Seongyun Lee

seongyun@kaist.ac.kr | github.com/sylee0520 | https://www.linkedin.com/in/seongyun-lee-647753233 |
Google Scholar: https://scholar.google.com/citations?user=T8Zid6YAAAAJ

PROFILE

I'm Ph.D student at **KAIST AI** and advisded by Prof. Minjoon Seo and Hyunwoo Kim. **I am currently interested in solving misalignment and safety issues in agents.** Previously I obtained B.S in Computer Science at Korea University, M.S. in AI at KAIST AI and also was previously a research intern at Twelve Labs and LG AI Research.

RESEARCH INTERESTS

My research focuses on the alignment and safety of agent, with particular emphasis on social and value alignment, as well as the safety challenges of LLM-/VLM-based agent systems. I am also interested in designing benchmarks and metrics to measure these concerns. The research that best supports this interest is as follows:

Safety & Alignment

- How Does Vision-Language Adaptation Impact the Safety of Vision Language Models? (ICLR 2025)
- Aligning to Thousands of Preferences via System Message Generalization (NeurIPS 2025)
- The CoT Encyclopedia: Analyzing, Predicting, and Controlling how a Reasoning Model will Think (Preprint)

Evaluation

- Prometheus-Vision: Vision-Language Model as a Judge for Fine-Grained Evaluation (ACL 2024 Findings)
- The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models (NAACL 2025 Main, Best Paper Award)

Multimodal & Agent

- Volcano: Mitigating Multimodal Hallucination through Self-Feedback Guided Revision (NAACL 2024)
- Zero-Shot Dense Video Captioning by Jointly Optimizing Text and Moment (Preprint)
- Paper2Code: Automating Code Generation from Scientific Papers in Machine Learning (Preprint)

EDUCATION

Sep 2025 —	Ph.D. **Kim Jaechul Graduate School of AI** Korea Advanced Institute of Science and Technology	Seoul
Mar 2024 — Aug 2025	M.S. **Kim Jaechul Graduate School of AI** Korea Advanced Institute of Science and Technology	Seoul
Mar 2020 — Feb 2024	B.S. **Computer Science and Engineering** Korea University	Seoul

PUBLICATIONS

May 2025

The CoT Encyclopedia: Analyzing, Predicting, and Controlling how a Reasoning Model will Think

Seongyun Lee, Seungone Kim, Minju Seo, Yongrae Jo, Dongyoung Go, Hyeonbin Hwang, Jinho Park, Xiang Yue, Sean Welleck, Graham Neubig, Moontae Lee, Minjoon Seo

Preprint

Apr 2025

Paper2Code: Automating Code Generation from Scientific Papers in Machine Learning

Minju Seo, Jinheon Baek, **Seongyun Lee**, Seong Ju Hwang

Preprint

Mar 2025

Scaling Evaluation-time Compute with Reasoning Models as Process Evaluators

Seungone Kim, Ian Wu, Jinu Lee, Xiang Yue, **Seongyun Lee**, Mingyeong Moon, Kiril Gashteovski, Carolin Lawrence, Julia Hockenmaier, Graham Neubig, Sean Welleck

Preprint

Dec 2024

Efficient Long Context Language Model Retrieval with Compression

Minju Seo, Jinheon Baek, **Seongyun Lee**, Seong Ju Hwang

ACL 2025 Main

Dec 2024

Evaluating Language Models as Synthetic Data Generators

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, **Seongyun Lee**, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, Graham Neubig

ACL 2025 Main

Oct 2024

How Does Vision-Language Adaptation Impact the Safety of Vision Language Models?

Seongyun Lee*, Geewook Kim*, Jiyeon Kim*, Hyunji Lee, Hoyeon Chang, Sue Hyun Park, Minjoon Seo

(*equal contribution)

ICLR 2025

Jun 2024

The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, **Seongyun Lee**, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, Minjoon Seo

NAACL 2025 Main (Oral and Winner of best-paper award)

May 2024

Aligning to Thousands of Preferences via System Message Generalization

Seongyun Lee*, Sue Hyun Park*, Seungone Kim, Minjoon Seo

(*equal contribution)

NeurIPS 2024

[Paper] [Code]

Apr 2024

LG AI Research & KAIST at EHRSQL 2024: Self-Training Large Language Models with Pseudo-Labeled Unanswerable Questions for a Reliable Text-to-SQL System on EHRs

Yongrae Jo*, **Seongyun Lee***, Minju Seo*, Sung Ju Hwang, Moontae Lee

(*equal contribution)

NAACL 2024 ClinicalNLP Workshop (Oral)

Feb 2024

Prometheus-Vision: Vision-Language Model as a Judge for Fine-Grained Evaluation

Seongyun Lee*, Seungone Kim*, Sue Hyun Park, Geewook Kim, Minjoon Seo

(*equal contribution)

ACL 2024 Findings

[Paper] [Code]

Nov 2023

Volcano: Mitigating Multimodal Hallucination through Self-Feedback Guided Revision

Seongyun Lee, Sue Hyun Park, Yongrae Jo, Minjoon Seo

NAACL 2024 Main

[Paper] [Code]

Jun 2023

Zero-Shot Dense Video Captioning by Jointly Optimizing Text and Moment

Yongrae Jo, **Seongyun Lee**, Aiden SJ Lee, Hyunji Lee, Hanseok Oh, Minjoon Seo

Preprint

Nov 2022

LIQUID: A Framework for List Question Answering Dataset Generation

Seongyun Lee*, Hyunjae Kim*, Jaewoo Kang (*equal contribution)

AAAI 2023

[Paper] [Code] [Leaderboard]

EXPERIENCE

Apr 2025 — Dec 2025

Research Intern, LG AI Research Seoul

Woking on code generation LLM and AI Agent

Jun 2023 — Oct 2023

ML Research Intern, Twelve Labs Seoul

Woking on building a video-language model at ML Modeling Team

One of the main inventor of Video-Language Foundation Model ‘Pegasus-1’

Dec 2022 — Feb 2024 **Research Intern, LK Lab, KAIST AI (Prof. Minjoon Seo)** Seoul

Working on validating various hypotheses related to video retrieval/captioning with advisor. Proposed some methods of zero-shot dense video captioning, conducted diverse experiments, visualized results. Finally, submitted a paper to the AI conference.

Jul 2021 — Dec 2022 **Research Intern, DMIS Lab, Korea University (Prof. Jaewoo Kang)** Seoul

Participated in BioASQ challenge and took 2nd place in some batches. Following the challenge, studied the methodology of augmenting the data needed for list question answering, and submitted the paper to AAAI 2023.

REVIEW EXPERIENCES

- NAACL 2024
- ACL 2024
- NeurIPS 2024
- ICLR 2025
- NeurIPS 2025
- ICLR 2026

AWARD

Apr 2024 **1st place, Reliable Text-to-SQL Modeling on Electronic Health Records**

NAACL 2024 ClinicalNLP Workshop Shared Task

Sep 2022 **2nd place, BioASQ 10B Batch 3: semantic QA**

BioASQ 2022

Sep 2022 **2nd place, BioASQ 10B collaborative: semantic QA**

BioASQ 2022

Mar 2020 — Feb 2024 **National Science & Technology Scholarship (Full-ride scholarship)**