

JIHYEON LEE

NLP Research Scientist | Master's degree, KAIST AI

✉ gina.ai@kakaobrain.com ☎ +82-10-9190-3894 🔗 <https://ginajihyeonlee.github.io>

SUMMARY

I am an NLP researcher at Kakao Brain. I received an Master at KAIST Graduate School of AI, under the advisement of Jaegul Choo. Before that, I obtained a Bachelor in Biz&Tech Management and Industrial Engineering at KAIST.

My recent efforts center on enhancing the coding and reasoning capabilities of LLMs and employing LLMs as tool agents. In the past, I have contributed to projects addressing large-scale foundation model training, in-context learning optimization, fair and robust language generation, multimodal question-answering, and unbiased representation learning. In these areas, I have meaningful research experiences that have been presented at top AI/ML venues (EACL, ICCV, NeurIPS, WACV).

EDUCATION

Master in Artificial Intelligence

Mar. 2020 – Feb. 2022

[DAVIAN Lab, KAIST](#)

- Keywords: NLP, Unbiased Representation Learning, Domain Adaptation
- Advisor: Prof. Jaegul Choo

Bachelor in Biz&Tech Management & double major in Industrial Engineering

Mar. 2014 – Feb. 2020

[KAIST \(Cum Laude\)](#)

- Exchange Student, TU Darmstadt, Hessen, Germany (Mar. 2017 – Feb. 2018)

PUBLICATIONS

P1 Exploiting the Potential of Seq2Seq Models as Robust Few-Shot Learners

Jihyeon Lee*, Dain Kim*, Dooha Jung*, Boseop Kim, Kyoung-woon On

ArXiv preprinted (ICLR 2024 UnderReview)

P2 Personalized Post-editing via User-generated Post-edits

Jihyeon Lee*, Yunwon Tae*, Taehee Kim*, Cheonbok Park, Jaegul Choo

European Chapter of the Association for Computational Linguistics ([EACL](#)), 2023, [Accepted as Findings](#)

P3 Dense but Efficient VideoQA for Intricate Compositional Reasoning

Jihyeon Lee*, Wooyoung Kang*, Eunsol Kim

Winter Conference on Applications of Computer Vision ([WACV](#)), 2023, [Accepted](#)

P4 Learning Debaised Representation via Disentangled Feature Augmentation

Eeungyeop Kim, Jungsoo Lee, Juyoung Lee, Jihyeon Lee, Jaegul Choo

Conference on Neural Information Processing Systems ([NeurIPS](#)), 2021, [Accepted as Oral Presentation \(< 1% acceptance rate\)](#)

P5 Removing Dataset Bias with Bias-Tailored Swapping Augmentation

Eeungyeop Kim*, Jihyeon Lee*, Jaegul Choo

Proc. of the IEEE international conference on computer vision ([ICCV](#)), 2021, [Accepted](#)

WORK EXPERIENCES

Project Leader

Jun. 2023 – Present

[Kakao Brain](#)

- Led the External Alignment Project, focusing on the exploration of LLM's emerging capabilities in code, reasoning, and agent functionalities. As the leader, established the project's direction and allocated tasks

W1-1 Enhancing the coding capability of LLMs

- Collected code pretraining data using BigQuery and explored optimal filtering methods with PySpark.
- Developed a code evaluation framework encompassing multiple benchmarks
- Curated diverse datasets for supervised fine-tuning, resulting in enhanced coding performance
- Implemented RLXF methodologies to align the model with human preferences
- Explored business directions for LLMs as tool (e.g., external APIs) learning agents

AI Research Scientist

Jul. 2021 – Present

Kakao Brain

W1-2 Large-scale language model training

- Managed collection, filtering, and preprocessing of large-scale text datasets for language modeling
- Executed distributed training for large-scale language models
- Enhanced model performance through ablative studies concerning datasets, hyperparameters, model architectures, and training objectives

W1-3 Exploiting the in-context learning ability of seq2seq models

- Significantly improved the few-shot learning capabilities of encoder-decoder models (P1)
- Developed a unified evaluation framework to assess the in-context learning ability of seq2seq models across diverse tasks

W1-4 Fair and robust language generation

- Addressed toxicity and fairness issues in LLM using controllable language generation techniques

W1-5 Compositional reasoning on multimodal question answering

- Conducted research on VideoQA problems that require intricate temporal reasoning (P3)

Data Analyst & Marketer

Jan. 2019 – Dec. 2019

CLASSUM

W2-1 Growth and data analysis

- Identified key raw data and KPI for the product and collected this data in the database
- Tracked user events on the web/app using analytics tools (GA, Optimizely, AppsFlyer)
- Significantly increased the active user and retention rate through the AARRR funnel

W2-2 Digital marketing

- Planned and developed code for marketing automation and increased user acquisition by over 300%
- Managed Facebook and Google Ads campaigns based on quantitative analysis

W2-3 Landing page development

- Managing the entire development process of the landing page, from initial planning and design to implementation

Research Assistant

Mar. 2020 – Jun. 2021

DAVIAN Lab, KAIST

W3-1 Research projects on domain adaptation

- Collaborated with Naver Papago to conduct research on personalized machine translation, utilizing real-world user data (P2)

W3-2 Research projects on unbiased representation learning

- Addressed the issue of shortcut bias in image classification, which has been causing social concerns (P4,5)

HONORS & AWARDS

Dean's List, KAIST

Fall, 2016

National Science & Engineering Scholarship

Mar. 2014 – Feb. 2018

TALKS & TEACHING EXPERIENCES

TechTalk on Developing Large Language Models In-House Singapore Management University (SMU)	Nov. 2023
Mentoring on Kaggle competitions in the field of Natural Language Processing KOSME (Korea Small & Medium Enterprises)	Dec. 2020 – Feb. 2021
Tutorial on Sentiment Analysis SK Telecom	Sep. 2020
Tutorial on Image Captioning Samsung DS	Aug. 2020

SKILLS & PERSONAL INFORMATION

Languages

- Korean (native), English (fluent; TOEIC 980, OPIC AL)

Programming & Framework

- Python, PyTorch, HuggingFace, Pandas, PySpark, SQL, LaTeX

Personality

- Self-assured, Driven, Logical, Positive, Unbiased, Smiling