Seanie **Lee**

LLMs · Safety Alignment · Efficiency

291 Daehak-ro, Yuseong-gu, Daejeon, Korea 34141

✉ 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 - Feb. 2025 (Expected)*

- Supervised by Sung Ju Hwang and Juho Lee
- Research interest: Safety and Efficiency of Large Language Models.

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

Mila *Montreal, Canada*

Internship *January 2024 - June 2024*

- Research internship at Mila, advised by Yoshua Bengio.

Apple *Cambridge, UK*

Internship *May 2023 - September 2023*

- Research internship at Siri team, hosted by Anders Johannsen.

Singapore National University *Singapore*

Internship *July 2022 - September 2022*

- Remote internship at Deep Learning lab, supervised by Kenji Kawaguchi.

Korea Advanced Institute of Science and Technology *Daejeon, S.Korea*

Teaching Assistant *Mar. 2020 - Dec. 2021*

- Deep Reinforcement Learning, AI611
- Mathematics for AI, AI503
- Deep Learning, AI502

42 Maru *Seoul, S.Korea*

Internship *Feb. 2019 - Jan. 2020*

- Research on Question Answering, Semi-supervised Learning, Domain Generalization

Awards

2023	**Apple AI/ML PhD Fellowship**, Recipient of Apple Scholars in AI/ML	*Coupertino, US*
2022	**Google Travel Grant**, NeurIPS 2022	*US*
2019	**Silver Medal**, Named Entity Recognition in NAVER NLP Challenge	*Seoul, Korea*

Presentation

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

Publication

(* indicates equal contribution)

PREPRINT

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

SEANIE LEE*, HAEBIN SEONG*, DONG BOK LEE, MINKI KANG, XIAOYIN CHEN, DOMINIK WAGNER, YOSHUA BENGIO, JUHO LEE, SUNG JU HWANG *2024*

- [paper]

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

SEANIE LEE, MINSU KIM, LYNN CHERIF, DAVID DOBRE, JUHO LEE, SUNG JU HWANG, KENJI KAWAGUCHI, GAUTHIER GIDEL, YOSHUA BENGIO, NIKOLAY MALKIN, MOKSH JAIN *2024*

- [paper]

CONFERENCES

Optimized Speculative Sampling for GPU Hardware Accelerators *EMNLP*

DOMINIK WAGNER, **SEANIE LEE**, ILJA BAUMANN, PHILIPP SEEBERGER, KORBINIAN RIEDHAMMER, TOBIAS BOCKLET *2024*

- [paper][code]

Drug Discovery with Dynamic Goal-aware Fragment *ICML*

SEUL LEE, **SEANIE LEE**, KENJI KAWAGUCHI, SUNG JU HWANG *2024*

- [paper][code]

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

SEANIE LEE, JIANPENG CHENG, JORIS DRIESEN, ALEXANDRU COCA, ANDERS JOHANNSEN *2024*

- [paper]

Self-Supervised Dataset Distillation for Transfer Learning *ICLR*

DONG BOK LEE*, **SEANIE LEE***, JOONHO KO, KENJI KAWAGUCHI, JUHO LEE, SUNG JU HWANG *2024*

- [paper][code]

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

SOHYUN AHN*, HAYEON LEE*, JAEHYEONG JO, **SEANIE LEE**, SUNG JU HWANG *2024*

- [paper][code]

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

JEFFREY WILLETTE*, **SEANIE LEE***, BRUNO ANDREIS, KENJI KAWAGUCHI, JUHO LEE, SUNG JU HWANG *2023*

- [paper][code]

Margin-based Neural Network Watermarking *ICML*

BYUNGJOO KIM, SUYOUNG LEE, **SEANIE LEE**, SOOEL SON, SUNG JU HWANG *2023*

- [paper]

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

DONG BOK LEE*, **SEANIE LEE***, KENJI KAWAGUCHI, YUNJI KIM, JIHWAN BANG, JUNG-WOO HA, SUNG JU HWANG *2023*

- [paper]

Self-Distillation for Further Pre-training of Transformers
ICLR

Seanie Lee, Minki Kang, Juho Lee, Sung Ju Hwang, Kenji Kawaguchi
2023

- [paper][code]

Set-based Meta-Interpolation for Few-Task Meta-Learning
NeurIPS

Seanie Lee*, Bruno Andreis*, Kenji Kawaguchi, Sung Ju Hwang
2022

- [paper] [code]

On Divergence Measures for Bayesian Pseudocoresets
NeurIPS

Balhae Kim, Jungwon Choi, **Seanie Lee**, Yoonho Lee, Jung-Woo Ha, Juho Lee
2022

- [paper]

Set Based Stochastic Subsampling
ICML

Bruno Andreis, **Seanie Lee**, A. Tuan Nguyen, Juho Lee, Eunho Yang, Sung Ju Hwang
2022

- [paper]

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

Seanie Lee*, Hae Beom Lee*, Juho Lee, Sung Ju Hwang
2022

- [paper]

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

Seanie Lee*, Minki Kang*, Juho Lee, Sung Ju Hwang
2021

- [paper][code]

Contrastive Learning with Adversarial Perturbations for Conditional Text Generation
ICLR

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

- [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]