Minjae Lee

mjbooo@kaist.ac.kr | minjae.lee.official@gmail.com | Homepage | LinkedIn

EDUCATION

Korea Advanced Institute of Science and Technology (KAIST) Seongnam, Korea
M.S in Kim Jaechul Graduate School of AI (GSAI) Mar. 2023
Thesis: Exploring Optimal Encoders for Electronic Health Records
Advisor: Edward Choi

Korea Advanced Institute of Science and Technology (KAIST) Daejeon, Korea
B.S in Industrial & Systems Engineering (ISysE); Minor in Economics Feb. 2021
Honors: Summa Cum Laude (GPA 4.0/4.3)

Technical University of Berlin (Technische Universität Berlin, TUB) Berlin, Germany
KAIST Outbound Exchange Program in Informatics (Informatik) Apr. ~ Jul. 2017
Honors: Mirae Asset Outbound Exchange Student Scholarship

RESEARCH INTERESTS

General Focus: Efficiently solving real-world problems with machine learning, with two main directions:

- **Inference in Large Language Models:** Developing methods for **efficient and agentic inference** in large language models, with a focus on speculative decoding, test-time compute scaling, and reasoning.
- **Applied Healthcare Machine Learning:** Enabling **automated clinical decision-making** by building on **prior work in Electronic Health Records** (EHRs), including their representation, encoding, synthesis, and sharing.

EMPLOYMENT

AI Algorithm Researcher, FuriosaAI, Seoul, Korea Mar. 2023 ~ Present
Conducted research with a focus on language modeling, inference acceleration, and test-time compute scaling

Research Assistant, KAIST, Seongnam, Korea Jul. 2023 ~ Jul. 2024
Researched EHR synthesis and transfer learning in healthcare with MD collaboration as a Post-MS Research Assistant

Research Intern, NAVER CLOVA, Seongnam, Korea Dec. 2021 ~ Feb. 2022
Investigated quantization and compression techniques for efficient AI models

PUBLICATIONS

(*: Equal Contribution, †: Equal Advising)

2026 **[C5, W1] TABED: Test-Time Adaptive Ensemble Drafting for Robust Speculative Decoding in LVLMs** [Paper]
Minjae Lee*, Wonjun Kang*, Byeongkeun Ahn, Christian Classen, Kevin Galim, Seunghyuk Oh, Minghao Yan, Hyung Il Koo, Kangwook Lee
*European Chapter of the Association for Computational Linguistics (**EACL**) Findings*
*International Conference on Learning Representations (**ICLR**) Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*

[C7] Draft-based Approximate Inference for LLMs [Paper]
Kevin Galim, Ethan Ewer, Wonjun Kang, **Minjae Lee**, Hyung Il Koo, Kangwook Lee
*International Conference on Learning Representations (**ICLR**)*

[C6] ParallelBench: Understanding the Trade-offs of Parallel Decoding in Diffusion LLMs [Paper]
Wonjun Kang, Kevin Galim, Seunghyuk Oh, **Minjae Lee**, Yuchen Zeng, Shuibai Zhang, Coleman

Hooper, Yuezhou Hu, Hyung Il Koo, Nam Ik Cho, Kangwook Lee
*International Conference on Learning Representations (**ICLR**)*

[W3] Inference-Aligned SFT for Diffusion LLMs via Group-based Trajectory Sampling [Paper]
Seunghyuk Oh, **Minjae Lee**, Kevin Galim, Minseo Kim, Hyung Il Koo, Wonjun Kang, Hanbaek Lyu, Kangwook Lee
*International Conference on Learning Representations (**ICLR**) Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*

[J4, W2] Transformers in the Dark: Navigating Unknown Search Spaces via Noisy Feedback [Paper]
Jungtaek Kim, Ziqian Lin, Thomas Zeng, **Minjae Lee**, Chungpa Lee, Jy-yong Sohn, Hyung Il Koo, Kangwook Lee
*Transactions on Machine Learning Research (**TMLR**)*
*Neural Information Processing Systems (**NeurIPS**) Workshop on What Can('t) Transformers Do?*

[J3] UNCAGE: Contrastive Attention Guidance for Masked Generative Transformers in Text-to-Image Generation [Paper]
Wonjun Kang, Byeongkeun Ahn, **Minjae Lee**, Kevin Galim, Seunghyuk Oh, Hyung Il Koo, Nam Ik Cho
*Institute of Electrical and Electronics Engineers (**IEEE**) Access*

2025 **[C4] Generating Multi-Table Time Series EHR from Latent Space with Minimal Preprocessing** [Paper]
Eunbyeol Cho, Jiyoun Kim, **Minjae Lee**, Sungjin Park, Edward Choi
*Neural Information Processing Systems (**NeurIPS**)*

[C3] VersaPRM: Multi-Domain Process Reward Model via Synthetic Reasoning Data [Paper]
Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, **Minjae Lee**, Heeju Kim, Wonjun Kang, Jackson Kunde, Ying Fan, Jungtaek Kim, Hyung Il Koo, Kannan Ramchandran, Dimitris Papailiopoulos, Kangwook Lee
*International Conference on Machine Learning (**ICML**)* 🏆***Oral (top ~1%)***

[C2] State-offset Tuning: State-based Parameter-Efficient Fine-Tuning for State Space Models [Paper]
Wonjun Kang*, Kevin Galim*, Yuchen Zeng, **Minjae Lee**, Hyung Il Koo, Nam Ik Cho
*Association for Computational Linguistics (**ACL**)*

[J2] A deep-learning algorithm using real-time collected intraoperative vital sign signals for predicting acute kidney injury after major non-cardiac surgeries: A modelling study [Paper]
Sehoon Park, Soomin Chung, Yisak Kim, Sun-Ah Yang, Soie Kwon, Jeong Min Cho, **Minjae Lee**, Eunbyeol Cho, Jiwon Ryu, Sejoong Kim, Jeonghwan Lee, Hyung Jin Yoon, Edward Choi, Kwangsoo Kim, Hajeong Lee
*Public Library of Science (**PLOS**) Medicine*

2023 **[C1] Rediscovery of CNN's Versatility for Text-based Encoding of Raw Electronic Health Records** [Paper]
Eunbyeol Cho*, **Minjae Lee***, Kyunghoon Hur, Jiyoun Kim, Jinsung Yoon, Edward Choi
*Conference on Health, Inference, and Learning (**CHIL**)* 🏆***Oral (top ~10 papers)***

[J1] Genhpf: General healthcare predictive framework for multi-task multi-source learning [Paper]
Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyoun Kim, **Minjae Lee**, Eunbyeol Cho, Seong-Eun Moon, Young-Hak Kim, Louis Atallah, Edward Choi
*IEEE Journal of Biomedical and Health Informatics (**JBHI**)*

PREPRINTS / UNDER REVIEWS / ONGOING RESEARCHES

2025 **[P1] Privacy-preserving synthetic data enhances postoperative acute kidney injury prediction in data-scarce clinical settings: a multicenter modeling study** [Paper]
Soie Kwon, Eunbyeol Cho, **Minjae Lee**, Yisak Kim, Sunah Yang, Jeong Min Cho, Sehoon Park, Chung Hee Baek, Jiwon Ryu, Sejoong Kim, Jeonghwan Lee, Ji In Park, Jin Ho Hwang, Ji Eun Kim, Kwangsoo Kim, Hyung-Jin Yoon, Edward Choi†, Hajeong Lee†
Under review

[P2] XQuant: Breaking the Memory Wall for LLM Inference with KV Cache Rematerialization [Paper]
Aditya Tomar*, Coleman Hooper*, **Minjae Lee**, Haocheng Xi, Rishabh Tiwari, Wonjun Kang, Luca Manolache, Michael W. Mahoney, Kurt Keutzer, Amir Gholami
Under review

[O2] Power-efficient Reinforcement Learning with Neural Processing Unit Oct. 2025 ~ Present
Enabling Group Relative Policy Optimization (GRPO) training with device segregation for rollout and backprop

[O1] Agentic AI: What We Need to Do Next Jul. 2025 ~ Present
Minjae Lee*, Joonwon Lee*, Seunghyuk Oh*, Jeehoon Kang
Investigating multi-dynamic inference with Neural Processing Unit (NPU) beyond single-inference benchmarks

TEACHING EXPERIENCE AND INVITED TALKS

Optimizing AI Efficiency from Silicon to Model, MODULABS 2026
Invited talk, on the growing importance of inference over training in the LLM era, and how FuriosaAI's vision and full-stack approach drive efficient inference, 100+ attendees

Language Models in Clinical ML: My Journey with Two Questions, 2025
Center for Advanced Medical Computing and Analysis (CAMCA), Harvard Medical School (HMS)/Massachusetts General Hospital (MGH)
Research talk, scalable and unified text-based approach for EHRs and cost-efficient inference for language models

How Will DeepSeek-R1 Impact Education's Future? College of Education, Chungnam National University 2025
Invited talk, LLM basics including DeepSeek-R1 and its educational impacts, 30+ educational staff and students

Summer Session for Medical AI, Korean Society of Artificial Intelligence in Medicine (KoSAIM) Summer 2023
Introductory course, basic theory and practice on CNN for Chest X-ray classification, 100+ medical staff and students

AI Short Course Program, Korean Artificial Intelligence Association (KAIA) Spring 2023
Introductory course, basic theory and practice for Diffusion models, 100+ attendees

Machine Learning for Healthcare (AI612), KAIST Spring 2022
Graduate level course, 200+ students (Instructor: Edward Choi)

Programming for AI (AI504), KAIST Fall 2021, Fall 2022
Graduate level course, 200+ students (Instructor: Edward Choi) 🏆***Best lecture award (Fall 2021)***

ACADEMIC AND LEADERSHIP INITIATIVES

Lab Synchronization Seminar, GSAI KAIST Sep. 2022 ~ Mar. 2023
Launched seminar session series for sharing and discussing the ongoing research topics

Lab Tech-talk for Diffusion models, GSAI KAIST Mar. 2022 ~ Mar. 2023
Proceeded Tech-talk seminar sessions for 'Improved Denoising Diffusion Probabilistic Model (DDPM)', 'Classifier-Free Diffusion Guidance', 'Structured Denoising Diffusion Models in Discrete State-Space (D3PM)'

Head of the Bureau of Social Participation, KAIST Undergraduate Student Council Sep. 2017 ~ Dec. 2018
Directed projects to ensure labor rights, learning rights and privacy of campus members

ACADEMIC SERVICES (PAPER REVIEWING)

International Conference on Machine Learning (ICML) 2026

Machine Learning for Health Symposium (ML4H) 2023-2025

Machine Learning for Healthcare (MLHC) 2025

AWARDS AND HONORS

Summa Cum Laude, KAIST 2021
National Science and Technology Scholarship, Ministry of Science, Korea 2021 ~ 2022
Outbound Exchange Student Scholarship, Mirae Asset 2017

Academic Excellence Scholarship *(2nd place)*, KAIST 2016
National Science and Technology Scholarship *(merit-based)*, Ministry of Science, Korea 2015 ~ 2016
Undergraduate School Fellowship, KAIST 2013 ~ 2016