Bruce W. Lee

Google Scholar: scholar.google.com/citations?user=a9HZkjMAAAAJ&hl=eny
Github: github.com/brucewlee Website: brucewlee.com
Email: phys.w.s.lee@gmail.com

Education

Bachelor of Applied Science, Computer Science
University of Pennsylvania, Philadelphia, PA, expected May 2026

Preprints & Reports

†: core contrib.

Distillation Robustifies Unlearning
Lee, B. W.†, Foote, A.†, Infanger, A.†, Shor, L.†, Kamath, H.†, ... & Turner, A. M.

Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs
Mazeika, M., Yin, X., Tamirisa, R., Lim, J., **Lee, B. W.**, ... & Hendrycks, D.

HyperCLOVA X Technical Report
Yoo, K. M., Han, J., In, S., Jeon, H., Jeong, J., ..., **Lee, B. W.**, ... & Jung, J.

Refereed Publications

∗: equal contrib.

Programming Refusal with Conditional Activation Steering
Lee, B. W., Padhi, I., Ramamurthy, K. N., Miehling, E., ..., & Dhurandhar, A.
ICLR 2025 (Spotlight)

Language Models Don't Learn the Physical Manifestation of Language
Lee, B. W., & Lim, J.
ACL 2024

Instruction Tuning with Human Curriculum
Lee, B. W.*, Cho, H.*, & Yoo, K. M.
NAACL 2024 (Findings)

Handcrafted Features in Computational Linguistics
Lee, B. W., & Lee, J. H. J.
BEA @ ACL 2023

Linguistic Properties of Truthful Response
Lee, B. W., Arockiaraj, B. F., & Jin, H.
TrustNLP @ ACL 2023

Prompt-based Learning for Text Readability Assessment
Lee, B. W., & Lee, J.
EACL 2023 (Findings)

Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features
Lee, B. W., Jang, Y. S., & Lee, J. H. J.
EMNLP 2021

Improving Text Readability Assessment Model for L2 English Students in Korea
Lee, B. W. & Lee, J. H. J.
NLP-TEA @ AACL 2020

A Low-cost Cryogenic Temperature Measurement System using Arduino Microcontroller
Lee, W. S.
Physics Education, 55(2)

Simplifying the Vacuum Bazooka
Lee, J., **Lee, W. S**., & Shin, E.
Physics Education, 54(3)

Experience

ML Alignment & Theory Scholars Berkeley, CA
Research Scholar Jun 2025 – Present

- **Mentor(s):** Tomek Korbak (UK AI Security Institute)
- Studying AI self-incrimination strategies

ML Alignment & Theory Scholars Berkeley, CA
Research Scholar Jan 2025 – Jun 2025

- **Mentor(s):** Alex Cloud & Alex Turner (Google DeepMind)
- Studied a special case of robust unlearning that erases mechanistic traces of supposedly unlearned information
- Developed experiment codebase, including pretraining, unlearning, and distillation PyTorch scripts for custom Gemma models

Center for AI Safety San Francisco, CA
Research Collaborator Sep 2024 – Jan 2025

- **Mentor(s):** Mantas Mazeika
- Developed preference elicitation methods that aim to quantify value representations in LLMs
- Wrote asynchronous Python evaluation scripts to assess value coherence and adversarial risk

IBM Research (Trustworthy AI) Yorktown Heights, NY
Research Intern May 2024 – Aug 2024

- **Mentor(s):** Inkit Padhi & Karthikeyan N. Ramamurthy
- Proposed Conditional Activation Steering as a safety technique allows a programmatic intervention on LLM behaviors
- Implemented IBM's first activation steering library, now integrated and used by other IBM papers

NAVER Cloud (Hyperclova AI) South Korea
Research Intern May 2023 – Aug 2023

- **Mentor(s):** Kang Min Yoo
- Proposed Curriculum Instruction Tuning that structures training data by cognitive complexity
- Helped implement synthetic data generation and instruction tuning pipeline for a proprietary LLM

LXPER South Korea
Research Engineer Apr 2020 – Apr 2023

- Led NLP research at an EdTech startup, architecting production-ready BERT variants for lexical analysis, grammatical error correction, and readability assessment
- Set up AWS-based serverless infrastructures to produce APIs, facilitating the complete lifecycle from research to production rollout

Center for Axion and Precision Physics Research / IBS South Korea
Research Scholar May 2019 – Aug 2019

- **Mentor(s):** Andrei Matlashov
- One of two high school students selected for a prestigious summer physics research program for undergraduate/graduate-level students
- Designed a low-cost Arduino-based cryogenic temperature measurement system, which shows a reasonable accuracy for superconducting quantum interference device (SQUID) experiments

Grants

Career Development and Transition Funding
Open Philanthropy, 2025

Gutmann-Doyle Research Opportunities Fund
University of Pennsylvania, 2025

Khan Family AI for Business Award
University of Pennsylvania, 2024
For an open-source LLM evaluation software, founded a non-profit org

Minister of Science and ICT Award
Government of South Korea, 2022
Top 10 submission out of 5420 at a Nationwide Startup Competition
For a transformer-based translator software that allows you to choose writing style

Minister of National Defense Award
Government of South Korea, 2022
Top 1 submission out of 953 at a MoND Startup Competition
For a transformer-based translator software that outperformed Google Translate for narrow technical/military use cases

Notable Softwares

IBM/Activation-Steering, **80+★**, 90% Contribution
A popular implementation of activation steering
github.com/IBM/activation-steering

LFTK, **100+★**, 100% Contribution
A multilingual, refactorized version of LingFeat. Cited and used internationally
github.com/brucewlee/lftk

LingFeat, **100+★**, 100% Contribution
A Python library that calculates 255 linguistic features from a text
github.com/brucewlee/lingfeat

Last updated: July 2025.