Hyeonbin Hwang
Ph.D Student at KAIST AI

Contact Information

E-mail hbin0701@kaist.ac.kr
Homepage https://hbin0701.github.io

Interest

My long-term goal is to build models that can reason beyond human data distribution. To this end, I study the learning signals, data structures, and evaluation methods that enable models to self-train and generalize systematically in multi-step reasoning. Currently, I am interested in iterative RL post-training and agentic systems, particularly autonomous research agents.

Experience

Research Intern
AI Companion Team, KRAFTON AI — *May 2025 – Nov 2025*
Seoul, Republic of Korea
- Game-Play Agent for PUBG (w. NVIDIA)
  (Link: www.youtube.com/watch?v=wEKUSMqrbzQ)

Research Intern
LK Lab, KAIST — *Dec 2022 – Feb 2024*
Seoul, Republic of Korea
- In-Context Learning
- Self-Feedback
- Fine-grained Evaluation of LMs

Research Intern
Naver CLOVA, Healthcare AI — *Mar 2022 – Aug 2022*
Seongnam, Republic of Korea
- EHR Summarization and embedding

Education

KAIST — *Aug 2026 (Exp.) -*
Ph.D., Graduate School of AI
Advisor : Seongjoon Oh

KAIST — *Feb 2024 – Aug 2026*
M.S., Graduate School of AI
Advisor : Minjoon Seo

KAIST — *Sep 2019 – Feb 2024*
B.S., School of Computing
Advisor : Tae-Kyun Kim
GPA: 3.96 / 4.3

Languages

- Korean — Native
- English — Fluent (TOEFL 113)
- Spanish — Fluent (DELE C1)

Honors & Awards

- Young-Han Kim Global Leadership Scholarship Awardee (2022)
- Qualcomm-KAIST Innovation Award Hackathon Winner (2023)

Reviewer Experience

- ICLR (2026)
- ICML (2026, Gold Reviewer [Top 25%])
- NeurIPS (2026)
- TMLR (2026)

Publications

[1] **Hwang, H., &** Park, Y. (2026). ***Intrinsic Task Symmetry Drives Generalization in Algorithmic Tasks***
(ICML 2026)

[2] **Hwang, H**., Jeon, B., Kim, S., Kim, J., Chang, H., Yang, S., Won, S., Lee, D., Ahn, Y., ... (2025). ***Let's Predict Sentence by Sentence*****.** arXiv preprint arXiv:2505.22202.
(COLM RAM 2 Workshop [Oral])

[3] Won, Y., Lee, H., **Hwang, H.**, & Seo, M. (2025). ***Differential Information: An Information-Theoretic Perspective on Preference Optimization*****.** arXiv preprint arXiv:2505.23761**.**
(Under Review)

[4] Chang, H., Park, J., Cho, H., Yang, S., Ko, M., **Hwang, H**., Won, S., Lee, D., Ahn, Y., ... (2025). ***The Coverage Principle: A Framework for Understanding Compositional Generalization*****.** arXiv preprint arXiv:2505.20278.
(ICLR 2026)

[5] Lee, S., Kim, S., Seo, M., Jo, Y., Go, D., **Hwang, H**., Park, J., Yue, X., Welleck, S., ... (2025). ***The CoT Encyclopedia: Analyzing, Predicting, and Controlling how a Reasoning Model will Think*****.** arXiv preprint arXiv:2505.10185.
(ICLR 2026)

[6] Kim, J., Lee, H., Cho, H., Jang, J., **Hwang, H**., Won, S., Ahn, Y., Lee, D., & Seo, M. (2024). ***Knowledge Entropy Decay during Language Model Pretraining Hinders New Knowledge Acquisition*****.**
(ICLR 2025 [Oral] & AAAI Workshop KnowFM 2025 [Best Paper])

[6] Kim, S., Suk, J., Cho, J. Y., Longpre, S., Kim, C., Yoon, D., Son, G., Cho, Y., & **Hwang, H.** (2024). ***The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models***
(NAACL 2025 [BEST PAPER])

[7] **Hwang, H**., Kim, D., Kim, S., Ye, S., & Seo, M.(2024). *Self-Explore to Avoid the Pit: Improving the Reasoning Capabilities of Language Models with Fine-grained Rewards*.
(EMNLP Findings (2024) & ACL Workshop NLRSE 2024 [Oral])

[8] Ye, S., **Hwang, H**., Yang, S., Yun, H., Kim, Y., & Seo, M. (2023). ***Investigating the Effectiveness of Task-Agnostic Prefix Prompt for Instruction Following.***
(AAAI 2024)

[9] Ye, S., Kim, D., Kim, S., **Hwang, H**., Kim, S., Jo, Y., Thorne, J., Kim, J., & Seo, M. (2023). ***FLASK: Fine-grained language model evaluation based on alignment skill sets.***
(ICLR 2024 [Spotlight])

[10] **Hwang, H**., Kim, S., Park, W. J., Seo, J., Ko, K., & Yeo, H. (2022). ***Vision transformer equipped with neural resizer on facial expression recognition task.***
(ICASSP 2022)