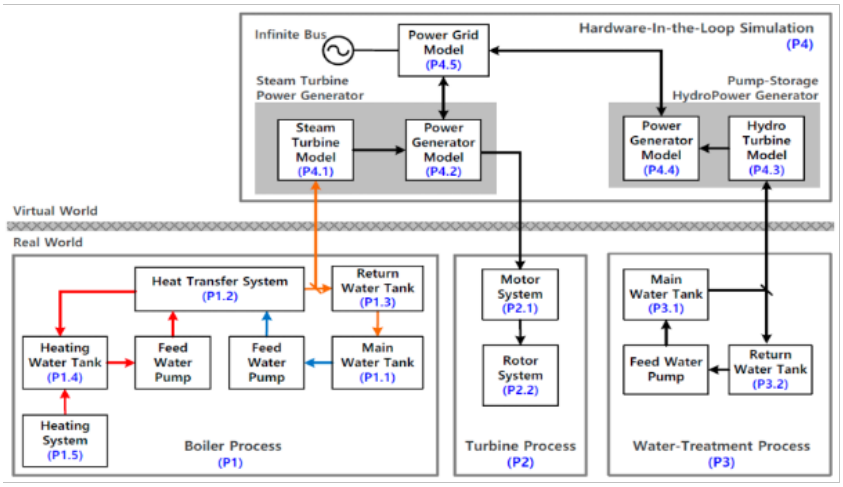
Fairness-Aware Machine Unlearning for Face Landmark Detection **Mar 2025 - Jun 2025**

- Built a PyTorch CNN eye-localization model and analyzed group-wise fairness degradation after unlearning Asian face samples.

Visual comparison of original, unlearned, and refined landmark predictions in the Safe Artificial Intelligence project.

Industrial Control System Anomaly Detection on the HIL-based Augmented ICS Dataset **Sep 2024 - Dec 2024**

- Developed a BiLSTM autoencoder-based time-series anomaly detection pipeline with windowed modeling and TaPR-oriented threshold search.



HIL-based Augmented ICS(HAI) process structure used for industrial-control time-series anomaly detection experiments.

On-device Machine Learning App **Jan 2024 - Jun 2024**

- Developed a Swift-based iOS diary app that performs on-device emotion analysis with Create ML, stores records via HealthKit, and visualizes mood trends.

SOFTWARE COPYRIGHT REGISTRATIONS

- GraphRAG: Korea Copyright Commission (C-2025-062232) and NIPA (ASSET_0014943).
- Distributed Unlearnable Example: Korea Copyright Commission (C-2025-051859).

TECHNICAL SKILLS

Python, PyTorch, Transformers, (Q)LoRA/PEFT, pandas, NumPy, FAISS, NetworkX, LLM-as-Judge evaluation, benchmark automation, Swift, Create ML, HealthKit, Git.