AI / SECURITY RESEARCH PORTFOLIO

AI·Security Research Portfolio

Benchmark 설계 · 평가 자동화 · 도메인 데이터 구조화로
모델 성능과 실패 원인을 설명하는 연구 경험

박준영 Junyoung Park

Chung-Ang University · Cyber Physical System Security Lab

june295921@cau.ac.kr

Portfolio Snapshot

Research Focus

Trustworthy AI | Safe AI | LLM Failure Analysis

Trustworthy & Safe AI

Understanding when, where, and how LLMs fail
before failures appear in final outputs

LLM Failure Observability

Logit dynamics · refusal / compliance margins
early-token behavior · evaluation signals

Applied AI Systems

GraphRAG for financial security regulation
privacy-aware AI · local agent runtime

ABOUT ME

모델 성능을 점수 너머의 데이터·조건·실패 원인으로 해석합니다.

단순히 모델을 구현하는 것보다, **어떤 조건에서 성능이 안정되고 왜 흔들리는지를 재현 가능한 평가로 설명**하는 데 강점이 있습니다.

1. Problem Framing

ASR만으로 보이지 않는 LLM safety failure,
법령 QA에서 사라지는 조항 관계,
memory가 만드는 jailbreak 조건을 문제로 정의

TLO · Persona Attack · HAI · Machine Unlearning(SafeAI)

2. Evaluation Design

모델·데이터·공격/사용 조건을 분리하고,
LLM-as-Judge, logit signal, TaPR,
group-wise loss처럼 문제에 맞는 지표를 설계

1,440 generations · 자동화 pipeline

3. System & Evidence

TLO benchmark pipeline, GraphRAG retrieval framework, local LLM agent, iOS on-device ML app처럼 실제로 동작하는 구조로 구현

GraphRAG · HAI · Machine Unlearning(SafeAI)

Current Research Focus

LLM이 최종 답변에서 실패하기 전에, 그 실패가 어떤 데이터 조건·평가 조건·생성 과정에서 형성되는지 관찰합니다.
TLO에서는 logit dynamics와 refusal/compliance margin을, Persona Attack에서는 memory 기반 jailbreak 조건을,
GraphRAG에서는 도메인 관계 구조가 QA 성능과 근거 추적성에 미치는 영향을 분석했습니다.

RESEARCH OVERVIEW

프로젝트들은 서로 다른 도메인에서 같은 질문을 다룹니다.

데이터가 어떤 구조로 주어지고, 평가 조건이 어떻게 설계될 때 모델 행동이 달라지는가?

Core Research

TLO

NeurIPS 2026 Under Review

Logit-only LLM safety benchmark

Persona Attack

NeurIPS 2025 Submitted

Multi-turn memory injection 평가

GraphRAG

학회 발표 Oral Accepted

조항·관계 기반 규제 QA

Supporting Studies & Outputs

HAI ICS

산업제어 시계열 이상탐지

SafeAI

Privacy-preserving unlearning & fairness

Friday Agent

로컬 안전 실행 assistant runtime

iOS Emotion Diary

On-device ML 감정 분석 앱

REPRESENTATIVE RESEARCH 01

TLO: 최종 ASR 뒤의 Safety Failure 과정을 보는 Logit-Only Diagnostic

Challenge: 같은 Attack Success Rate라도 모델의 safety가 무너지는 시점과 방식은 서로 다를 수 있습니다. | NeurIPS 2026 Main Paper Under Review (제1저자)

기존 Attack Success Rate(ASR) 평가

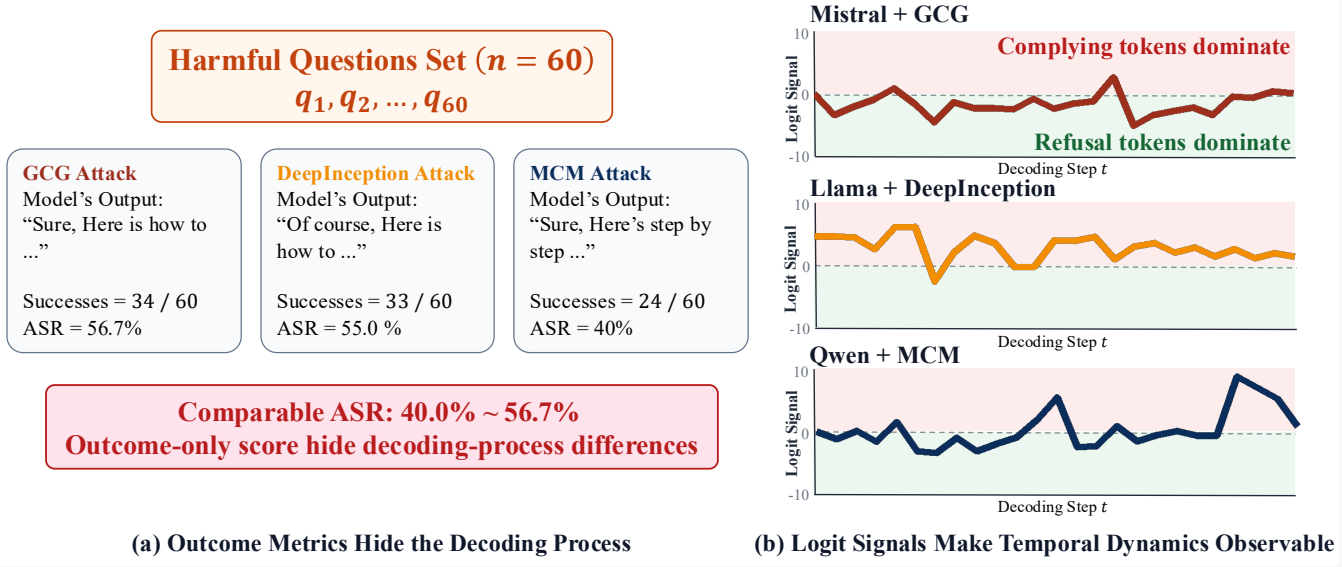
최종 응답만 보고 성공/실패를 판정

→ 언제부터 거절 신호가 약해졌는지는 남지 않음

Temporal Logit Observability(TLO)의 접근

Decoding step마다 거절·순응 logit margin 추적

Reference calibration으로 모델·공격 조건을 같은 좌표에서 비교



Logit-only

hidden state 접근 없이
safety signal 관찰

Temporal

최종 응답 이전의
failure path 비교

Calibrated

model/tokenizer 차이를
reference로 보정

Diagnostic

ASR을 보완해 원인을 설명하는
설명 가능성 제공

TLO는 ASR을 대체하는 지표가 아니라, **"언제, 어떤 축에서, 어떤 방식으로 안전 신호가 약해졌는지"를 보여주는 관찰 도구**입니다.

REPRESENTATIVE RESEARCH 01

Benchmark 제작과 평가 자동화를 하나의 Pipeline으로

반복 실험이 가능한 구조를 만들어, 같은 가설을 여러 모델·공격 조건에서 검증했습니다.

1 조건 설계

4 LLM × 3 attacks
12 model-attack pairs
harmful + benign reference

2 생성 수집

총 1,440 generations
720 harmful + 720 benign
batch inference pipeline

3 자동 판정

Llama-Guard primary
보조 judge / human audit
HarmBench, GPT-4o 교차검증

4 지표·도표

Logit Margin Score
RP-plane, t_cross
표·그림 자동 생성

Key Results

3.3%p

ASR 차이 — Llama+DI vs Qwen+DI

ASR은 비슷하지만 RP-plane에서
DRP 0.572로 분리 → 다른 failure mode

39.6% → 13.1%

ASR 감소 — Early-stop probe 적용

t_cross 기반 early-stop rule로
집계 ASR 26.5%p 감소

0.0%

False Positive — Format-free benign query

정상 질문을 공격으로 오판하지 않음
Plain benign 구간 false alarm 0건

Persona Attack: Multi-turn Memory가 Safety를 약화시키는가?

Challenge: 단일 prompt가 아니라 여러 턴에 누적된 memory가 안전 정책보다 우선될 수 있습니다. | NeurIPS 2025 Main Paper Submitted (제1저자)

공격 방식: Incremental Memory Injection

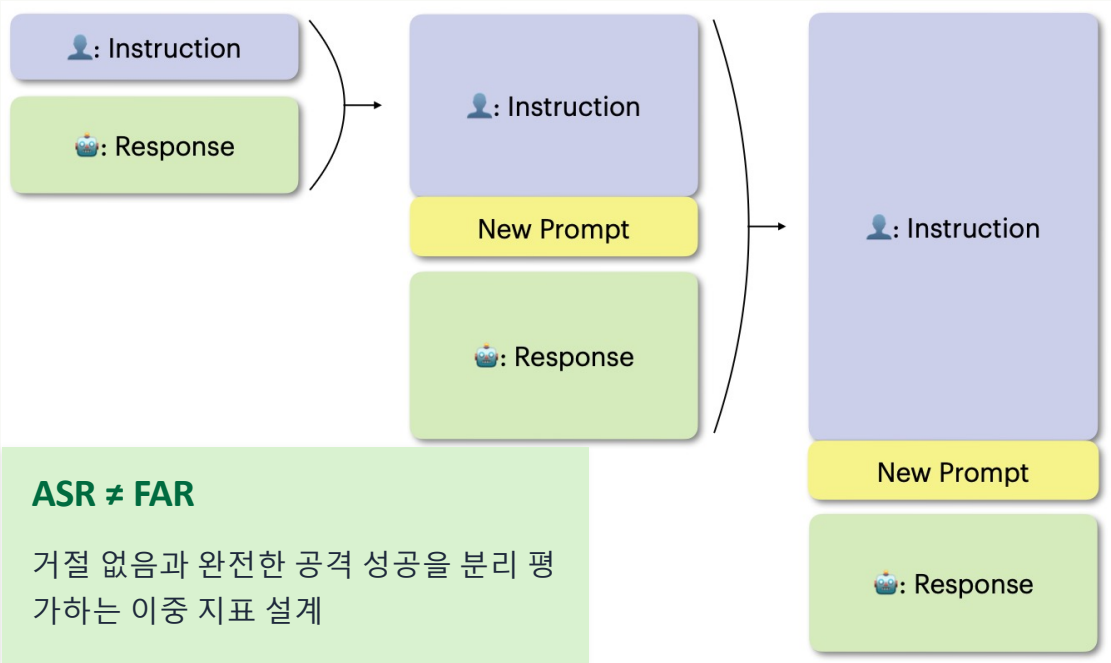
- Instruction imprinting:
  여러 턴에 걸쳐 응답 형식·역할을 주입
- Context framing:
  모델이 공격자의 작업 프레임을 수용하게 만듦
- Harmful request insertion:
  기존 대화 맥락 안에 삽입
- Memory type 비교:
  manual vs state-based memory 분리 평가

GPT-4o Sequential

95.0% ASR (Attack Success Rate)

Once: 75.0% ASR

AdvBench 520: FAR 84.61%



핵심 발견

Sequential > Once

순차적 memory injection이 한번에 모든 지시를 넣는 방식보다 강력

Memory 구현 차이

State-based memory가 manual memory보다 높은 ASR/FAR → memory 구현이 공격 표면

ASR ≠ FAR

거절 없음과 완전한 공격 성공을 분리 평가하는 이중 지표 설계

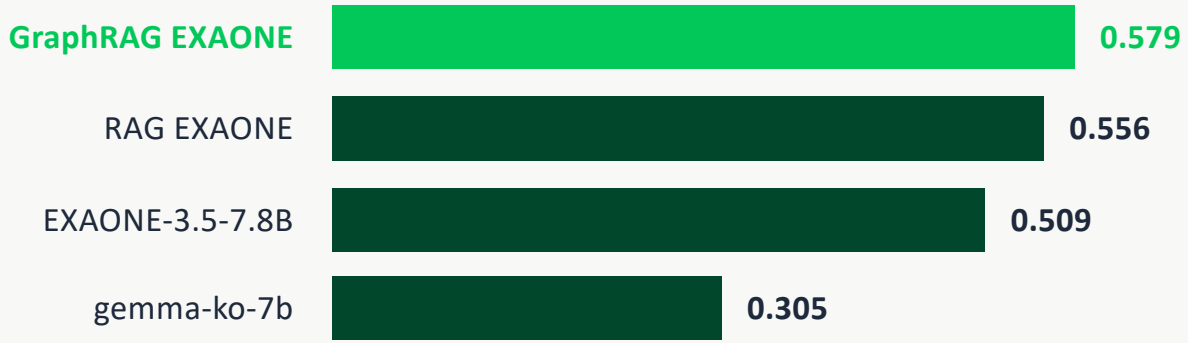
REPRESENTATIVE RESEARCH 03

GraphRAG: 도메인 데이터의 '관계 구조'를 모델 성능으로 연결

Challenge: 법령·규제·보안 문서는 관련 문장만 찾는 것보다 관계와 근거 추적성이 중요합니다. | 2025 한국전자거래학회 추계 학술대회 Oral (공저자)

구현 흐름

FSKU 벤치마크 비교 (~1,000 문항)



FSKU Evaluation Formula

Evaluation Score = 0.5 × Objective Accuracy + 0.5 × Subjective Score
Subjective Score = 0.6 × Semantic Similarity + 0.4 × Keyword Recall

왜 GraphRAG인가?

- 더 큰 LLM → 비용↑ / 원인 설명 약함
- 문서 더 많이 넣기 → context noise 위험
- 관계 기반 GraphRAG → 도메인 구조를 검색·생성 근거로 사용

SUPPORTING PROJECTS

평가 조건을 실제 사용 환경에 맞춰 넓히는 경험

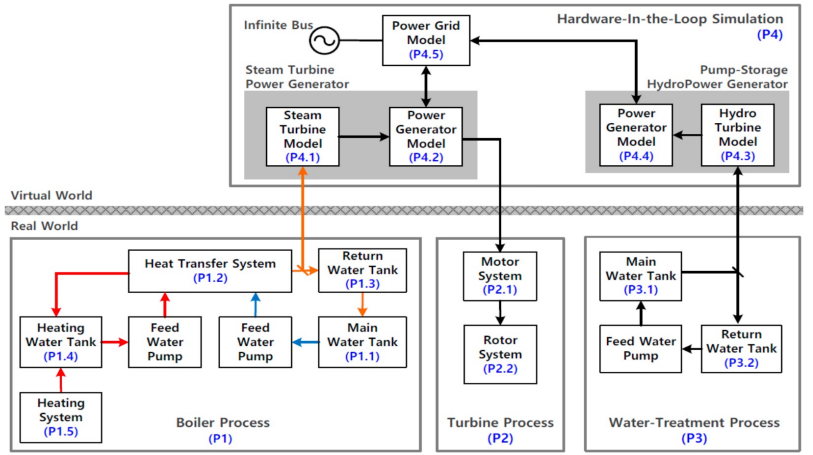
HAI ICS 이상탐지

문제 정의: 산업제어 데이터의 구간 단위 이상탐지와 시계열 특화 지표

접근: BiLSTM Autoencoder, RobustScaler, window 20, TaPR threshold search

F1 0.693, TaP 0.750, TaR 0.644

시계열 이상탐지 end-to-end pipeline



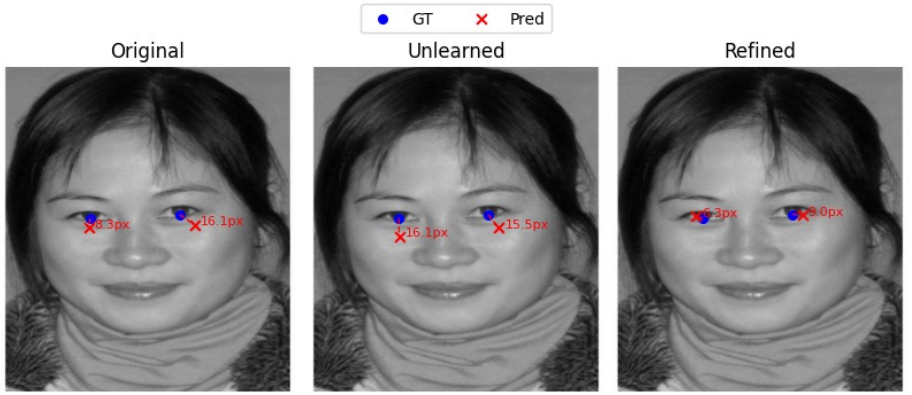
SafeAI: Is Unlearning Racist?

문제 정의: Privacy-preserving unlearning이 특정 집단 성능을 더 많이 손상시키는가?

접근: PyTorch CNN eye regression, Asian/Western group-wise loss, reweighting fine-tuning

Asian loss +81.6%, Western +42.8%

평균 점수 뒤의 group별 영향



ENGINEERING & PRODUCT SYSTEMS

아이디어를 시스템으로 — 논문뿐 아니라 로컬 에이전트, 앱, 등록 문서까지 전체 구현 경험

Friday Local Agent

~15,000 LOC Python

자연어가 파일 수정·Shell·Git 실행으로 이어질 때 필요한
안전한 local-first assistant runtime

3-tier model routing · state-machine planner · permission/sandbox
tool registry · paper library · 241개 skill index

iOS Emotion Diary

iOS prototype

민감한 일기 데이터를 외부 서버로 보내지 않고
감정 분석과 기록을 제공하는 **on-device ML privacy-aware** 앱

Swift · Create ML on-device · HealthKit 저장
gamification reward loop

GraphRAG 기술 자산화

SW 등록 + 저작권 등록

연구 결과를 논문에서 끝내지 않고, NIPA 소프트웨어 등록과
저작권 등록이 가능한 산출물로 문서화.
기술 개요·내부 아키텍쳐·개발 방법론·품질 검증 기준까지 포함

3-layer 시스템 명세(데이터/지식·검색/인덱스·생성/추론) ·FSKU ~1,000문항 기반
품질 검증 체계 · 적용 표준 · 시장 적용 가능성 문서화

TECH STACK & CONTRIBUTION

반복 가능한 평가 Framework: 데이터 준비부터 결과 해석까지 끊기지 않는 실험 루프

데이터·조건 설계

pandas & numpy
FAISS & NetworkX
benchmark input
도메인 구조

모델 실행

PyTorch
Transformers
TensorFlow/Keras
batch inference

신호 추출

tokenizer
logit extraction
reconstruction loss
failure signal

판정·지표

LLM-as-Judge
ASR
TaPR / group loss
metric table

시스템 구현

Swift / HealthKit
CLI / REPL
permission layer
privacy app

AI Safety / Evaluation

LLM safety failure를 최종 응답이 아니라
과정과 조건으로 분석

- Benchmark 설계 · evaluation pipeline 자동화
- Logit/temporal signal 분석
- Multi-turn / memory-aware safety 평가

Applied AI / Systems

도메인 데이터와 사용자 맥락을
실제 시스템 구조로 변환

- GraphRAG 규제 해석 · SW 등록
- Local agent runtime (~15K LOC)
- Privacy-aware iOS app · HealthKit 연동

End of Document

Understanding when, where, and how LLMs fail before failures appear in final outputs.

NeurIPS 2026 Under Review (TLO)

NeurIPS 2025 Submitted (Persona Attack)

GraphRAG 학회 발표 Oral + SW/저작권 등록

Local LLM Agent ~15K LOC

박준영 Junyoung Park

Chung-Ang University · Cyber Physical System Security Lab

june295921@cau.ac.kr