DongGeon Lee

Updated December 16, 2025

Email: dg.lee@postech.ac.kr **Homepage**: https://donggeon.github.io **Google Scholar**: /DongGeon Lee

Research Interests

Empirical AI safety and alignment for large language and vision-language models, with a focus on automated red teaming, jailbreak robustness, and scalable safety and alignment evaluations for multimodal, culturally grounded adversarial inputs.

Education

M.S. in Artificial Intelligence — Feb 2024 - Present
Pohang University of Science and Technology (POSTECH) — *Pohang, South Korea*

- Master's Thesis: *Evaluating the Safety of Vision-Language Models against Meme Images*

B.S. in Information and Communication Engineering — Mar 2018 - Feb 2024
Inha University — *Incheon, South Korea*

Research Experiences

Graduate Research Assistant — Feb 2024 - Present
Data Intelligence Lab, POSTECH (Advisor: Prof. Hwanjo Yu) — *Pohang, South Korea*

- Research on Vision-Language Model safety benchmarks and evaluation methodologies.

Research Scientist (Freelance) — Jul 2025 - Present
AIM Intelligence — *Seoul, South Korea*

- Research on automated multi-turn red teaming for LLMs.
- Research on safety guardrails for multimodal/multilingual LLMs.
- Research on dynamic evaluation of LLM compliance with organization-defined policies.

Invited Safety Researcher (External) — Dec 2025 - Present
Meta — *Remote*

- Selected as a pilot researcher for Meta's private bug bounty program.

Research Intern — Jan 2025 - Feb 2025
KT Corporation — *Seoul, South Korea*

- Research on mathematical data synthesis for pre-training Korea-centric LLM.

International Publications

Selected Publications (Full list available on Google Scholar.)
* Equal contribution; † Equal advising

[6] COMPASS: A Framework for Evaluating Organization-Specific Policy Alignment in LLMs
Dasol Choi*, DongGeon Lee*, Brigitta Jesica Kartono*, Helena Berndt, Haon Park, Hwanjo Yu†, Minsuk Kahng†
Under Review, 2025.10

[5] Are Vision-Language Models Safe in the Wild? A Meme-Based Benchmark Study
DongGeon Lee*, Joonwon Jang*, Jihae Jeong, Hwanjo Yu
EMNLP'25 | The 2025 Conference on Empirical Methods in Natural Language Processing

[4] Everyday Physics in Korean Contexts: A Culturally Grounded Physical Reasoning Benchmark
Jihae Jeong*, DaeYeop Lee*, DongGeon Lee, Hwanjo Yu
MRL @ EMNLP'25 | 5th Multilingual Representation Learning Workshop (Co-located with the 2025 Conference on Empirical Methods in Natural Language Processing)

[3] When Good Sounds Go Adversarial: Jailbreaking Audio-Language Models with Benign Inputs

Bodam Kim*, Hiskias Dingeto*, Taeyoun Kwon*, Dasol Choi, DongGeon Lee, Haon Park, Jae-Hoon Lee, Jongho Shin

arXiv Preprint, 2025.08

[2] Typed-RAG: Type-Aware Decomposition of Non-Factoid Questions for Retrieval-Augmented Generation
DongGeon Lee*, Ahjeong Park*, Hyeri Lee, Hyeonseo Nam, Yunho Maeng

XLLM @ ACL'25 | The First Workshop on Structure-aware Large Language Models (Co-located with the 63rd Annual Meeting of the Association for Computational Linguistics)

NAACL'25 SRW (Non-Archival) | 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Student Research Workshop

[1] REFIND at SemEval-2025 Task 3: Retrieval-Augmented Factuality Hallucination Detection in Large Language Models
DongGeon Lee, Hwanjo Yu

SemEval @ ACL'25 | The 19th International Workshop on Semantic Evaluation (Co-located with the 63rd Annual Meeting of the Association for Computational Linguistics)

Patents

Method and system for automated evaluation of a conversational language-model assistant against organization-specific policies
EP Patent Application 25216794.5 (Nov 18, 2025)

Method and apparatus for evaluating safety of an artificial intelligence model
KR Patent Application 10-2025-0148814 (Oct 15, 2025)

Multimodal content policy violation decision system
KR Patent Application 10-2025-0141965 (Sep 30, 2025)

System and method for evaluating safety of multimodal AI models using user-generated visual content
KR Patent Application 10-2025-0086059 (Jun 27, 2025)

System for providing parking information and control method
KR Patent Application 10-2021-0178090 (Dec 13, 2021)

Academic Services

Reviewer of AAAI'25 (The Association for the Advancement of Artificial Intelligence) 2025
Reviewer of International Journal of Nursing Studies 2025
Reviewer of MELT (Workshop on Multilingual and Equitable Language Technologies) at COLM'25 2025
Reviewer of SemEval (International Workshop on Semantic Evaluation) at ACL'25 2025
Student Volunteer of ACL'25 (Annual Meeting of the Association for Computational Linguistics) 2025

Honors and Awards

Excellent Paper Award 2025
HCLT 2025 (The 37th Annual Conference on Human & Cognitive Language Technology)

Gold Prize (Director's Award of the NIKL) 2024
Korean AI Language Proficiency Challenge, NIKL (National Institute of Korean Language)

Excellent Paper Award 2024
HCLT 2024 (The 36th Annual Conference on Human & Cognitive Language Technology)

Top Engineering Student Award 2024
Inha University

Technical Skills

Professional working proficiency
Python, PyTorch, transformers, vLLM, Git

Limited working proficiency
Google ADK, CrewAI, Shell Script, Keras, LaTeX

Elementary proficiency
DeepSpeed, TensorFlow, C++, C, MySQL