Jongwon Lim

Ph.D. Student, Graduate School of Data Science, Seoul National University
elijah0430@snu.ac.kr
Homepage · Google Scholar

RESEARCH INTERESTS

My research focuses on **using interpretability to build an actionable understanding of foundation models**. I am exploring two complementary directions:

1. Understanding **what** models learn
   - Which learned mechanisms enable or limit model capabilities in practice? [2]
   - How do models process concepts differently from humans, and when do these differences lead to failures? [4]

2. Understanding **how** models learn
   - Which changes in a model's internal computations drive improvements during training? [3]
   - How can we use these insights to improve training algorithms and pipelines? [1], [4]

SELECTED PAPERS

* Equal contribution

[1] **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
NeurIPS 2026 Poster [arXiv] [Project page]

[2] **Dual Mechanisms of Value Expression: Intrinsic vs. Prompted Values in Large Language Models**
Jongwook Han*, **Jongwon Lim***, Injin Kong, Yohan Jo
ICML 2026 Regular [arXiv] [Project page] [OpenReview]

[3] **The Low-Rank Structure of VLA Reinforcement Learning**
Minjae Oh*, Yoonah Park*, **Jongwon Lim***, Yohan Jo
Preprint (under review) [arXiv]

[4] **How Do LLMs Change Predictions Under Negation?**
Jongwook Yoon*, **Jongwon Lim***, Sungjib Lim, Woojin Cho, Yohan Jo
Preprint (under review) [arXiv]

EDUCATION

Seoul National University 2025 - Present
Ph.D. in Data Science, Graduate School of Data Science. Advised by Yohan Jo at Human-Oriented Language Intelligence (HOLI) Lab.

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

AWARDS

Research Scholarship 2025.09 - 2026.08
National Research Foundation of Korea (NRF)

Outstanding Bachelor's Thesis 2024
Evaluating the Understanding of Blend Morphology in Large Language Models

TEACHING

Teaching Assistant Fall 2026
Seoul National University. Course: AI at Work.

Teaching Assistant Spring 2026
Seoul National University. Course: Large Language Models and Conversational AI.

SERVICES

Conference Reviewer
ICLR 2027

Workshop Reviewer
- Pluralistic Alignment Workshop @ ICML 2026
- Mechanistic Interpretability Workshop @ ICML 2026 · also emergency reviewer
- 3rd AI for Math Workshop @ ICML 2026
- BlackboxNLP Workshop @ EMNLP 2026
- Interpretability for Discovery Workshop @ NeurIPS 2026 · also emergency reviewer
- Transitioning from Pre-Training to Post-Training Workshop @ NeurIPS 2026

Mentor, 3rd LG AI Youth Camp 2026
Mentored student participants on AI projects and research-oriented problem solving.

EXPERIENCES

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

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

Military Service, Republic of Korea Army Apr 2021 - Sep 2022
Instructor for armored vehicle operation at the Republic of Korea Army Armor School.

SKILLS & COMPETENCIES

- **Mathematics:** Linear algebra, optimization, calculus
- **Programming:** Python, C/C++
- **Languages:** Korean (native), English (fluent, TOEIC 990)