Minjae Lee

mjbooo@kaist.ac.kr | Homepage | LinkedIn

RESEARCH INTERESTS

Developing active agents for medical decision-making that learn generalizable knowledge from heterogeneous healthcare data and adapt to local contexts through inference.

- **Inference in Language Models:** Efficient and agentic inference for scalable language models, with a focus on speculative decoding, test-time compute scaling, and reasoning.
- **Healthcare Machine Learning:** Automated clinical decision-making using electronic health records (EHRs), including representation learning, encoding, synthesis, and privacy-preserving data sharing.

EDUCATION

Korea Advanced Institute of Science and Technology (KAIST), *Seongnam, South Korea* 03/2023
M.S. in Kim Jaechul Graduate School of AI (GSAI)
Thesis: Exploring Optimal Encoders for Electronic Health Records
Advisor: Edward Choi

Korea Advanced Institute of Science and Technology (KAIST), *Daejeon, South Korea* 02/2021
B.S. in Industrial & Systems Engineering (ISE); Minor in Economics
Honors: Summa Cum Laude (ranked 1st in department; GPA 4.0/4.3)

Technical University of Berlin (Technische Universität Berlin, TUB), *Berlin, Germany* 04/2017 – 07/2017
KAIST Outbound Exchange Program in Informatics (Informatik)
Honors: Mirae Asset Outbound Exchange Student Scholarship

SELECTED PUBLICATIONS AND RESEARCH

(* Equal contribution; † Equal advising)

[17] TABED: Test-Time Adaptive Ensemble Drafting for Robust Speculative Decoding in LVLMs
Minjae Lee*, Wonjun Kang*, Byeongkeun Ahn, Christian Classen, …, Kangwook Lee [Paper] [arXiv:2601.20357]
European Chapter of the Association for Computational Linguistics Findings **[EACL 2026 Findings]**
ICLR Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models **[ICLRW 2025]**

[16] EfficientRollout: System-Aware Self-Speculative Decoding for RL Rollouts
Minseo Kim*, **Minjae Lee***, …, Coleman Hooper, Harman Singh, Amir Gholami, Wonjun Kang [Paper] [arXiv:2606.18967]
ICML Workshop on Resource-Adaptive Foundation Model Inference **[ICMLW 2026]**

[15] VersaPRM: Multi-Domain Process Reward Model via Synthetic Reasoning Data
Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, **Minjae Lee**, …, Kannan Ramchandran, Dimitris Papailiopoulos, Kangwook Lee [Paper] [arXiv:2502.06737]
International Conference on Machine Learning ***Oral (top ~1%)*** **[ICML 2025 Oral]**

[14] Rediscovery of CNN's Versatility for Text-based Encoding of Raw Electronic Health Records
Eunbyeol Cho*, **Minjae Lee***, Kyunghoon Hur, …, Jinsung Yoon, Edward Choi [Paper] [arXiv:2303.08290]
Conference on Health, Inference, and Learning ***Oral (top ~10 papers)*** **[CHIL 2023 Oral]**

[13] Privacy-preserving synthetic data enhances postoperative acute kidney injury prediction in data-scarce clinical settings: a multicenter modeling study
Soie Kwon (MD), Eunbyeol Cho, **Minjae Lee**, …, Kwangsoo Kim, Hyung-Jin Yoon (MD), Edward Choi†, Hajeong Lee (MD)† [Lancet Preprint:5416409]
Under review

[12] Generating Multi-Table Time Series EHR from Latent Space with Minimal Preprocessing
Eunbyeol Cho, Jiyoun Kim, **Minjae Lee**, Sungjin Park, Edward Choi [Paper] [arXiv:2507.06996]
Conference on Neural Information Processing Systems **[NeurIPS 2025]**

EMPLOYMENT

AI Algorithm Researcher, FuriosaAI, *Seoul, South Korea* 03/2023 – Present
Conducted research on efficient LLM inference and test-time scaling, contributing to publications at leading AI/ML venues.
Research Partnerships & Academic Collaborations: Berkeley AI Research (BAIR), UW–Madison.

Post-M.S. Research Assistant, KAIST, *Seongnam, South Korea* 07/2023 – 07/2024
Conducted research on EHR synthesis and transfer learning with clinical collaborators, contributing to a multicenter study.

Research Intern, NAVER CLOVA, *Seongnam, South Korea* 12/2021 – 02/2022
Investigated model quantization and compression techniques for efficient AI inference.

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 FuriosaAI's full-stack approach to 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 on scalable, unified text-based approaches 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 on LLM fundamentals, 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 on CNN theory and practice for chest X-ray classification, 100+ medical staff and students.

AI Short Course Program, Korean Artificial Intelligence Association (KAIA) Spring 2023
Introductory course on diffusion model theory and practice, 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 09/2022 – 03/2023
Launched a seminar series for sharing and discussing ongoing research topics.

Lab Tech-talk for Diffusion models, GSAI KAIST 03/2022 – 03/2023
Led technical seminar sessions on diffusion models, including Improved DDPM, classifier-free guidance, and structured discrete-state diffusion models.

Head of the Bureau of Social Participation, KAIST Undergraduate Student Council 09/2017 – 12/2018
Directed projects to support labor rights, learning rights, and privacy for campus members.

ACADEMIC SERVICE

Reviewer: ICML (2026); Machine Learning for Health Symposium (2023–2025); Machine Learning for Healthcare (2025–2026)

SELECTED AWARDS AND HONORS

Summa Cum Laude (ranked 1st in department; GPA 4.0/4.3), KAIST 2021

National Science and Technology Scholarship, Ministry of Science, Korea 2021 – 2022

Outbound Exchange Student Scholarship, Mirae Asset 2017

Dean's List (Departmental Academic Excellence Scholarship, *2nd place*), ISE, KAIST 2016

National Science and Technology Scholarship *(merit-based)*, Ministry of Science, Korea 2015 – 2016

National Science and Technology Scholarship, Ministry of Science, Korea 2013 – 2016

ADDITIONAL PUBLICATIONS

Healthcare Machine Learning

[2] 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
Sehoon Park (MD), Soomin Chung, Yisak Kim, Sun-Ah Yang, Soie Kwon (MD), Jeong Min Cho, **Minjae Lee**, …, Hyung Jin Yoon (MD), Edward Choi, Kwangsoo Kim, Hajeong Lee (MD) [Paper]
Public Library of Science Medicine **[PLOS Medicine 2025]**

[1] Genhpf: General healthcare predictive framework for multi-task multi-source learning
Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyoun Kim, **Minjae Lee**, …, Young-Hak Kim (MD), Louis Atallah, Edward Choi [Paper] [arXiv:2207.09858]
IEEE Journal of Biomedical and Health Informatics **[JBHI 2023]**

Efficient and Agentic Language Model Inference

[11] AsyncOPD: How Stale Can On-Policy Distillation Be?
Wonjun Kang, Kevin Galim, Seunghyuk Oh, Minjun Kang, Sanghyun Park, Donghoon Kim, **Minjae Lee**, …, Kangwook Lee [Paper]
ICML Workshop on Black-Box Optimization to Reinforcement Learning **[ICMLW 2026]**

[10] XQuant: Breaking the Memory Wall for LLM Inference with KV Cache Rematerialization
Aditya Tomar*, Coleman Hooper*, **Minjae Lee**, …, Michael W. Mahoney, Kurt Keutzer, Amir Gholami [arXiv:2508.10395]
Under review

[9] LoSA: Locality Aware Sparse Attention for Block-Wise Diffusion Language Models
Haocheng Xi, Harman Singh, Yuezhou Hu, Coleman Hooper, Rishabh Tiwari, Aditya Tomar, **Minjae Lee**, …, Michael Mahoney, Chenfeng Xu, Kurt Keutzer, Amir Gholami [Paper] [arXiv:2604.12056]
International Conference on Machine Learning **[ICML]**

[8] Draft-based Approximate Inference for LLMs
Kevin Galim, Ethan Ewer, Wonjun Kang, **Minjae Lee**, …, Kangwook Lee [Paper] [arXiv:2506.08373]
International Conference on Learning Representations **[ICLR]**

[7] ParallelBench: Understanding the Trade-offs of Parallel Decoding in Diffusion LLMs
Wonjun Kang, Kevin Galim, Seunghyuk Oh, **Minjae Lee**, …, Coleman Hooper, Kangwook Lee [Paper] [arXiv:2510.04767]
International Conference on Learning Representations **[ICLR]**

[6] Inference-Aligned SFT for Diffusion LLMs via Group-based Trajectory Sampling
Seunghyuk Oh, **Minjae Lee**, Kevin Galim, …, Hanbaek Lyu, Kangwook Lee [Paper]
ICLR Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy **[ICLRW 2026]**

[5] Transformers in the Dark: Navigating Unknown Search Spaces via Bandit Feedback
Jungtaek Kim, Thomas Zeng, Ziqian Lin, **Minjae Lee**, …, Kangwook Lee [Paper] [arXiv:2603.24780]
Transactions on Machine Learning Research **[TMLR]**
NeurIPS Workshop on What Can't Transformers Do? **[NeurIPSW 2025]**

[4] UNCAGE: Contrastive Attention Guidance for Masked Generative Transformers in Text-to-Image Generation
Wonjun Kang, Byeongkeun Ahn, **Minjae Lee**, Kevin Galim, …, Nam Ik Cho [Paper] [arXiv:2508.05399]
Institute of Electrical and Electronics Engineers Access **[IEEE Access]**

[3] State-offset Tuning: State-based Parameter-Efficient Fine-Tuning for State Space Models
Wonjun Kang*, Kevin Galim*, Yuchen Zeng, **Minjae Lee**, …, Nam Ik Cho [Paper] [arXiv:2503.03499]
Association for Computational Linguistics **[ACL 2025]**