Kevin Lu

lu.kev@northeastern.edu linkedin.com/in/kevinlu4588 kevinlu4588.github.io

EDUCATION

Northeastern University Expected April 2026
B.S. in Computer Science & Mathematics, Honors Program GPA: 3.96/4.00

RESEARCH EXPERIENCE

Bau Lab, Interpretable Neural Networks — Northeastern University June 2024 — Present
Undergraduate Researcher

- **Concept Erasure in Diffusion Models (NeurIPS 2025):** Developed framework distinguishing mechanisms in diffusion model unlearning, and designed probing methods to detect residual knowledge (i.e. latent classifier guidance).
- Evaluated 7 erasure methods under 9 probing strategies and showed that concepts appearing erased under standard prompts often persist through alternative pathways, revealing limits of current unlearning methods.

- **Layer-Wise Analysis of Protein Language Models (NEMI Workshop 2025):** Localized binding-relevant information in ESM-2 using difference-of-means representation analysis, linear classifiers, and causal attention-head ablations.
- Found that binding signal peaks in intermediate layers and degrades toward the output; identified late-layer attention heads whose zero-shot ablation improves binding prediction.

- **Mechanistic Interpretability in AI Protein Folding:** Investigating how ESMFold computes secondary structure predictions through gradient attribution and activation patching of intermediate representations. (In Progress)

Takeda Pharmaceuticals June 2025 — Present
Machine Learning Research Intern

- **Distributed Language Model Training**: Finetuned ESM-2 protein language model for masked language modeling over antibody sequences on 8xA100 node, improving CDR region reconstruction accuracy by 32%
- **Sparse Autoencoders for Antibody Optimization**: Trained sparse autoencoders to disentangle biologically interpretable antibody features and applied feature steering during ESM inference, generating over 10,000 high-affinity candidates; 192 variants experimentally validated with 20% increased binding specificity. (Manuscript In Preparation)

Neural Systems Group — Harvard Medical School June 2023 — May 2024
Undergraduate Researcher

- **Signal Processing for Biomedical ML (RISE Expo 2024):** Implemented Kalman filtering algorithms to extract features from ECG waveforms and trained SVR models achieving Grade B AAMI blood pressure prediction.

PUBLICATIONS & PRESENTATIONS

Published Papers

- **Kevin Lu**, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, & Niv Cohen (2025). *When Are Concepts Erased From Diffusion Models?* In Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).
- **Kevin Lu**, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, & Niv Cohen (2025) *Where Do Erased Concepts Go?* In CVPR 2025 2nd Workshop on Visual Concepts.

Manuscripts

- Kevin Lu, Shipra Malhotra. *Sparse Autoencoder Feature Steering Enables Targeted Antibody Optimization.* (In preparation).

Additional Presentations

- **New England Mechanistic Interpretability (NEMI) Workshop** August 2025
  Kevin Lu, David Bau *To Bind or Not to Bind: A Layer-Wise Dissection of Binding Information in ESM*

- **3rd CVPR Workshop on Generative Models for Computer Vision** June 2025
  K. Lu, N. Kriplani, R. Gandikota, M. Pham, D. Bau, C. Hegde, N. Cohen *Where Do Erased Concepts Go?*

- **Northeastern University Research Expo (RISE)** April 2024
  Kevin Lu, Ye Yang, Quan Zhang *Predicting Blood Pressure Using AI Models for Physiological Signals*

INDUSTRY EXPERIENCE

Babel Street July 2024 — December 2024
Machine Learning Engineer Intern

- **High-Throughput NLP Pipeline**: Designed multithreaded data streaming service that filters, embeds, and indexes over 1 TB of multilingual news data per week into an Elasticsearch vector database for downstream analysis.
- **Agentic Annotation System**: Developed a multi-agent LLM data annotation system using LangGraph tooling and human-in-the-loop feedback, automating data labeling workflows and reducing annotation costs by 25%.

PROJECTS

Latent Space Exploration for Concept Erasure in Diffusion Models February 2025 — May 2025
Course Project, MATH 7223: Riemannian Optimization

- **Latent Perturbation Optimization**: Investigated recovery of erased concepts in diffusion models via latent-space optimization; used Taylor-series expansion to diagnose gradient plateaus that cause naive optimization to stall.
- **Geometry-Aware Concept Recovery**: Applied Riemannian pullback metrics to identify latent directions of maximal model sensitivity via Jacobian SVD, constraining optimization to semantically meaningful subspaces and improving concept recovery from 7% to 43%.

ConcussionMute June 2024 – December 2024

- **Transformer-Based Drum Separation**: Trained a 700M-parameter transformer model to separate percussive components from music using negative matrix factorization and fourier transform features.
- **Percussive Noise Suppression:** Reduced harsh percussive sounds by 4 dB while preserving melody and harmony, aimed at helping listeners with post-concussion sound sensitivity.

LEADERSHIP & SERVICE

Reviewer, ICLR Main Conference Oct 2025 - Nov 2025
Invited and served as reviewer for ICLR main track submission on geometric deep learning.
Reviewer, NeurIPS Mechanistic Interpretability Workshop Aug 2025
President, Northeastern Veritas Forum Chapter Sep 2023 – Present
Lead annual forums engaging 150+ students on philosophy, ethics, and faith in everyday college life.
Community Lead, InterVarsity Christian Fellowship Sep 2023 – Present
Teaching Aide, Dr. Martin Luther King Jr. K-8 School Jan 2023 – Apr 2023
Assisted in K-8 STEM program teaching math, reading, and writing skills in 90-minute weekly classes.
Mandarin Elder Service Intern, Action for Boston Community Development Oct 2022 – Jan 2023
Assisted Mandarin-speaking elders in nursing homes and community centers with learning basic technology skills.

AWARDS

Honors Global Support Fund *$6000 grant to conduct research on software infrastructure in Taiwan* December 2025
Khoury Travel Award (2X) *$1000 Travel grant for undergraduate conference presentations* July 2025, October 2025
Veritas Forum Grant (3X) *$1,250 grant for speaker travel, venue costs, and marketing.* September 2023, 2024, 2025
John Martinson Honors Scholarship *Annual $40,000 merit scholarship* September 2022 — May 2026

SKILLS

Programming Languages: Python, Java, C++, MATLAB, Bash, JavaScript
Tools & Frameworks: PyTorch, Pandas, Transformers, LangChain, Docker, BioPython, Accelerate, Scikit, FastAPI
Data & Cloud: HuggingFace, S3, RabbitMQ, ElasticSearch, DynamoDB, Amazon Web Services (AWS)