

Alexander Loftus

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; meeting lead for [Harvard AI Safety technical fellowship](#)

Strategic Advisory: Led a red-teaming sprint for OpenAI; 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

PhD Student, Computer Science — AI/ML | *Advisor:* [Dr. David Bau](#)

Research: mechanistic interpretability, data attribution, evaluation of large language models.

Boston, MA

2024–Present

Johns Hopkins University

MSE, Biomedical Engineering | ML & Data Science Focus | *Advisor:* [Dr. Joshua Vogelstein](#)

GPA 3.97/4.0, highest honors. Thesis: [Hands-On Network Machine Learning](#).

Baltimore, MD

2020–2022

Western Washington University

BS, Behavioral Neuroscience | Minors: Chemistry, Philosophy

Founded Computational Neuroscience Club; taught weekly seminars.

Bellingham, WA

2014–2018

EXPERIENCE

Red-Teaming Lead

OpenAI (Contract)

Agents of Chaos: Viral paper led to contract work designing a project and then leading a team of 10 in an internal red-teaming sprint for OpenAI.

Remote

March 2026

Data Scientist

Creon Bio

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.

San Diego, CA

2023–2024

Machine Learning Research Engineer

Blue Halo

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.

Rockville, MD

2022–2023

Research Software Engineer

NeuroData Lab, Johns Hopkins University | [Dr. Joshua Vogelstein](#)

Baltimore, MD

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

iD Tech Camps | University of Washington

Leader and Manager: Managed 10+ instructors/week and 300+ students.

Curriculum Designer: Authored game development curriculum deployed to 50+ locations, impacting 10k+ students nationwide.

Seattle, WA

2014–2018 summers

SKILLS SUMMARY

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

Tools & Frameworks: PyTorch, NumPy, scikit-learn, transformers, 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, LaTeX

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.

Agents of Chaos: *N. Shapiro et al.* preprint, 2026.

Viral red-teaming investigation of LLMs in a social Discord server. Covered by IEEE Spectrum, Cisco, Kaspersky, and Constellation Research within days of publication. Led to leadership in an OpenAI red-teaming sprint.

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

Developing funding strategy for EleutherAI's 2025-2026 philanthropic effort. Paid position.

EleutherAI

2025

Technical Lead

Paid facilitator for the Spring 2026 Introductory Technical AI Safety Fellowship at Harvard

Harvard AI Safety Team

2026

Conference Organizer

Organized 200+ person interpretability conference; Raised \$17,000 grant funding.

NEMI

2025

Research Mentor

Mentoring Harvard/MIT students in Summer 2025

CBAI

2025

Strategic Advisor

Advisor to cybersecurity-focused startup specializing in interpretability tooling for AI systems.

Krnel.ai

2025

Meetup Speaker

Speaker & organizer for San Diego AI Meetups.

SDML

2023–2024

Hackathon Organizer

Helped organize hackathon & workshop to explore statistics for high-dimensional testing.

NeuroData Workshop

2019

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

Harvard fellowship for technical work in AI safety.

2025

GCP Research Grant

\$5,000 grant for computational research.

2025

Khoury Distinguished Fellowship

Northeastern University PhD fellowship.

2024

First Place Winner

Kaggle Vesuvius Competition, \$100,000.

2023

Best Poster Award

NeurIPS 2023 LatinX AI Workshop.

2023

AWS Research Grant

\$10,000 grant for computational research on cloud services.

2019

TEACHING

Head Teaching Assistant

Foundations of Computational Biology and Bioinformatics, *EN.BME.410/634*

Johns Hopkins University

Spring 2021

Teaching Assistant

NeuroData Design II, *EN.BME.438/638*

Johns Hopkins University

Spring 2020

Teaching Assistant

NeuroData Design I, *EN.BME.437/637*

Johns Hopkins University

Fall 2019

Teaching Assistant

Introduction to Behavioral Neuroscience, *PSY.220*

Western Washington University

Winter 2017

Curriculum Designer

Built curriculum used across 50 locations in the United States by tens of thousands of students.

iD Tech Camps

Spring 2017

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)