

Alexander Loftus

I strive to develop new machine learning and algorithmic techniques that help humans understand, evaluate, and control LLMs in more meaningful, powerful, and trustworthy ways. In the past, I have worked on representation learning, scalable systems, volumetric segmentation, and network statistics. Some highlights:

Textbook author: Publishing contract with Cambridge University Press.

1st place ranking, \$100,000 Machine Learning competition: [Work featured on cover of Scientific American](#). Competed against 1249 teams. Vesuvius scroll ink detection.

Publications in top conferences: Best poster award at NeurIPS 2023 LatinX workshop, first author work in ICLR 2024

Open-source contributions: Primary contributor to microsoft network statistics package graspologic.

Teaching and Leadership: Led a team of three to develop an object detection augmentation algorithm; a team of five to contribute to a brain network estimation pipeline; assistant director managing 8-12 instructors.

EDUCATION

Northeastern University

Boston, MA

PhD Computer Science

2024-

Advisor: [Dr. David Bau](#)

Interpretability, evaluations, and training dynamics in Code LLMs.

Johns Hopkins University

Baltimore, MD

MSE Biomedical Engineering: Machine Learning & Data Science Focus

2020-2022

Advisor: [Dr. Joshua Vogelstein](#)

Thesis: [Hands-On Network Machine Learning](#)

dean's list, highest honors, GPA 4.0/4.0.

Western Washington University

Bellingham, WA

BS Behavioral Neuroscience — *Minors:* Chemistry, Philosophy

2014-2018

Founder & President, Computational Neuroscience Club

Vice President, Neuroscience Club

Built computational neuroscience club from scratch, taught weekly seminars.

TEXTBOOK

[Hands-on Network Machine Learning](#): *Eric Bridgeford, Alexander R. Loftus, Joshua Vogelstein*. Cambridge University Press, in copy-editing phase. To be printed November 2025.

Spectral representation theory on networks. 530 pages, 147 figures.

PUBLICATIONS

* indicates equal contribution.

[NNSight and NDIF: Democratizing Access to Open-Weight Foundation Model Internals](#): *A.R. Loftus**, *J. Fiotto-Kaufman**, et al. ICLR 2025.

Infrastructure project to easily explore and manipulate foundation model internals with no 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. **Won best poster.**

Use embeddings of saliency map crops to identify predictions caused by spurious features.

[A low-resource reliable pipeline to democratize multi-modal connectome estimation and analysis](#): *J. Chung, R. Lawrence, A.R. Loftus*, et al. Paper, in review at Nature Methods, 2024

Turn diffusion MRI scans into adjacency matrices. [Code](#) on github.

[Role of CAMKII in Associative Conditioning and GLR-1 Expression in C. Elegans](#): *M. Pribic, A.R. Loftus*, et al. Poster, Society for Neuroscience, 2017.

Removing a protein involved in learning blocks associative conditioning in worms.

TALKS

State of the Art in Knowledge Editing: *A.R. Loftus*, 2023

Current techniques in knowledge localization and editing in LLMs and diffusion models.

1st Place Solution - Vesuvius Ink Competition: *R. Chesler, A.R. Loftus, A. Tersol Montserrat, T. Kyi*, 2023

Presenting on our winning solution to a \$100,000 Kaggle competition, part of the \$1,000,000 Vesuvius competition.

ICML Conference Highlights: *A.R. Loftus*, 2023

Machine learning techniques in drug discovery and medicine at ICML 2023.

Working with LLMs: *A.R. Loftus*, 2023.

Introduction to LLM engineering. Talk given to 100 people at the AI/ML San Diego meetup.

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.

Effects of an *unc-43* (CaMKII) Gene Deletion on Short-Term Memory for Associative Conditioning in *C. elegans*: *A.R. Loftus*, Psychfest 2017.

Mechanistic understanding of worm neural circuitry.

FELLOWSHIPS & AWARDS

First Place Winner

Kaggle Vesuvius Competition, \$100,000.

2023

Khoury Distinguished Fellowship

Northeastern University PhD fellowship.

2024

Best Poster Award

NeurIPS 2023 LatinX AI Workshop.

2023

MIT EECS GAAP

MIT mentorship program.

2023

Harvard AI Safety Technical Fellowship

Harvard fellowship.

2025

AWS Research Grant

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

2019

EXPERIENCE

Data Scientist

Creyon Bio

San Diego, CA

2022-2024

ESP Embeddings: Developed contrastive feature representation learning approach for electrostatic potential data. Helped classify toxic drugs to save on drug testing costs. Resulted in classification accuracy improvement from an AUC of 78 to 89.

Neuron Toxicity Detection: Built deconvolution and segmentation pipeline to detect toxicity in neurons.

Machine Learning Research Engineer

Blue Halo

Rockville, MD

2021-2022

Conditional Image Generation with Generative Adversarial Networks: Synthetic data augmentation. Led the switch to diffusion-based methods over GANs.

Detecting Objects with Enhanced Yolo and Knowledge Graphs: Used knowledge graphs to enhance object detection on videos.

Geometric Multi-Resolution Analysis: Infrastructure for a hierarchical clustering method.

Research Engineer

Johns Hopkins University — Dr. Joshua Vogelstein

Baltimore, MD

2018-2021

Graspologic: Built dimensionality reduction models on networks.

ndmg: Diffusion MRI to graphs pipeline. AWS cloud-computing integration with pytest CI/CD infrastructure. Eliminated 1000 lines of code and halved computation time.

Assistant Director

iD Tech Camps — University of Washington

Seattle, WA

2014-2018 summers

Leader and Manager: Administrator for a STEM education camp which taught C++, Python, Java, game design, and robotics at the University of Washington. Managed 8-12 instructors with 80-120 students per week.

SERVICE & ADVISORY

Consultant Advisory role for cybersecurity-based interpretability startup.	Krnel.ai Spring 2025
Organizer Organized the 200-person New England Mechanistic Interpretability conference. Obtained \$17,000 grant.	NEMI Summer 2025

TEACHING

Head Teaching Assistant Foundations of Computational Biology and Bioinformatics, <i>EN.BME.410/634</i>	Johns Hopkins University Spring 2021
Teaching Assistant NeuroData Design II , <i>EN.BME.438/638</i>	Johns Hopkins University Spring 2020
Teaching Assistant NeuroData Design I , <i>EN.BME.437/637</i>	Johns Hopkins University Fall 2019
Teaching Assistant Introduction to Behavioral Neuroscience, <i>PSY.220</i>	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
Instructor Taught programming and game design to high school students.	iD Tech Camps 2014-2018 summers

SKILLS SUMMARY

Languages: Python, R, Rust, Bash, CSS, Mojo, English, Broken Spanish

Tools & Frameworks: pytorch, pytorch-lightning, tensorflow, jax, numpy, scipy, pandas, polars, sklearn, seaborn, matplotlib, docker, AWS, google cloud (GCP), photoshop, SQL, weights & biases, mlflow, kubernetes, linux

Areas of Expertise: Linear algebra, probability & statistics, deep learning, information theory, transformers, diffusion models, convolutional autoencoders, GPUs and CUDA, public speaking, leadership & management, teaching, natural language processing, computer vision

FUN

Gaming: Starcraft 2 grandmaster in high school, competed and won in Seattle-area tournaments.

Music: Fingerstyle guitarist. Played at open mic nights.

Dancing: Partner dance instructor and competition winner. Fusion, West Coast Swing, Zouk, Salsa, Bachatta.