Hyunjun Kim

+41 076 268 52 54| hyunjun1121@kaist.ac.kr | **[Portfolio]** | **[Linkedin]** | **[Scholar]**

EDUCATION

KAIST (Korea Advanced Institute of Science and Technology) — *B.S., Computer Science* Mar. 2023 – (Expected) Aug. 2026
Daejeon, South Korea

- Dean's List (Top 2%)
- CoE Leadership Award (Mar. 2025)
- All courses conducted in English; planning Ph.D. in AI
- Relevant Coursework : **Artificial Intelligence and Machine Learning (A), AI & Simulation in Biological Sciences (A+), Intelligent Robot Design and Programming (A+), Data Structure (A-), System Programming (A-)**

RESEARCH EXPERIENCE

CREST - The University of Adelaide [Link] May.2024 - Feb.2025
Research Intern, *Adelaide, Australia*

- Led a 36-participant study on LLM-based software engineering tasks with multi-class dependencies and iterative refinement.
- Analyzed GPT logs, test outcomes, and screen recordings to create data-driven guidelines, boosting developer productivity by ~20%.
- Received a strong recommendation letter acknowledging first-author-level contributions.

AIM Intelligence [Link] Oct 2024 - Feb.2025
Al Red Team Researcher, LLM jailbreak red team, *Seoul, South Korea*

- Investigated advanced LLM jailbreak techniques (achieved up to 95.9% success).
- **Co-first author** of *"One-Shot is Enough…" (ACL 2025 main)*, introducing M2S (Multi-turn-to-Single-turn) adversarial methods.
- Co-authored *"Breaking the Guardrails with Personality…" (ACL 2025 under review)*, exploring personality-driven jailbreak.

MLAI@KAIST AI Graduate school [Link] Mar. 2025 – August 2025
Research Intern, *Seoul, South Korea*

- Developing retrieval-augmented generation (RAG) for AI-driven drug discovery, integrating multi-document summarization, molecular pathway analysis, multi-LLM collaboration ("Chain of Agents"), and long-context summarization to accelerate research.

Georgia Institute of Technology (with NVIDIA) August. 2025 – Present
Research Intern, Atlanta, GA

- Coordinating a cross-institution agentic-LLM benchmark effort with NVIDIA across 10+ benchmarks and 8+ model families.

PUBLICATIONS

(*: equal contribution, †: Corresponding author)

Published

1. **One-Shot is Enough: Consolidating Multi-Turn Attacks into Efficient Single-Turn Prompts for LLMs**
   Junwoo Ha*, **Hyunjun Kim***, Sangyoon Yu, Haon Park, Ashkan Yousefpour, Yuna Park, Suhyun Kim†
   ACL main, 2025. **[PAPER]**
   - Introduced M2S method for efficient LLM jailbreaking (95.9% success).
   - Reduced LLM security testing time by 80%+ by consolidating multi-turn attacks.
2. **Optimizing Retrieval Strategies for Financial Question Answering Documents in Retrieval-Augmented Generation Systems [PAPER]**
   Sejong Kim*, Hyunseo Song*, Hyunwoo Seo*, **Hyunjun Kim***†
   ICLR Advances in Financial AI Workshop, 2025
   - Achieved 86% improvement in retrieval accuracy across 7 real-world finance QA datasets.
   - Reduced LLM hallucinations and accelerated compliance-critical insights at scale.
3. **Humanity's Last Exam**
   Long Phan et al. (including **Hyunjun Kim**)
   arXiv:2501.14249, 2025. **[PAPER]**
   - Contributed to large-scale collaborative evaluation of AI capabilities.
4. **The Impact of Prompt-Based Personality Induction on LLMs Safety**
   Yuna Park, Yujin Kim, Sangyoon Yu, Junwoo Ha, **Hyunjun Kim**, Wonwoo Ro, Suhyun Kim
   KCC 2025

Under Review

1. **OBJEX(MT): Objective Extraction and Metacognitive Calibration for LLM-as-a-Judge under Multi-Turn Jailbreaks**
   Hyunjun Kim, Junwoo Ha, Jiyoung Park, Sangyoon Yu, Haon Park
   Preprint, (Under review) NeurIPS 2025 Workshop on Multi-Turn Interactions in Large Language Models.
   - First author. Proposed OBJEX(MT), a benchmark for objective extraction + metacognitive calibration of LLM-as-a-Judge under multi-turn jailbreaks; uses a single human-aligned threshold ($\tau^* = 0.61$) and calibration metrics (ECE, Brier, Wrong@High-Conf).

- Findings. Claude-Sonnet-4 achieves the best accuracy 0.515 and calibration (ECE 0.296 / Brier 0.324); GPT-4.1 and Qwen3 are overconfident (mean confidence ≈ 0.88 vs. accuracy ≈ 0.44). Recommend providing explicit objectives to judges and using selective prediction/abstention.

2. **Personality Plug: How System Level Behavioral Customization Boosts Jailbreak Vulnerabilities**
   Yuna Park, Yujin Kim, Won Woo Ro, **Hyunjun Kim**, Junwoo Ha, Jiyoung Park, Garam Kim, Suhyun Kim
   (Under review) AAAI, 2026 AI Alignment Track.
   - Explores personality-driven jailbreak techniques in collaboration with KIST, KAIST, SNU, Yonsei, and University of Seoul.

3. **Experimental Analysis of Productive Interaction Strategy with ChatGPT: User Study on Function and Project-level Code Generation Tasks**
   Sangwon Hyun, Hyunjun Kim, Jinhyuk Jang, Hyojin Choi, and M. Ali Babar
   Preprint, (Under review) TOSEM 2026.
   - Investigating how different prompting strategies optimize LLM performance in real-world software engineering.

Awards & Honors

Gold Award (out of 200+ teams) – 4th UNIST-KAIST-POSTECH Data Science Competition [PPT] Jan. 2025
- Led development of an advanced retrieval-augmented generation (RAG) system for financial data analytics and multi-agent collaboration.
- Expanded into a research paper published at ICLR Advances in Financial AI Workshop, 2025.

CoE leadership - KAIST College of Engineering Mar. 2025
- Recognized for outstanding achievements beyond academics and research.

Dean's List Dec. 2023
- Achieved top 2% ranking among freshmen.

Excellence Award(out of over 200 teams) – Generative AI Application Contest Nov. 2024
- Honored for innovative application of generative AI technologies.

Excellence Award – Korea Tourism Data Lab Best Practices Nov. 2024
- Awarded for "Transforming Daejeon via Tourism Data Analysis," validated by a 288-student survey

Excellence Award – Youth SW Mentorship Essay Contest Dec. 2024
- Mentored middle/high school students in coding (Fall 2024) and authored a recognized essay on their growth.

Excellence Award – Youth SW Mentorship Essay Contest Dec. 2024
- Mentored middle/high school students in coding (Fall 2024) and authored a recognized essay on their growth.

ADDITIONAL EXPERIENCE

Intelligent Robot Design and Programming <Project VIRUS> [GitHub] [Midterm PPT, Final Demo]
Spring 2025
(Grade: A+)
- Developed VIRUS (Versatile Intelligent Robotic Unit for Strategy), an autonomous combat robot integrating an LLM and computer vision to execute complex tactical commands (e.g., hallway clearance, target pursuit) and minimize human risk in Close Quarters Battle (CQB) simulations.
- Led the design of the voice command and natural language understanding pipeline, using Whisper for real-time speech-to-text, an LLM to translate commands into executable JSON actions, and ElevenLabs for synthesized audio responses.
- Significantly enhanced command-following accuracy by fine-tuning a gpt-4.1-mini model on a custom tactical dataset. This model achieved a 70.8% win rate on the AlpacaEval 2.0 benchmark, effectively doubling the performance of the baseline model.

Special Issues: Entrepreneurship & Innovation <Core Skills for Entrepreneurs> [PPT] Fall 2023
(Grade: A0 | Instructor: Prof. Jongchul Kim, Founding Member, McKinsey Seoul)
- Analyzed early-adopter segments for "Answerock" using McKinsey's seven-step framework.
- Delivered a top-graded report recommending high-impact feature enhancements and a phased rollout strategy

Teaching Assistant Feb. 2024 – Present
- [CS101] Introduction to Programming, TA
  - Streamlined Python lab assignments for 50 students weekly with hands-on English guidance.
  - Delivered English-language mentoring and led discussions, enhancing coding proficiency for a diverse cohort.

Languages

- **English: Capable of presenting and debating complex topics in academic meetings and conferences.**
- **Korean: Native**