

Hyunbin Jin

hyunbin.jin@snu.ac.kr | <https://hynbjn.github.io/>

RESEARCH INTERESTS

AI Safety, Computational Social Science

My research interests lie in understanding how large language models behave throughout their reasoning processes, particularly how and why failures emerge. I am especially interested in safety failures, including how unsafe behaviors develop during reasoning and how these insights can inform methods for detecting and preventing them. I am also interested in how model behavior varies across languages, users, and social contexts, as well as the underlying factors that shape these differences.

EDUCATION

Seoul National University

M.S. in Data Science (Advisor: Prof. Taesup Kim)

Sep. 2023 ~ Feb. 2026

Seoul, South Korea

New York University

Exchange Student, College of Arts and Science

Sep. 2021 ~ Dec. 2021

New York, United States

Yonsei University

B.A. in Political Science and Diplomacy

Mar. 2018 ~ Feb. 2023

Seoul, South Korea

PUBLICATIONS (* : EQUAL CONTRIBUTION)

1. TrASE: Trajectory-Aware Safety Evaluation for Large Reasoning Models

[Hyunbin Jin*](#), Jungwhan Kim*

Under Review

2. Beyond Case Law: Evaluating Structure-Aware Retrieval and Safety in Statute-Centric Legal QA

Kyubyung Chae, Je Won Yeom, Jeongjae Park, Seunghyun Bae, Ijun Jang, [Hyunbin Jin](#), Jinkwan Jang, Taesup Kim

ACL 2026 Oral

3. What If TSF: A Benchmark for Reframing Forecasting as Scenario-Guided Multimodal Forecasting

Jinkwan Jang*, [Hyunbin Jin*](#), Hyungjin Park, Kyubyung Chae, Taesup Kim

Under Review

4. From Threat to Tool: Leveraging Refusal-Aware Injection Attacks for Safety Alignment

Kyubyung Chae*, [Hyunbin Jin*](#), Taesup Kim

arXiv preprint

5. ATAS: Any-to-Any Self-Distillation for Enhanced Open-Vocabulary Dense Prediction

Juan Yeo*, Soonwoo Cha*, Jiwoo Song*, [Hyunbin Jin](#), Taesup Kim

ICCV 2025

6. "Well, Keep Thinking": Enhancing LLM Reasoning with Adaptive Injection Decoding

[Hyunbin Jin*](#), Je Won Yeom*, Seunghyun Bae*, Taesup Kim

ACL 2025 Findings

7. HyperCLOVA X 8B Omni

NAVER Cloud HyperCLOVA X Team - Contributor

Technical Report, 2026

WORK EXPERIENCE

Residency Program, NAVER Cloud

Oct. 2025 ~ Jul. 2026

- Constructed distillation-based synthetic SFT data to improve LRM reasoning and designed a safety evaluation framework for assessing LRM reasoning trajectories.

NLP Research Engineer, Twigfarm

Oct. 2022 ~ Feb. 2023

- Developed an image-guided machine translation (IMT) pipeline for Korean-Japanese webtoon translation and constructed a parallel dataset for video-guided machine translation (VMT).

Data Analyst Intern, UpennSolution

Feb. 2021 ~ Jul. 2021

- Conducted quantitative analysis on real-world data to extract insights into social phenomena.

RESEARCH PROJECTS

Clinical Information Extraction

Oct. 2025 ~ Feb. 2026

Research Collaboration with Seoul National University Bundang Hospital

- Constructed an LLM-based pipeline that extracts and structures key clinical information from epilepsy patient notes, ensuring robustness to heterogeneous physician documentation styles.

Legal RAG System

Mar. 2025 ~ Aug. 2025

Research Collaboration with Naedam C&C

- Constructed a legal retrieval corpus and an evaluation QA dataset for Korean fire safety law, and developed a domain-specific RAG system validated on this benchmark.

Face Blur Estimation

Jul. 2022 ~ Sep. 2022

Research Collaboration with Alchera

- Developed and evaluated a regression model estimating blur levels in facial images to filter motion-blurred inputs for reliable face recognition.

PATENTS

Method and Apparatus for Adaptive Injection Decoding to Prevent Incomplete Reasoning in Large Language Models

Patent No. 10-2025-0165083, South Korea (Patent Pending)

AWARDS AND HONORS

Fellowships/Scholarships

Support for Next-Generation Researchers Program – Ministry of Education, South Korea 2024 ~ 2025

Merit-based Scholarship – Seoul National University 2024

Jilli Scholarship (Merit-based Scholarship) – Yonsei University 2018, 2020

Awards

Honors (Top 10% in the College of Social Science) – Yonsei University 2018, 2020