Jiwoo **Hong**

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

Redwood City, California, US

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*

- GPA: 4.12 / 4.3 (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*)

Industry Experiences

Amazon — *Palo Alto, CA, US*

Applied Scientist Intern — *May. 2025 - Present*

- **Topic**: Multi-objective steerability injection in LLMs via reinforcement learning and closing the Reward-Benchmark Gap.

Naver Cloud — *Seoul, S.Korea*

AI Research Scientist Intern — *Feb. 2025 - May. 2025*

- **Topic**: Core contributor in developing a proprietary large-scale reasoning model, HyperCLOVA X THINK.
- Built a theoretical background in applying online difficulty filtering in Reinforcement learning with verifiable rewards (RLVR) for maximum efficiency.

Publications

Selected Works

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

Jiwoo Hong, Noah Lee & James Thorne — ***Topic***: *Offline Preference Learning*

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

Jiwoo Hong, Noah Lee, Eunki Kim, Guijin Son, Woojin Chung, Aman Gupta, Shao Tang, & James Thorne — ***Topic***: *RLHF, Generalizability*

Online Difficulty Filtering for Reasoning Oriented Reinforcement Learning — *Preprint*

Sanghwan Bae*, **Jiwoo Hong***, Min Young Lee, Hanbyul Kim, JeongYeon Nam, & Donghyun Kwak — ***Topic***: *RLVR, Reasoning*

2025

HyperCLOVA X THINK Technical Report — *Technical Report*

NAVER Cloud HyperCLOVA X Team, Core contributor — ***Topic***: *Proprietary Reasoning Model*

When AI Co-Scientists Fail: SPOT-a Benchmark for Automated Verification of Scientific Research — *Under Review for NeurIPS 2025*

Guijin Son, **Jiwoo Hong**, Honglu Fan, Heejeong Nam, Hyunwoo Ko, Seungwon Lim, Jinyeop Song, Jinha Choi, Gonçalo Paulo, Youngjae Yu, & Stella Biderman — ***Topic***: *AI for Science*

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

Jiwoo Hong, Noah Lee, Eunki Kim, Guijin Son, Woojin Chung, Aman Gupta, Shao Tang, & James Thorne — ***Topic***: *RLHF, Generalizability*

AlphaPO: Reward Shape Matters for LLM Alignment — *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 — ***Topic***: *RLHF, Interpretability*

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

Guijin Son, **Jiwoo Hong**, Hyunwoo Ko, & James Thorne — ***Topic***: *Reasoning, Multilingual*

Online Difficulty Filtering for Reasoning Oriented Reinforcement Learning *Preprint*

Sanghwan Bae*, **Jiwoo Hong***, Min Young Lee, Hanbyul Kim, JeongYeon Nam, & Donghyun Kwak ***Topic***: *RLVR, Reasoning*

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

Jiwoo Hong*, Noah Lee*, Rodrigo Martínez-Castaño, César Rodríguez, & James Thorne ***Topic***: *RLHF, Generalizability*

2024

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

Jiwoo Hong, Noah Lee & James Thorne ***Topic***: *Offline Preference Learning*

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

Woojin Chung, **Jiwoo Hong**, Na Min An, James Thorne, & Se Young Yoon ***Topic***: *Pre-training*

Margin-aware Preference Optimization for Aligning Diffusion Models without Reference *ICLR 2025 SCOPE Workshop*

Jiwoo Hong*, Sayak Paul*, Noah Lee, Kashif Rasul, James Thorne & Jongheon Jeong ***Topic***: *RLHF, Diffusion*

Evaluating the Consistency of LLM Evaluators *COLING 2025*

Noah Lee*, **Jiwoo Hong***, & James Thorne ***Topic***: *LLM-as-a-Judge*

2023

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 ***Topic***: *NLP Application*

2022

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

Hyeoksoo Lee, **Jiwoo Hong** & Jongpil Jeong ***Topic***: *Multi-agent RL*

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-recipe instruction-following LLMs with fine-grained synthetic data and Mixtral-8x22B-v0.1.
- In collaboration with Argilla and Hugging Face, published Zephyr-ORPO-141B-A35B-v0.1.

HyperCLOVA X THINK *Seoul, S.Korea*

RLVR Pipeline Development in Proprietary Large Reasoning Model *Feb.-May. 2025*

- Participated in developing a proprietary large-scale reasoning model, HyperCLOVA X THINK, specialized in Korean.
- Presented theoretical background on online difficulty filtering in reinforcement learning with verifiable rewards (RLVR), contributing to sample-efficient improvements in complex reasoning tasks.

ML/NLP Conference Reviewer -

Reviewer *Nov. 2024 - Present*

- TMLR
- ICLR 2025
- ACL Rolling Review (Since February, 2025)