Bruce W. Lee

Email: brucelws@seas.upenn.edu
Google Scholar: scholar.google.com/citations?user=a9HZkjMAAAAJ&hl=en
Github: github.com/brucewlee
Website: brucewlee.com

Education

Exp. May 2026 Bachelor's Degree in Computer Science and Cognitive Science, **University of Pennsylvania**

Experience

Fall 2024 ~ Research Scholar, **ML Alignment & Theory Scholars** Berkeley, CA
Learning language model mechanistic interpretability (training phase).

Summer 2024 Research Intern, **IBM Research** (Trustworthy AI) Yorktown Heights, NY
Introduced conditional activation steering, a lightweight contextual language model behavior control using activation patterns. Wrote IBM's activation steering toolkit.

Summer 2023 Research Intern, **NAVER Cloud** (Hyperclova AI) South Korea
Developed instruction tuning and synthetic data pipelines. Conducted large-scale fine-tuning experiments on in-house language models for performance optimization.

2020 ~ 2021
2022 ~ 2023 Research Engineer, **LXPER** South Korea
Led educational natural language processing research at an early-stage startup. Developed lexical databases, grammatical error correction, and text classification models.

Summer 2019 Research Scholar, **Institute for Basic Science** (Quantum Sensing) South Korea
Developed low-cost replacements for temperature measurement systems connected to superconducting quantum interference devices. Won top research award.

Grants, honours & awards

2024 AI for Business Award, **University of Pennsylvania** 10,000 USD
Large-scale LLM evaluation toolkit.

2022 Minister of Science and ICT Award, **Government of South Korea** ~ 40,000 USD
Transformer-based multilingual translator.

2022 Minister of National Defense Award, **Government of South Korea** ~ 8,000 USD
Transformer-based multilingual translator.

2019 Honorable Mention at Space Settlement Contest, **NASA**
Donut-structured space station design plan.

Softwares

15+ Stars **IBM/Activation-Steering** github.com/IBM/activation-steering
General toolchain for activation steering and conditional activation steering.

100+ Stars **LFTK** github.com/brucewlee/lftk
Lightweight linguistic features extractor for general text style analysis.

100+ Stars **LingFeat** github.com/brucewlee/lingfeat
Linguistic features extractor for readability classification.

Publications

Preprints and Technical Reports

2024 Programming Refusal with Conditional Activation Steering
B. W. Lee, I. Padhi, E. Miehling, M. Nagireddy, P. Dognin, A. Dhurandhar, K. N. Ramamurthy
Under Review

2024 Language Models Show Stable Value Orientations Across Diverse Role-Plays
B. W. Lee, Y. Lee, and H. Cho
Under Review

2024 HyperCLOVA X Technical Report
HyperCLOVA X Team
Technical Report

Conferences and Workshops

ACL 2024 Language Models Don't Learn the Physical Manifestation of Language
B. W. Lee and J. Lim
Association for Computational Linguistics (ACL), 2024

NAACL 2024 Instruction Tuning with Human Curriculum
B. W. Lee, H. Cho, and K. M. Yoo
North American Chapter of the Association for Computational Linguistics (NAACL), 2024

EACL 2023 Prompt-based Learning for Text Readability Assessment
B. W. Lee and J. H. J. Lee
European Chapter of the Association for Computational Linguistics (EACL), 2023

CODI 2023 A Side-by-side Comparison of Transformers for Implicit Discourse Relation Classification
B. W. Lee, B. Yang, and J. Lim
Workshop on Computational Approaches to Discourse (CODI), 2023

BEA 2023 LFTK: Handcrafted Features in Computational Linguistics
B. W. Lee and J. H. J. Lee
Workshop on Innovative Use of NLP for Building Educational Applications (BEA), 2023

TrustNLP 2023 Linguistic Properties of Truthful Response
B. W. Lee, B. F. Arockiaraj, and H. Jin
Workshop on Trustworthy Natural Language Processing (TrustNLP), 2023

EMNLP 2021 Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features
B. W. Lee, Y. S. Jang, and J. H. J. Lee
Empirical Methods in Natural Language Processing (EMNLP), 2021

NLPTEA 2020 LXPER Index 2.0: Improving Text Readability Assessment Model for L2 English Students in Korea
B. W. Lee and J. H. J. Lee
Natural Language Processing Techniques for Educational Applications (NLPTEA), 2020

Journals

2019 A low-cost cryogenic temperature measurement system using Arduino microcontroller
W. Lee
Physics Education, Volume 55, Number 2, 2019

2019 Simplifying the vacuum bazooka
J. Lee, W. Lee, E. Shin
Physics Education, Volume 54, Number 3, 2019

Patents

2023 Text Analysis Method Using LDA Topic Modeling Technique and Text Analysis Apparatus Performing The Same
Korean Intellectual Property Office, 2023

2022 Natural Language Processing Method and Natural Language Processing Device Using Neural Network Model and Non-Neural Network Model
Korean Intellectual Property Office, 2022

Others

2022 - Varsity Athlete, **University of Pennsylvania - Men's Lightweight Rowing**

2021 - 2022 Sergeant, Aircrew, **Marine Corps**

Last updated: October 4, 2024