Jiwoo **Hong**

jiwoo_hong@kaist.ac.kr | jiwooya1000 | jiwoohong09 | @jiwoohong98 | Google Scholar

Education

KAIST (Korea Advanced Institute of Science and Technology) — *Seoul, S.Korea*

Master's Degree in Graduate School of Artificial Intelligence — *Sep. 2023 - Aug. 2025 (Expected)*

- GPA: 4.2 / 4.3 ($\sim 3^{rd}$ semester), Supervised by Professor James Throne

SKKU (SungKyunKwan University) — *Seoul, S.Korea*

Bachelor's Degree in Statistics & Industrial Engineering — *Mar. 2017 - Feb. 2023*

- GPA : 4.38 / 4.5 (*Summa Cum Laude*)

Research Interest

Natural Language Processing Large Language Model, Preference Alignment, Reasoning

Reinforcement Learning Reinforcement Learning with Human/AI Feedback

Publications (2025)

On the Robustness of Reward Models for Language Model Alignment — *Under Review for ICML 2025*

Jiwoo Hong, Noah Lee, Eunki Kim, Guijin Son, Woojin Chung, Aman Gupta, Shao Tang, & James Thorne — ***Released***: -, ***Citation***: -

- Keywords: Reward Modeling, Large Language Models, Preference Alignment, RLHF

AlphaPO - Reward Shape Matters for LLM Alignment — *Under Review for ICML 2025*

Aman Gupta, Shao Tang, Qingquan Song, Sirou Zhu, **Jiwoo Hong**, Ankan Saha, Viral Gupta, Noah Lee, Eunki Kim, Siyu Zhu, Parag Agrawal, Natesh Pillai, & S. Sathiya Keerthi — ***Released***: 24.01.07, ***Citation***: -

- Keywords: Large Language Models, Preference Alignment, RLHF

Controlling the Geometry of Token Embeddings in Language Models — *Under Review for ACL 2025*

Woojin Chung*, **Jiwoo Hong***, Suchir Salhan*, Jeonghoon Kim, Richard Diehl Martinez, James Thorne, & Paula Buttery — ***Released***: -, ***Citation***: -

- Keywords: Language Model Pre-training, Interpretability

Linguistic Generalizability of Test-Time Scaling in Mathematical Reasoning — *Under Review for ACL 2025*

Guijin Son, **Jiwoo Hong**, Hyunwoo Ko, & James Thorne — ***Released***: 25.02.25, ***Citation***: -

- Keywords: Language Model Reasoning, Language Model Generalizability

Publications (Until 2024)

ORPO: Monolithic Preference Optimization without Reference Model — *EMNLP 2024*

Jiwoo Hong, Noah Lee & James Thorne — ***Released***: 24.03.12, ***Citation***: 231

- Keywords: Large Language Models, Preference Alignment, Instruction-tuning

Stable Language Model Pre-training by Reducing Embedding Variability — *EMNLP 2024*

Woojin Chung, **Jiwoo Hong**, Na Min An, James Thorne, & Se Young Yoon — ***Released***: 24.09.12, ***Citation***: 1

- Keywords: Language Model Pre-training, Foundation Models

Cross-lingual Transfer of Reward Models in Multilingual Alignment — *NAACL 2025*

Jiwoo Hong*, Noah Lee*, Rodrigo Martínez-Castaño, César Rodríguez, & James Thorne — ***Released***: 24.10.24, ***Citation***: -

- Keywords: Reward Modeling, Language Model Generalizability, RLHF

Margin-aware Preference Optimization for Aligning Diffusion Models without Reference — *Under Review for CVPR 2025*

Jiwoo Hong*, Sayak Paul*, Noah Lee, Kashif Rasul, James Thorne & Jongheon Jeong — ***Released***: 24.06.11, ***Citation***: 5

- Keywords: Text-to-image diffusion models, Preference Alignment, Instruction-tuning

Evaluating the Consistency of LLM Evaluators — *COLING 2025*

Noah Lee*, **Jiwoo Hong***, & James Thorne — ***Released***: 24.11.30, ***Citation***: 1

- Keywords: LLM as evaluator, Evaluation Bias

Disentangling Structure and Style: Political Bias Detection in News by Inducing Document Hierarchy *Findings of EMNLP 2023*

Jiwoo Hong, Yejin Cho, Jaemin Jung, Jiyoung Han & James Thorne ***Released***: *23.12.06,* ***Citation***: 5

- Keywords: Bias Detection, Explainable Document Understanding, Discourse Analysis

MARL-Based Dual Reward Model on Segmented Actions for Multiple Mobile Robots in Automated Warehouse Environment *Applied Science*

Hyeoksoo Lee, **Jiwoo Hong** & Jongpil Jeong ***Released***: *22.05.07,* ***Citation***: 11

- Keywords: Multi-Agent Reinforcement Learning, Reward Shaping

Industry Experiences

Amazon - Rufus *Palo Alto, CA*

Applied Scientist Internship *May. 2025 - Aug. 2025 (Incoming)*

- **Topic**: Multi-objective optimization in reinforcement learning for language models (To be specified).

Naver Cloud - Generative Chatbot *Seoul, S.Korea*

Research Internship *Feb. 2025 - Present (May. 2025)*

- **Topic**: Reinforcement learning with verifiable rewards (RLVR) algorithms and pipeline development.

Additional Experiences

Multiple Invited Talks on LLM Alignment *Seoul, S.Korea*

Invited Speaker

- Talk on recent alignment techniques and ORPO at:
- (1) Kakao Brain, (2) AI EXPO Korea, (3) KISTI (Korean Institute of Science and Technology Information), (4) Tweleve Labs,
- and (5) Trillion Parameter Consortium Seminar.

Zephyr-ORPO Project *Seoul, S.Korea*

Open-source Language Model Development *April. 2024*

- Developing data-efficient open-source instruction-following large language models with fine-grained synthetic data.
- In collaboration with Argilla and Hugging Face.

ML/NLP Conference Reviewer *Seoul, S.Korea*

Reviewer *Nov. 2024 - Present*

- Served as a reviewer for ICLR 2025 on the papers related to LLM alignment.
- Serving as a reviewer for ACL Rolling Review (ARR) February cycle on the papers related to LLM alignment.