DongGeon Lee

Updated October 18, 2026

Email: dg.lee@postech.ac.kr **Homepage**: https://donggeon.github.io **LinkedIn**: /dong-geon-lee

Research Areas

Data-centric Natural Language Processing (Data synthesis, Evaluation),
Multimodal AI Safety and Alignment (Post-training, Benchmarking)

Technical Skills

Professional working proficiency
Python, PyTorch, transformers, vLLM, Git, Slurm
Limited working proficiency
Shell Script, Docker, Megatron-LM, FastAPI, LaTeX
Elementary proficiency
DeepSpeed, Unsloth, Elasticsearch, MySQL

Work Experiences

Research Engineer Mar 2026 - Present
Frontier AI Lab, Korea Telecom (KT) *Seoul, South Korea*

- Research on enhancing agentic capabilities of foundation models (data synthesis for agentic mid-training and post-training).

Research Scientist Jul 2025 - Feb 2026
AIM Intelligence *Seoul, South Korea*

- Research on automated multi-turn red teaming for large language models (LLMs).
- Research on self-evolving safety guardrails for multimodal/multilingual LLMs.
- Research on dynamic evaluation of LLM compliance with organization-defined policies. [News]

Research Intern Jan 2025 - Feb 2025
Gen AI Lab, Korea Telecom (KT) *Seoul, South Korea*

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

Education

M.S. in Artificial Intelligence Feb 2024 - Feb 2026
Pohang University of Science and Technology (POSTECH) *Pohang, South Korea*

- Master's Thesis: *Evaluating the Safety of Vision-Language Models against Meme Images* (Advisor: Prof. Hwanjo Yu)

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

International Publications

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

[8] Global PIQA: Evaluating Commonsense Reasoning Across 100+ Languages and Cultures
Tyler A. Chang*, Catherine Arnett*, ... , DongGeon Lee, ...
(Author list truncated for brevity; 339 authors in total.)
NeurIPS'26 Evaluations Datasets Track | The 40th Annual Conference on Neural Information Processing Systems

[7] Multimodal Safety Evaluation Should Measure Controllability Beyond Classification
Junhyeong Park*, Hanwool Lee,* DongGeon Lee, Dasol Choi, Yejin Son, Haon Park†, Youngjae Yu†
arXiv Preprint, 2026.10

[6] COMPASS: A Framework for Evaluating Organization-Specific Policy Alignment in LLMs
Dasol Choi*, DongGeon Lee*, Brigitta Jesica Kartono*, Helena Berndt, Taeyoun Kwon, Joonwon Jang, Haon Park, Hwanjo Yu†, Minsuk Kahng†
ACL'26 | The 64th Annual Meeting of the Association for Computational Linguistics

[5] CJK Representation across arXiv Papers, Hugging Face Datasets, and Models
Dasol Choi*, DongGeon Lee*, Dokyoon Yoon , Youngsook Song
PACLIC'26 | The 40th Pacific Asia Conference on Language, Information and Computation

[4] 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

[3] When Good Sounds Go Adversarial: Jailbreaking Audio-Language Models with Benign Inputs
Hiskias Dingeto*, Taeyoun Kwon*, Dasol Choi*, Bodam Kim, 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

Selected Patents (Full list available on Google Scholar.)

Device, method and computer program for generating math problem data based on Korean
KR Patent Application 10-2026-0008110 (Jan 15, 2026)

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)

Academic Services

Program Committee of AAAI'27 (The Association for the Advancement of Artificial Intelligence) 2026
Reviewer of GenAI4World Workshop at COLM'26 2026
Reviewer of Workshop on Efficient Reasoning at COLM'26 2026
Reviewer of AI4GOOD Workshop at NeurIPS'26 2026
Program Committee of AAAI'26 (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)
NAACL 2025 Registration Grant 2025
NAACL 2025 SRW (Student Research Workshop)
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