Seanie Lee

Date of Birth: 1992.04.17, Nationality: Korean

126 Yangjae-dong, Seocho District, Seoul

(+82) 10-4475-2273 | lsnfamily02@kaist.ac.kr | seanie12.github.io | seanie12 | Seanie Lee

Education

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

Ph.D in Artificial Intelligence — *Mar. 2022 - present*

- Supervised by Sung Ju Hwang and Juho Lee
- Research interest: AI safety, responsible AI, and evaluation

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

M.S. in Artificial Intelligence — *Mar. 2020 - Feb. 2022*

- Supervised by Sung Ju Hwang and Juho Lee
- Master Thesis: Data augmentation for natural language processing

Yonsei University — *Seoul, S.Korea*

B.A. in Library and Information Science — *Mar. 2011 - Feb. 2018*

Experience

Apple — *Seattle, US*

Internship — *October 2025 - May 2026*

- Machine Learning Research, hosted by Raviteja Vemulapalli.
- Synthetic data generation for tool-calling LLMs.

Krafton — *Seoul, Korea*

Internship — *July 2025 - Oct 2025*

- Research Internship
- Safety alignment of long reasoning models.

Mila — *Montreal, Canada*

Internship — *January 2024 - June 2024*

- Research internship at Mila, advised by Yoshua Bengio.
- Robust red-teaming of LLMs.

Apple — *Cambridge, UK*

Internship — *May 2023 - September 2023*

- Research internship at Siri team, hosted by Anders Johannsen.
- Few-shot example retrieval for ICL.

National University of Singapore (NUS) — *Singapore*

Internship — *July 2022 - September 2022*

- Remote internship at Deep Learning lab, supervised by Kenji Kawaguchi.
- Regularization for continual pretraining.

Awards

2023 **Apple AI/ML PhD Fellowship**, Recipient of Apple Scholars in AI/ML — *Cupertino, US*

2022 **Google Travel Grant**, NeurIPS 2022 — *US*

2019 **Silver Medal**, Named Entity Recognition in NAVER NLP Challenge — *Seoul, Korea*

Presentation

Seminar at Korea University. *Seoul, Korea*

Presentation of large scale set-encoding *May. 2025*

- Synthetic Data Generation for LLM Safeguards
- ICLR 2025, ACL Findings 2025

Seminar at Hanyang University. *Seoul, Korea*

Presentation of large scale set-encoding *April. 2025*

- Synthetic Data Generation for LLM Safeguards
- ICLR 2025, ACL Findings 2025

Tech. Talk, Nuremberg Institute of Technology Georg Simon Ohm. *Nürnberg, Germany*

Presentation of large scale set-encoding *Oct. 2023*

- Scalable Set Encoding with Universal Mini-Batch Consistency and Unbiased Full Set Gradient Approximation
- ICML 2023

Tech. talk, Samgsung SDS. *Seoul, South Korea*

Presentation of large scale set-encoding *22.May. 2023*

- Scalable Set Encoding with Universal Mini-Batch Consistency and Unbiased Full Set Gradient Approximation
- ICML 2023

Tech. talk, NAVER corp. *Online, South Korea*

Presentation of Info-HCVAE *04.Dec. 2020*

- Generating Diverse and Consistent QA pairs from Contexts with Information-Maximizing Hierarchical Conditional VAEs
- ACL 2020 Long paper

Selected Publication

(* indicates equal contribution)

Preprint

HoliSafe: Holistic Safety Benchmarking and Modeling with Safety Meta Token for Vision-Language Model *Arxiv*

Youngwan Lee, Kangsan Kim, Kwanyong Park, Ilcahe Jung, Soojin Jang, **Seanie Lee**, Yong-Ju Lee and Sung Ju Hwang *2025*

- [paper][code]

Conferences

Learning Diverse Attacks on Large Language Models for Robust Red-teaming and Safety Tuning *ICLR*

Seanie Lee, Minsu Kim, Lynn Cherif, David Dobre, Juho Lee, Sung Ju Hwang, Kenji Kawaguchi, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, Moksh Jain *2025*

- [paper][code]

HarmAug: Effective Data Augmentation for Knowledge Distillation of Safety Guard Models *ICLR*

Seanie Lee*, Haebin Seong*, Dong Bok Lee, Minki Kang, Xiaoyin Chen, Dominik Wagner, Yoshua Bengio, Juho Lee, Sung Ju Hwang *2025*

- [paper][code]

SafeRoute: Adaptive Model Selection for Efficient and Accurate Safety Guardrails in Large Language Models *ACL Findings*

Seanie Lee*, Dong Bok Lee*, Dominik Wagner, Minki Kang, Haebin Seong, Tobias Bocklet, Juho Lee, Sung Ju Hwang *2025*

- [paper][code]

Reliable Decision-Making via Calibration-Oriented Retrieval-Augmented Generation *NeurIPS*

Chaeyun Jang, Deukhwan Cho, **Seanie Lee**, Juho Lee and Hyungi Lee *2025*

- [paper][code]

FedSVD: Adaptive Orthogonalization for Private Federated Learning with LoRA *NeurIPS*

Seanie Lee*, Sangwoo Park*, Dong Bok Lee*, Dominik Wagner, Haebin Seong, Tobias Bocklet, Juho Lee, Sung Ju Hwang *2025*

- [paper][code]

Margin-based Neural Network Watermarking — *ICML* 2023

Byungjoo Kim, Suyoung Lee, **Seanie Lee**, Sooel Son, Sung Ju Hwang

- [paper]

Publication

(* indicates equal contribution)

Preprint

Rethinking Reward Models for Multi-Domain Test-Time Scaling — *Arxiv* 2025

Dong Bok Lee*, **Seanie Lee***, Sangwoo Park, Minki Kang, Jinheon Baek, Dongki Kim, Dominik Wagner, Jiongdao Jin, Heejun Lee, Tobias Bocklet, Jinyu Wang, Jingjing Fu, Sung Ju Hwang, Jiang Bian and Lei Song

- [paper][code]

HoliSafe: Holistic Safety Benchmarking and Modeling with Safety Meta Token for Vision-Language Model — *Arxiv* 2025

Youngwan Lee, Kangsan Kim, Kwanyong Park, Ilcahe Jung, Soojin Jang, **Seanie Lee**, Yong-Ju Lee and Sung Ju Hwang

- [paper][code]

Conferences

FedSVD: Adaptive Orthogonalization for Private Federated Learning with LoRA — *NeurIPS* 2025

Seanie Lee*, Sangwoo Park*, Dong Bok Lee*, Dominik Wagner, Haebin Seong, Tobias Bocklet, Juho Lee, Sung Ju Hwang

- [paper][code]

Distilling LLM Agent into Small Models with Retrieval and Code Tools — *NeurIPS Spotlight* 2025

Minki Kang, Jongwon Jeong, **Seanie Lee**, Jaewoong Cho and Sung Ju Hwang

- [paper][code]

Reliable Decision-Making via Calibration-Oriented Retrieval-Augmented Generation — *NeurIPS* 2025

Chaeyun Jang, Deukhwan Cho, **Seanie Lee**, Juho Lee and Hyungi Lee

- [paper][code]

Trajectory Balance with Asynchrony: Decoupling Exploration and Learning for Fast, Scalable LLM Post-Training — *NeurIPS* 2025

Brian R. Bartoldson, Siddarth Venkatraman, James Diffenderfer, Moksh Jain, Tal Ben-Nun, **Seanie Lee**, Minsu Kim, Johan Obando-Ceron, Yoshua Bengio and Bhavya Kailkhura

- [paper][code]

SafeRoute: Adaptive Model Selection for Efficient and Accurate Safety Guardrails in Large Language Models — *ACL Findings* 2025

Seanie Lee*, Dong Bok Lee*, Dominik Wagner, Minki Kang, Haebin Seong, Tobias Bocklet, Juho Lee, Sung Ju Hwang

- [paper][code]

Personalized Fine-Tuning with Controllable Synthetic Speech from LLM-Generated Transcripts for Dysarthric Speech Recognition — *Interspeech* 2025

Dominik Wagner, Ilja Baumann, Natalie Engert, **Seanie Lee**, Elmar Nöth, Korbinian Riedhammer and Tobias Bocklet

- [paper]

HarmAug: Effective Data Augmentation for Knowledge Distillation of Safety Guard Models — *ICLR* 2025

Seanie Lee*, Haebin Seong*, Dong Bok Lee, Minki Kang, Xiaoyin Chen, Dominik Wagner, Yoshua Bengio, Juho Lee, Sung Ju Hwang

- [paper][code]

Learning Diverse Attacks on Large Language Models for Robust Red-teaming and Safety Tuning — *ICLR* 2025

Seanie Lee, Minsu Kim, Lynn Cherif, David Dobre, Juho Lee, Sung Ju Hwang, Kenji Kawaguchi, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, Moksh Jain

- [paper][code]

Optimized Speculative Sampling for GPU Hardware Accelerators — *EMNLP* 2024

Dominik Wagner, **Seanie Lee**, Ilja Baumann, Philipp Seeberger, Korbinian Riedhammer, Tobias Bocklet

- [paper][code]

Drug Discovery with Dynamic Goal-aware Fragment — *ICML* 2024

Seul Lee, **Seanie Lee**, Kenji Kawaguchi, Sung Ju Hwang

- [paper][code]

Effective and Efficient Conversation Retrieval for Dialogue State Tracking with Implicit Text Summaries — *NAACL* 2024

Seanie Lee, Jianpeng Cheng, Joris Driesen, Alexandru Coca, Anders Johannsen

- [paper]

Self-Supervised Dataset Distillation for Transfer Learning — *ICLR* 2024

Dong Bok Lee*, **Seanie Lee***, Joonho Ko, Kenji Kawaguchi, Juho Lee, Sung Ju Hwang

- [paper][code]

DiffusionNAG: Task-guided Neural Architecture Generation with Diffusion Models — *ICLR* 2024

Sohyun Ahn*, Hayeon Lee*, Jaehyeong Jo, **Seanie Lee**, Sung Ju Hwang

- [paper][code]

Scalable Set Encoding with Universal Mini-Batch Consistency and Unbiased Full Set Gradient Approximation — *ICML* 2023

Jeffrey Willette*, **Seanie Lee***, Bruno Andreis, Kenji Kawaguchi, Juho Lee, Sung Ju Hwang

- [paper][code]

Margin-based Neural Network Watermarking — *ICML* 2023

Byungjoo Kim, Suyoung Lee, **Seanie Lee**, Sooel Son, Sung Ju Hwang

- [paper]

Self-Supervised Set Representation Learning for Unsupervised Meta-Learning — *ICLR* 2023

Dong Bok Lee*, **Seanie Lee***, Kenji Kawaguchi, Yunji Kim, Jihwan Bang, Jung-Woo Ha, Sung Ju Hwang

- [paper]

Self-Distillation for Further Pre-training of Transformers — *ICLR* 2023

Seanie Lee, Minki Kang, Juho Lee, Sung Ju Hwang, Kenji Kawaguchi

- [paper][code]

Set-based Meta-Interpolation for Few-Task Meta-Learning — *NeurIPS* 2022

Seanie Lee*, Bruno Andreis*, Kenji Kawaguchi, Sung Ju Hwang

- [paper] [code]

On Divergence Measures for Bayesian Pseudocoresets — *NeurIPS* 2022

Balhae Kim, Jungwon Choi, **Seanie Lee**, Yoonho Lee, Jung-Woo Ha, Juho Lee

- [paper]

Set Based Stochastic Subsampling — *ICML* 2022

Bruno Andreis, **Seanie Lee**, A. Tuan Nguyen, Juho Lee, Eunho Yang, Sung Ju Hwang

- [paper]

Sequential Reptile: Inter-Task Gradient Alignment for Multilingual Learning — *ICLR* 2022

Seanie Lee*, Hae Beom Lee*, Juho Lee, Sung Ju Hwang

- [paper]

Learning to Perturb Word Embeddings for Out-of-distribution QA — *ACL* 2021

Seanie Lee*, Minki Kang*, Juho Lee, Sung Ju Hwang

- [paper][code]

Contrastive Learning with Adversarial Perturbations for Conditional Text Generation — *ICLR* 2021

Seanie Lee*, Dong Bok Lee*, Sung Ju Hwang

- [paper][code]

Meta-GMVAE: Mixture of Gaussian VAE for Unsupervised Meta-Learning *ICLR*

Dong Bok Lee, Dongchan Min, **Seanie Lee**, Sung Ju Hwang *2021*

- [paper][code]

Generating Diverse and Consistent QA pairs from Contexts with Information-Maximizing Hierarchical Conditional VAEs *ACL*

Dong Bok Lee*, **Seanie Lee***, WooTae Jeong, Donghwan Kim, Sung Ju Hwang *2020*

- [paper] [code][video]

g2pM: A Neural Grapheme-to-Phoneme Conversion Package for Mandarin Chinese Based on a New Open Benchmark Dataset *INTERSPEECH*

Kyubyong Park*, **Seanie Lee*** *2020*

- [paper][code]

References

Sung Ju Hwang *Advisor*

Associate Professor in KAIST. *2020-2025*

e-mail: sjhwang82@kaist.ac.kr.

Juho Lee *Advisor*

Associate Professor in KAIST. *2020-2025*

e-mail: juholee@kaist.ac.kr.

Yoshua Bengio *Collaborator*

Full Professor at Université de Montréal and Scientific Director of Mila – Quebec AI Institute. *2024-2025*

e-mail: yoshua.bengio@mila.quebec.

Kenji Kawaguchi *Collaborator*

Presidential Young Professor in the Department of Computer Science at NUS. *2022-present*

e-mail: kenji@comp.nus.edu.sg

Nikolay Malkin *Collaborator*

Chancellor's Fellow at University of Edinburgh, School of Informatics *2024*

e-mail: nmalkin@ed.ac.uk