

Seungyoon Lee

NLP Researcher · Graduate Student, Korea University

🏠 Sep. 30th, 1996 · ✉️ dltmddb100@korea.ac.kr

🐙 Github · 🔗 LinkedIn · 🎓 Google Scholar

BIOGRAPHY

Integrated M.S./Ph.D. researcher at the Korea University NLP & AI Lab, advised by Prof. Heuseok Lim, specializing in large language models and natural language processing — with a focus on Cross-lingual Transfer, Policy Optimization, Information Retrieval, and Retrieval Augmented Generation. My research philosophy centers on bridging the gap between theoretical Natural Language Processing (NLP) and its real-world execution, ensuring that foundational advances translate into scalable, high-impact applications. My primary focus lies in the mitigation of language disparity and advancement of Foundation Models, where I investigate scalable Representation Learning by exploring the way of transferring the capacity of high-resource to low-resource and developing new training paradigm for language model.

EDUCATION

- **Korea University** Seoul, Republic of Korea
Ph.D. in Computer Science & Engineering *Aug. 2023 - Aug. 2028 (Expected)*
Advisor: Heuseok Lim
- **Chung-ang University** Seoul, Republic of Korea
Bachelor of Applied Statistics *Mar. 2016 - Aug. 2023*

INTERNATIONAL PUBLICATIONS

*, †: Equal Contribution, Corresponding Author

2026

- **Language Chain in Alignment: Cross-Lingual Ranking Preference Optimization [pdf]**
Seungyoon Lee, Minhyuk Kim, Jungseob Lee, Heuseok Lim[†]
EMNLP 2026 - Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing
- **CLEAR: Cross-Lingual Enhancement in Retrieval via Reverse-training [pdf] [code]**
Seungyoon Lee, Minhyuk Kim, Seongtae Hong, Youngjoon Jang, Dongsuk Oh, Heuseok Lim[†]
ACL 2026 - Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
- **CultureConverse: A Multilingual Multi-turn Simulation Harness for Culturally Grounded Assistance in East and Southeast Asia**
Bryan Chen Zhengyu Tan, Weihua Zheng, Thong T. Doan, Bich Ngoc Doan, Jia Wang Peh, . . . , Seungyoon Lee, et al.
EMNLP 2026 - Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing
- **Query-Synergy: Leveraging High-Resource Languages for Improving Retrieval Performance Across Multiple Languages [pdf]**
Seongtae Hong*, Jungseob Lee*, Hyeonseok Moon, Seungyoon Lee, Youngjoon Jang, Heuseok Lim[†]
ACL 2026 MeLLM Workshop - Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models
- **Improving Korean-English Cross-Lingual Retrieval: A Data-Centric Study of Language Composition and Model Merging [pdf]**
Youngjoon Jang, Junyoung Son, Taemin Lee, Seongtae Hong, Hyeonseok Moon, Seungyoon Lee, Andrew Matteson, Heuseok Lim[†]
ACL 2026 MeLLM Workshop - Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models

2025

- **Semantic Aware Linear Transfer by Recycling Pre-trained Language Models for Cross-lingual Transfer [pdf]**
Seungyoon Lee, Seongtae Hong, Hyeonseok Moon, Heuseok Lim[†]
ACL 2025 Findings - Findings of the Association for Computational Linguistics: ACL 2025
- **TORSO: Template-Oriented Reasoning Towards General Tasks [pdf]**
Minhyuk Kim, Seungyoon Lee, Heuseok Lim[†]
EMNLP 2025 (Oral) - Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing
- **Find the Intention of Instruction: Comprehensive Evaluation of Instruction Understanding for Large Language Models [pdf]**
Hyeonseok Moon, Jaehyung Seo, Seungyoon Lee, Chanjun Park[†], Heuseok Lim[†]
NAACL 2025 Findings - Findings of the Association for Computational Linguistics: NAACL 2025
- **FLEX: A Benchmark for Evaluating Robustness of Fairness in Large Language Models [pdf]**
Dahyun Jung*, Seungyoon Lee*, Hyeonseok Moon, Chanjun Park[†], Heuseok Lim[†]
NAACL 2025 Findings - Findings of the Association for Computational Linguistics: NAACL 2025
- **MIGRATE: Cross-Lingual Adaptation of Domain-Specific LLMs through Code-Switching and Embedding Transfer [pdf]**
Seongtae Hong, Seungyoon Lee, Hyeonseok Moon, Heuseok Lim[†]
COLING 2025 - Proceedings of the 31st International Conference on Computational Linguistics

2024

- **Translation of Multifaceted Data without Re-Training of Machine Translation Systems** [pdf]
Hyeonseok Moon, **Seungyoon Lee**, Seongtae Hong, Seungjun Lee, Chanjun Park, Heuiseok Lim[†]
EMNLP 2024 Findings - Findings of the Association for Computational Linguistics: EMNLP 2024
- **Length-aware Byte Pair Encoding for Mitigating Over-segmentation in Korean Machine Translation** [pdf]
Jungseob Lee*, Hyeonseok Moon*, Seungjun Lee, Chanjun Park, Sugyeong Eo, Hyunwoong Ko, Jaehyung Seo, **Seungyoon Lee**, Heuiseok Lim[†]
ACL 2024 Findings - Findings of the Association for Computational Linguistics: ACL 2024
- **Exploring Inherent Biases in LLMs within Korean Social Context: A Comparative Analysis of ChatGPT and GPT-4** [pdf]
Seungyoon Lee, Dongjun Kim, Dahyun Jung, Chanjun Park[†], Heuiseok Lim[†]
NAACL 2024 SRW - Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)
- **Leveraging Pre-existing Resources for Data-Efficient Counter-Narrative Generation in Korean** [pdf] [code]
Seungyoon Lee, Chanjun Park, Dahyun Jung, Hyeonseok Moon, Jaehyung Seo, Sugyeong Eo, Heuiseok Lim[†]
LREC-COLING 2024 (Oral) - Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation

~2023

- **Alternative Speech: Complementary Method to Counter-Narrative for Better Discourse** [pdf]
Seungyoon Lee*, Dahyun Jung*, Chanjun Park*, Seolhwa Lee, Heuiseok Lim[†]
ICDM 2023 DCAI Workshop - Proceedings of the 23rd IEEE International Conference on Data Mining Workshops (ICDMW 2023)
- **Focus on FoCus: Is FoCus focused on Context, Knowledge and Persona?** [pdf]
Seungyoon Lee*, Jungseob Lee*, Chanjun Park, Sugyeong Eo, Hyeonseok Moon, Jaehyung Seo, Jeongbae Park, Heuiseok Lim[†]
COLING 2022 - Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge

PRE-PRINTED PUBLICATIONS

- **LAMAR: An Open Language-Aware Multilingual Alignment Reranker** [pdf] [model]
Seongtae Hong, Youngjoon Jang, Jungseob Lee, **Seungyoon Lee**, Heuiseok Lim[†]
arXiv preprint arXiv:2607.22042
- **Answer-Conditioned Chains of Thought Degrade Verifiable-Reasoning Distillation in Large Language Models** [pdf]
Jungseob Lee, **Seungyoon Lee**, Suhyune Son, Dongyub Jude Lee, Sungbin Han, Sugyeong Eo, Heuiseok Lim[†]
arXiv preprint arXiv:2607.14552
- **Beyond Penalizing Mistakes: Stabilizing Efficiency Training in Large Reasoning Models via Adaptive Correct-Only Rewards** [pdf]
Jungseob Lee*, **Seungyoon Lee***, Seongtae Hong, Minhyuk Kim, Chanjun Park, Heuiseok Lim[†]
arXiv 2026 - arXiv preprint arXiv:2606.22716
- **PRISM: Pruning via Interference-Signal Maximization for LLM Merging** [pdf]
Jungseob Lee, **Seungyoon Lee**, Sugyeong Eo, Hyeonseok Moon, Jaehyung Seo, Heuiseok Lim[†]
- **Skin-Deep: A Geometric Diagnostic for Alignment Fragility in Large Language Model Representations** [pdf] [code]
Dongyub Jude Lee, Jungseob Lee, **Seungyoon Lee**, Seongtae Hong, Suhyune Son, Sugyeong Eo, Jaehyung Seo, Heuiseok Lim[†]
arXiv 2026 - arXiv preprint arXiv:2606.22676

MAJOR PROJECT

- **VAETKI — Korean Sovereign AI Foundation Model** NC AI (NCSoft) - led consortium
Post-training Team Leader Jun. 2025 – Jan. 2026
 - Served as Post-training Team Leader representing Korea University in Korea's national **Sovereign AI Foundation Model** program, organized by the Ministry of Science and ICT — an approximately KRW 200 billion national initiative with staged six-month evaluations running through 2027.
 - Led the full post-training — Supervised Fine-Tuning, Preference Optimization, and Multi-node Distributed Training.
 - **VAETKI model**: an industry-specialized sovereign foundation model, open-sourced on Hugging Face on Dec 30, 2025. A 100B+ parameter Mixture-of-Experts architecture (11B active parameters at inference) with a custom Multi-head Latent Attention. [Model] [Technical Report]
- **KULLM-R — Efficient Korean Reasoning Model** NLP & AI Lab
Project Manager Feb. 2025 – Oct. 2025
 - Served as Project Manager for KULLM-R (Efficient Korean Reasoning Model) project.
 - Efficient Korean reasoning model based on Qwen3-8B via Reinforcement Learning.
 - **KULLM-R**: shows superior reasoning efficiency and mitigates language confusion phenomenon in Korean. [Model]
- **KULLM3 — Language Transfer in Large Language Model** NLP & AI Lab
Lead Researcher Jan. 2024 – July 2024
 - Served as Lead Researcher for KULLM3 (Korea University Large Language Model) project — developing Korean-specialized LLM.
 - Developed an embedding transfer method by diluting the embeddings of existing Language models into target LLMs for smooth language adaptation.
 - **KULLM3**: an advanced instruction-following and fluent chat model specialized in Korean, closely reaching gpt-3.5-turbo. [Model] [Github]

HONORS AND AWARDS

- **Best Paper Award**
Annual Conference on Human and Cognitive Language Technology (HCLT 2025)
KULLM-RAG: A RAG-Specialized Large Language Model with Intelligent Self-Verification
J. Lee, M. Kim, J. Yoon, S. Hong, Y. Jang, S. Lee, J. Seo, C. Park, J. Park, H. Lim[†]
- **Best Paper Award**
Annual Conference on Human and Cognitive Language Technology (HCLT 2023)
KoCheckGPT: Korean LLM written document detector
K. Myeong, J. Lee, S. Lee, S. Hong, J. Park, H. Lim[†]
- **Best Paper Award**
Annual Conference on Human and Cognitive Language Technology (HCLT 2023)
Analysis of Toxicity and Bias of ChatGPT within Korean Social Context
S. Lee, C. Park, K. Kim, H. Lim[†]
- **Ministry of Science and ICT Awards**
ICT Challenge 2nd, 2022
- **Statistics Korea Awards**
Artificial Intelligence Competition 1st, 2022
Designed HiBERT, an innovative architecture that captures hierarchical information of Korea Standard Industry Code

ETC

- **Academic Service**
NeurIPS, ARR (ACL Rolling Review) Reviewer
- **Military obligation**
Republic of Korea Air Force, Sep. 2020 - June 2022