

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	<p>M.S. in Artificial Intelligence <i>Pohang University of Science and Technology (POSTECH)</i> Feb 2024 - Present • Master's Thesis: <i>Evaluating the Safety of Vision-Language Models against Meme Images</i></p> <p>B.S. in Information and Communication Engineering <i>Inha University</i> Mar 2018 - Feb 2024 Incheon, South Korea</p>		
RESEARCH EXPERIENCES	<p>Graduate Research Assistant <i>Data Intelligence Lab, POSTECH</i> (Advisor: Prof. Hwanjo Yu) Feb 2024 - Present Pohang, South Korea</p> <ul style="list-style-type: none">Research on Vision-Language Model safety benchmarks and evaluation methodologies. <p>Research Scientist (Freelance) <i>AIM Intelligence</i> Jul 2025 - Present Seoul, South Korea</p> <ul style="list-style-type: none">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. <p>Research Intern <i>KT Corporation</i> Jan 2025 - Feb 2025 Seoul, South Korea</p> <ul style="list-style-type: none">Research on mathematical data synthesis for pre-training Korea-centric LLM.		
INTERNATIONAL PUBLICATIONS	<p>Selected Publications (Full list available on Google Scholar.) * Equal contribution; † Equal advising</p> <p>[6] COMPASS: A Framework for Evaluating Organization-Specific Policy Alignment in LLMs Dasol Choi*, <u>DongGeon Lee*</u>, Brigitta Jesica Kartono*, Helena Berndt, Haon Park, Hwanjo Yu†, Minsuk Kahng† Under Review, 2025.10</p> <p>[5] Are Vision-Language Models Safe in the Wild? A Meme-Based Benchmark Study <u>DongGeon Lee*</u>, Joonwon Jang*, Jihae Jeong, Hwanjo Yu EMNLP'25 The 2025 Conference on Empirical Methods in Natural Language Processing</p> <p>[4] Everyday Physics in Korean Contexts: A Culturally Grounded Physical Reasoning Benchmark Jihae Jeong*, DaeYeop Lee*, <u>DongGeon Lee</u>, Hwanjo Yu MRL @ EMNLP'25 5th Multilingual Representation Learning Workshop (Co-located with the 2025 Conference on Empirical Methods in Natural Language Processing)</p> <p>[3] When Good Sounds Go Adversarial: Jailbreaking Audio-Language Models with Benign Inputs Bodam Kim*, Hiskias Dingeto*, Taeyoun Kwon*, Dasol Choi, <u>DongGeon Lee</u>, Haon Park, Jae-Hoon Lee, Jongho Shin arXiv Preprint, 2025.08</p>		

- [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 <i>EP Patent Application 25216794.5 (Nov 18, 2025)</i>
	Method and apparatus for evaluating safety of an artificial intelligence model <i>KR Patent Application 10-2025-0148814 (Oct 15, 2025)</i>
	Multimodal content policy violation decision system <i>KR Patent Application 10-2025-0141965 (Sep 30, 2025)</i>
	System and method for evaluating safety of multimodal AI models using user-generated visual content <i>KR Patent Application 10-2025-0086059 (Jun 27, 2025)</i>
	System for providing parking information and control method <i>KR Patent Application 10-2021-0178090 (Dec 13, 2021)</i>

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 <i>HCLT 2025 (The 37th Annual Conference on Human & Cognitive Language Technology)</i>	2025
	Gold Prize (Director's Award of the NIKL) <i>Korean AI Language Proficiency Challenge, NIKL (National Institute of Korean Language)</i>	2024
	Excellent Paper Award <i>HCLT 2024 (The 36th Annual Conference on Human & Cognitive Language Technology)</i>	2024
	Top Engineering Student Award <i>Inha University</i>	2024

TECHNICAL SKILLS	Professional working proficiency Python, PyTorch, transformers, vLLM, Git
	Limited working proficiency Google ADK, CrewAI, Shell Script, Keras, L ^A T _E X
	Elementary proficiency DeepSpeed, TensorFlow, C++, C, MySQL