Jongwon Lim

Ph.D. Student, Graduate School of Data Science, Seoul National University

elijah0430@snu.ac.kr

RESEARCH INTERESTS

I am interested in developing methods for understanding the internal mechanisms of language models, and using those insights to improve practical NLP and LLM applications.

EDUCATION

2025 - Present
Seoul National University
Ph.D. in Data Science, Graduate School of Data Science.
Advisor: Prof. Yohan Jo, HOLI Lab.

2019 - 2025
Seoul National University
B.A. in Linguistics and Data Science for Humanities.

PUBLICATIONS

arXiv preprint, 2026
Your Language Model is Its Own Critic: Reinforcement Learning with Value Estimation from Actor's Internal States
Yunho Choi*, Jongwon Lim*, Woojin Ahn, Minjae Oh, Jeonghoon Shim, Yohan Jo
POISE estimates RLVR baselines from the actor's internal hidden states and entropy statistics, reducing rollout overhead while matching DAPO-level performance.

ICML 2026 Regular Paper; Mechanistic Interpretability Workshop @ NeurIPS 2025
Dual Mechanisms of Value Expression: Intrinsic vs. Prompted Values in Large Language Models
Jongwook Han*, Jongwon Lim*, Injin Kong, Yohan Jo
Mechanistic analysis of how language models internally represent and express values under intrinsic and prompted settings.

ACL Findings 2026
Learning to Retrieve User History and Generate User Profiles for Personalized Persuasiveness Prediction
Sejun Park, Yoonah Park, Jongwon Lim, Yohan Jo
Context-aware user profiling framework for retrieving persuasion-relevant history and generating user profiles for persuasiveness prediction.

FEVER Workshop @ EMNLP 2024
DAHL: Domain-specific Automated Hallucination Evaluation of Long-Form Text through a Benchmark Dataset in Biomedicine
Jean Seo, Jongwon Lim, Dongjun Jang, Hyopil Shin
Biomedical benchmark and automated evaluation pipeline for factuality assessment in long-form LLM outputs.

* Equal contribution

RESEARCH EXPERIENCE

Oct 2023 - Sep 2024
Research Intern, CL_NLP Lab, Seoul National University
Worked on language model training and evaluation, retrieval-augmented generation, experiments, analysis, and academic writing.

May 2024 - Dec 2024
Head Researcher, SNU Faculty of Liberal Education
Led work on a benchmark dataset for evaluating morphological capabilities of large language models.

Jan 2024 - Mar 2024
Kaggle Silver Medalist, LLM - Detect AI-generated Text
Built ensemble systems for detecting AI-generated text.

TEACHING AND SERVICE

Spring 2026	Teaching Assistant, Large Language Models and Conversational AI Seoul National University.
2026	Mentor, 3rd LG AI Youth Camp Mentored student participants on AI projects and research-oriented problem solving.
2026	Reviewer, ICML 2026 Workshops Pluralistic Alignment Workshop @ ICML 2026; Mechanistic Interpretability Workshop @ ICML 2026.

AWARDS AND OTHER EXPERIENCE

2024	Outstanding Bachelor's Thesis SNU Linguistics Department.
Apr 2021 - Sep 2022	Military Service, Republic of Korea Army Instructor for armored vehicle operation at the Republic of Korea Army Armor School.