Bruce W. Lee

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

Education

Expected 2026 — Bachelor's in Computer Science + Master's (Pending), **University of Pennsylvania**

Experience

Current — Research Scholar, **ML Alignment & Theory Scholars** — Berkeley, CA
Training phase: Language model mechanistic interpretability fundamentals. Research phase: Working on agent interpretability.

Summer 2024 — Research Intern, **IBM Research** (Trustworthy AI) — Yorktown, NY
Developed representation-level control method for language models. Built IBM's activation engineering framework.

Summer 2023 — Research Intern, **NAVER Cloud** (Hyperclova AI) — South Korea
Built instruction tuning and synthetic data pipelines. Fine-tuned in-house language models for optimal performance.

2020∼2023 — Research Engineer, **LXPER** — South Korea
Led NLP research at an early-stage startup, building lexical databases and text classification models.

Selected Publications

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

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

Report — HyperCLOVA X Technical Report
HyperCLOVA X Team
Technical Report

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

Honors

2024 — AI at Wharton Award, **University of Pennsylvania** — 10,000 USD
Language model evaluation toolkit.

2024 — Venture Lab Innovation Fund, **University of Pennsylvania** — 1,000 USD
Language model 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.

Softwares

2024 — 25+ Stars — **IBM/Activation-Steering** — github.com/IBM/activation-steering
General toolchain for lightweight language model behavior control.

2023 — 100+ Stars — **LFTK** — github.com/brucewlee/lftk
Lightweight linguistic features extractor for general text style analysis.

2021 — 100+ Stars — **LingFeat** — github.com/brucewlee/lingfeat
Linguistic features extractor for readability classification.

Last updated: November 19, 2024