Kevin Lu

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

EDUCATION

Northeastern University Expected May 2025
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:** Introduced evaluation framework for diffusion models using latent trajectory perturbations and image-space injections. (Published at NeurIPS 2025, CVPR 2025 Visual Concepts Workshop)
- **Mechanistic Interpretability in AI Protein Folding:** Performed gradient attribution and patching interventions to isolate components of ESMFold that relate to hairpin structures in proteins. (In Progress)
- **Binding Prediction in Protein Language Models:** Optimized TCR-peptide binding prediction through linear probing of language model layers and ablation of late layer attention heads. (Presented at NEMI Workshop 2025)

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

- **Signal Processing for Biomedical ML:** Implemented Kalman filtering algorithms to extract features from ECG and PPG signals and trained SVR regressors for blood pressure prediction (Presented at Northeastern RISE 2024 Expo).

PUBLICATIONS

Published Papers

- Kevin Lu, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hedge, & 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 Hedge, & Niv Cohen (2025) *Where Do Erased Concepts Go?* In CVPR 2025 Workshop on Visual Concepts

Manuscripts

- Kevin Lu, Shipra Malhotra. (in preparation). *Sparse Autoencoder Feature Steering for Antibody Sequence Generation and Optimization.*

PRESENTATIONS & CONFERENCES

NeurIPS 2025 (Conference on Neural Information Processing Systems) December 2025
K. Lu, N. Kriplani, R. Gandikota, M. Pham, D. Bau, C. Hegde, N. Cohen *When Are Concepts Erased From Diffusion Models?*

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*

CVPR 2025 (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?*

CVPR 2025 (Second Workshop on Visual Concepts) June 2025
K. Lu., N. Kriplani, R. Gandikota, M. Pham, D. Bau, C. Hegde, N. Cohen *Where Do Erased Concepts Go?*

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

INDUSTRY EXPERIENCE

Takeda Pharmaceuticals July 2025 — Present
Machine Learning Research Co-op

- **Distributed Language Model Training**: Finetuned ESM-2 protein language model for masked language modeling inference on 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% improved binding efficacy. Manuscript in preparation.

Babel Street July 2024 — December 2024
Machine Learning Engineer Co-op

- **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 annotation framework using the LangGraph architecture, 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**: Optimized perturbation vectors to recover erased concepts in diffusion models, using Taylor-series loss landscape analysis to characterize gradient plateaus
- **Geometry-Aware Concept Recovery**: Applied Riemannian pullback metrics to model local diffusion manifolds and extract semantically meaningful latent directions via Jacobian SVD, improving erased-concept recovery from 7% to 43%.

ConcussionMute June 2024 — October 2024

- **Signal Processing for Music**: Trained a Transformer model to isolate and suppress percussive components in audio signals, reducing high-frequency drum noise for users with sound sensitivity.

LEADERSHIP & SERVICE

Reviewer, ICLR (Geometric Deep Learning)	Nov 2025
Reviewer, NeurIPS Mechanistic Interpretability Workshop (Linear Probing Papers)	Aug 2025
President, Northeastern Veritas Forum Chapter	Sep 2023 – Present
Community Lead, InterVarsity Christian Fellowship	Sep 2023 – Present
Mandarin Teaching Aide, Dr. Martin Luther King Jr. K–8 School	Jan 2023 – Apr 2023
Mandarin Elder Service Intern, Action for Boston Community Development	Oct 2022 – Jan 2023

AWARDS

Khoury Travel Award (2X) *$2000 Travel grant for undergraduate conference presenations*	July 2025, October 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, TensorFlow, OpenCV, LangChain, Docker, MLFlow, Accelerate, Scikit, FastAPI
Data & Cloud: S3, RabbitMQ, ElasticSearch, DynamoDB, Amazon Web Services (AWS), Microsoft Azure