alexloftus2004@gmail.com alex-loftus.com
Alexander Loftus
alex-loftus loftusa google scholar

AI researcher & communicator with 7+ years of experience in deep learning & machine learning. Research in code interpretability, attribution, and evaluation for Large Language Models. Transitioning to industry to apply research depth to real-world AI systems; seeking role where technical depth, community-building, and teaching ability can combine. Chat with this CV!

Career highlights:

Textbook Authorship: Authored a 524-page technical book on statistical network ML (Cambridge University Press, November 2025)

Community-building: Organized the New England Mechanistic Interpretability (NEMI) conference; Linear Algebra YouTube lecture series creator; delivered 10+ invited talks to 20–300 attendees; taught hundreds of students through meetups, summer camps, and tutorials.

Competitions: Part of a 4-person team that won 1st place in the $1,000,000 Vesuvius Kaggle competition. Won $100k Kaggle Ink Detection Progress Prize. (1,249 teams); featured on the cover of *Scientific American*.

Research: Subliminal learning work featured in YouTube video with 1m+ subscribers; best poster award for a NeurIPS workshop paper; publications in top conferences; meeting lead for Harvard AI Safety technical fellowship

Strategic Advisory: Head of Growth at EleutherAI; CBAI mentor for Harvard/MIT students; advisor for cybersecurity/mechanistic interpretability startup.

AI Infrastructure: First author on ICLR paper on scaling up AI systems for interpretability; Scaled up an AI pipeline for computational neuroscience.

Education

Northeastern University — **Boston, MA**
PhD Student, Computer Science — AI/ML | *Advisor*: Dr. David Bau — 2024–Present
Research: mechanistic interpretability, data attribution, evaluation of large language models.

Johns Hopkins University — **Baltimore, MD**
MSE, Biomedical Engineering | ML & Data Science Focus | *Advisor*: Dr. Joshua Vogelstein — 2020–2022
GPA 3.97/4.0, highest honors. Thesis: Hands-On Network Machine Learning.

Western Washington University — **Bellingham, WA**
BS, Behavioral Neuroscience | Minors: Chemistry, Philosophy — 2014–2018
Founded Computational Neuroscience Club; taught weekly seminars.

Experience

Data Scientist — **San Diego, CA**
Creyon Bio — 2023–2024
Large Protein Models For Splice-Site Prediction: Explored splice site prediction for LLMs trained on protein sequences. Pre-training, fine-tuning, and benchmarking+evals.
ML for Toxicity Prediction: Developed a novel contrastive learning pipeline to predict drug toxicity from 3-D electrostatic maps of molecules; increased classification AUC from 0.73 to 0.88.
Neuron Toxicity Detection: Developed scalable neuron segmentation and toxicology classification pipeline.

Machine Learning Research Engineer — **Rockville, MD**
Blue Halo — 2022–2023
Conditional Image Generation with Generative Adversarial Networks: Built diffusion-model synthetic data generator.
Detecting Objects with Enhanced Yolo and Knowledge Graphs: Led knowledge graph effort for object detection project. Delivered live demos to program officers.
Geometric Multi-Resolution Analysis: Led infra for document clustering & analysis method.

Research Software Engineer — **Baltimore, MD**
NeuroData Lab, Johns Hopkins University | Dr. Joshua Vogelstein — 2018–2020
MRI-to-Graphs: Optimized a diffusion MRI pipeline with docker and AWS Batch. Halved runtime and cut cloud costs by 40%.
Graspologic: Worked on an open-source graph statistics library. Later adopted by Microsoft Research for large-scale network analysis.

Assistant Director — **Seattle, WA**
iD Tech Camps | University of Washington — 2014–2018 summers
Leader and Manager: Managed 10+ instructors/week and 300+ students.
Curriculum Designer: Authored game development curriculum deployed to 50+ locations, impacting 10k+ students nationwide.

Skills Summary

Languages: Python (advanced), Bash (intermediate), R, JavaScript, SQL

Tools & Frameworks: PyTorch, NumPy, scikit-learn, nnsight, inspect-evals, lm-evaluation-harness, pandas, Polars, matplotlib, seaborn, Weights & Biases, PyTorch Lightning, vLLM, Docker, AWS, Google Cloud (GCP), Photoshop, Linux, Cursor, Codex CLI, Claude Code

Areas of Expertise: LLMs for code, interpretability, transformers, GPUs and CUDA, linear algebra, probability & statistics, deep learning, information theory, diffusion models, convolutional autoencoders, natural language processing, computer vision

Soft Skills: Public speaking, technical writing, leadership, mentorship, community-building

Textbook

Hands-On Network Machine Learning with Python: *Eric Bridgeford,* ***Alexander R. Loftus****, Joshua Vogelstein.* Cambridge University Press. Printed November 2025.

Statistics + spectral representation theory on networks. 524 pages, 147 figures.

Selected Publications

* indicates equal contribution.
🏆 indicates best poster.

Token Entanglement in Subliminal Learning: *A. Zur, Z. Ying,* ***A.R. Loftus****, et al.* NeurIPS mechanistic interpretability workshop 2025.

Investigation on token entanglement in LLMs. Featured in Welch Labs video on YouTube.

NNsight and NDIF: Democratizing Access to Open-Weight Foundation Model Internals: ***A.R. Loftus****, J.Fiotto-Kaufman*, et al.* ICLR 2025.

Open source fabric for probing & manipulating LLM weights without engineering overhead.

🏆 **A Saliency-based Clustering Framework for Identifying Aberrant Predictions**: *A. Tersol Montserrat,* ***A.R. Loftus****, Y. Daihes.* Paper, **NeurIPS** LatinX AI Workshop, 2023.

Detects spurious feature reliance via saliency embeddings.

A low-resource reliable pipeline to democratize multi-modal connectome estimation and analysis: *J. Chung, R. Lawrence,* ***A.R. Loftus****, et al.* Paper, Under review, 2024.

Transforms diffusion MRI scans into graphs; open-sourced (code).

Leadership & Community Engagement

Head of Growth — **EleutherAI**, 2025
Developing funding strategy for EleutherAI's 2025-2026 philanthropic effort. Paid position.

Technical Lead — **Harvard AI Safety Team**, 2026
Paid facilitator for the Spring 2026 Introductory Technical AI Safety Fellowship at Harvard

Conference Organizer — **NEMI**, 2025
Organized 200+ person interpretability conference; Raised $17,000 grant funding.

Research Mentor — **CBAI**, 2025
Mentoring Harvard/MIT students in Summer 2025

Strategic Advisor — **Krnel.ai**, 2025
Advisor to cybersecurity-focused startup specializing in interpretability tooling for AI systems.

Meetup Speaker — **SDML**, 2023–2024
Speaker & organizer for San Diego AI Meetups.

Hackathon Organizer — **NeuroData Workshop**, 2019
Helped organize hackathon & workshop to explore statistics for high-dimensional testing.

Talks & Demos

White-Box Techniques for Code LLMs: Influence Benchmarking, the Attendome, and Variable State Debugging: *Lawrence Livermore National Laboratory, 2025*

Invited talk on interpretability for code LLMs.

A Shared Infrastructure for Interpretability: *FAR AI Tech. Innovations for AI Policy Conf., 2025*

Invited demo for DC policymakers; showcased live editing of GPT2 internals

State of the Art in Knowledge Editing: *A.R. Loftus*, 2024

Survey talk on LLM knowledge-editing methods.

1st Place Solution — Vesuvius Ink Competition: *R. Chesler,* ***A.R. Loftus****, A. Tersol Montserrat, T. Kyi*, 2023
Walkthrough of winning $100,000 ink-detection model.

ICML Conference Highlights: ***A.R. Loftus***, 2023
Selected breakthroughs from ICML. Presented to biotech execs and SDML meetup group.

Working with LLMs: AI San Diego Conference, 2023.
Invited talk: Introduction to LLM engineering. 300+ attendees

Linear Algebra, from Dot Products to Neural Networks: ***A.R. Loftus***, 2023.
Created a YouTube tutorial series on the fundamentals of linear algebra for machine learning.

Fellowships & Awards

Harvard AI Safety Technical Fellowship 2025
Harvard fellowship for technical work in AI safety.

GCP Research Grant 2025
$5,000 grant for computational research.

Khoury Distinguished Fellowship 2024
Northeastern University PhD fellowship.

First Place Winner 2023
Kaggle Vesuvius Competition, $100,000.

Best Poster Award 2023
NeurIPS 2023 LatinX AI Workshop.

AWS Research Grant 2019
$10,000 grant for computational research on cloud services.

Teaching

Head Teaching Assistant — **Johns Hopkins University**, Spring 2021
Foundations of Computational Biology and Bioinformatics, *EN.BME.410/634*

Teaching Assistant — **Johns Hopkins University**, Spring 2020
NeuroData Design II, *EN.BME.438/638*

Teaching Assistant — **Johns Hopkins University**, Fall 2019
NeuroData Design I, *EN.BME.437/637*

Teaching Assistant — **Western Washington University**, Winter 2017
Introduction to Behavioral Neuroscience, *PSY.220*

Curriculum Designer — **iD Tech Camps**, Spring 2017
Built curriculum used across 50 locations in the United States by tens of thousands of students.

Fun

Gaming: Starcraft 2 grandmaster, local tournament winner; WoW 10-man server first ToGC (off-tank)

Music: Fingerstyle guitarist; performed at open mic nights.

Dancing: Partner dance instructor and competition winner (Fusion, West Coast Swing, Zouk)