

Aniket Ghosh

ghosh.anik@northeastern.edu • Boston, MA • [Portfolio](#) • [LinkedIn](#) • [GitHub](#)

Education

M.S. in Artificial Intelligence, ML Concentration – Northeastern University, Khoury College – GPA: 4.0/4.0 Aug 2025 – May 2027
Coursework: CS 7180 Actionable Interpretable Methods (PhD seminar, Prof. Byron Wallace) • CS 5180 Reinforcement Learning • CS 5100 Foundations of AI • CS 5800 Algorithms • IE 7374 Machine Learning Operations (MLOps).

B.Tech in CS & Business Systems – Institute of Engineering & Management, Kolkata – GPA: 4.0/4.0 Jul 2020 – Jun 2024
Ranked **2nd of 180+** • **Director's Award** • *Key Courses:* Neural Networks, NLP, Data Structures, OS.

Research, Publications & Patents

Function Vectors as a Model-Diffing Tool: 17 Heads Repair a Bad Fine-Tune *Mech Interp* • [LessWrong](#) 2026 • [GitHub](#)

- Patched function vectors **across models** rather than contexts: transplanting **17 attention heads** took a misaligned medical LoRA from **3 to 27 of 36** safe responses (clean: 30), **ARC flat at 100%**, beating crosscoder diffing (1), few-shot (23), steering (10).
- Established a **repair/injection asymmetry**: the intervention cures misalignment but cannot induce it, collapsing the metric to **-0.2 logits** under injection where a matched random direction holds **+19.9**. Replicated on **Llama-3.1-8B**.

Linear Probes Tell You Where Quantization Will Hurt

LLM Compression • [LessWrong](#) 2026 • [GitHub](#)

- Trained per-layer linear probes (NER, POS, chunking) to localize token-level signal, then allocated precision by layer: held **99–100% accuracy at a 5-bit average** where uniform quantization collapsed to **16–41%** on unseen datasets.

Intelligent Traffic Sign Detection Using YOLOv8 with Hybrid CNN Filtering

First Author

AISC 2024 • Springer Proceedings in Information and Communication Technologies (in press, Oct 2026)

Lightweight Hybrid DNN-GNN for Network Intrusion Detection with Adaptive Late Fusion

Co-author

Published – [IEEE Xplore](#) • **Patent** Filed by SRM Institute of Science and Technology

Technical Skills

Interpretability: nnsight, NDIF, PyTorch hooks, DAS / Boundless DAS / MDAS, causal patching, linear probes, crosscoders, SAEs

Training: PyTorch, Hugging Face Transformers, accelerate, distributed multi-GPU training, SLURM, LoRA/QLoRA, quantization

GenAI & Agents: RAG, LangChain, LangGraph, CrewAI, OpenAI Agents SDK, MCP, ChromaDB, Claude Code, Cursor

MLOps: Docker, Kubernetes, AWS, GCP, CI/CD, MLflow, FastAPI, Terraform

Core Domains: Mechanistic Interpretability, AI Safety, Reinforcement Learning, CV, NLP, Multi-Agent Systems, Medical Imaging

Languages: Python, C/C++, SQL, Bash

Experience

AI/ML Engineer – *Varosync, Inc., New York, NY* (remote from Boston)

May 2026 – Present

- **Sole ML engineer** on an early-stage drug-discovery platform, building the stack from scratch and reporting directly to the CEO/CTO; architected a **biomedical knowledge graph spanning millions of molecules**.
- **Trained 7 to 8B parameter molecular embedding models** on an **800M-molecule** corpus across a **multi-GPU cluster**, with complete control over the training loop (data curation, distributed training, checkpointing, and evaluation) on a hybrid cloud stack.
- Owned everything downstream of the gold labels: featurization, **vector indexing**, retrieval logic, and the live user query path.
- Applied **mechanistic interpretability** techniques to refine model architecture, improving **inference efficiency** and performance.

Teaching Co-Lead, ML Lab – *AI & Data Ethics Program (AIDE), Northeastern Ethics Institute*

May 2026 – Aug 2026

- Designed and taught a **12-session ML lab** for a residential cohort of philosophy PhD students in this **Sloan Foundation-funded** program on fairness, differential privacy, transformers, epistemic simulations and mechanistic interpretability.
- Taught **sparse autoencoders, the logit lens, and linear probing** hands-on using **nnsight** and **NDIF**.

AI Researcher – *CMATER Lab, Jadavpur University, Kolkata*

Jun 2023 – May 2025

- Built an unsupervised pipeline for histopathological cell segmentation that reached **85%+ accuracy** with **30% fewer false positives** than a Mask R-CNN baseline, enabling annotation-free pathology screening.
- Designed the preprocessing stage: **color quantization** plus **DBSCAN** to separate nuclei that intensity thresholding merges together.
- Worked two years under Dr. Kaushiki Roy and [Prof. Debotosh Bhattacharjee](#); mentored juniors in deep learning for medical imaging.

Computer Vision Research Intern – *North-Eastern Hill University* (Remote)

Feb 2023 – Jul 2023

- Ran a **50+ paper** literature review on sign language recognition to define annotation and quality standards for a novel **Indian Sign Language** benchmark dataset, adapting practices from Microsoft ASL and MNIST to underrepresented language contexts.

Selected Projects

Sufficient Cause Disambiguation

Mechanistic Interpretability • CS 7180 (PhD seminar) • [GitHub](#)

- Proved **probing accuracy and causal relevance are distinct properties** on Llama-3-8B: separator tokens hit **1.000 LDA accuracy across all 32 layers** yet drive a **0% flip rate** under causal patching.
- Built the pipeline (LDA, gradient attribution, causal patching): **448× compression at 99.6% accuracy, 100% flip rate** at $\alpha=2$.
- Audited **LLM resume screening for demographic bias**: name-swapped resumes triggered a **95.8% flip rate** and accounted for **62%** of the model's highest-uncertainty Fit decisions on otherwise identical content.

Biomedical Knowledge Graph Link Prediction

Healthcare AI • CS 5100 • [GitHub](#)

- Engineered graph features (PageRank, structural metrics) on the **BioRED** biomedical corpus for a Random Forest link predictor at **0.94 ROC-AUC**, with **5-hop** path reasoning for relationship explainability.
- Ran a classical ML vs. **TransE** embedding benchmark (**0.94 vs. 0.61 ROC-AUC**), showing hand-crafted graph features dominate learned representations on low-resource biomedical corpora, a practical finding for data-scarce settings.

Autonomous Multi-Agent Trading Simulation

[GitHub](#)

- Designed and deployed a multi-agent trading simulation on the **OpenAI Agents SDK: 4 autonomous AI traders** coordinated across **6 MCP servers** with **44 distinct tools**, consuming live market data via **Polygon.io** for autonomous portfolio management.

Achievements & Service

BlueDot Impact Technical AI Safety — certified July 2026 • **Graduate TA, NLP**, Northeastern Khoury — Ran PyTorch and deep learning labs for **40+** graduate students • **Volunteer Educator, The Hope Foundation** — 7 years teaching math and CS to underprivileged students, Kolkata • **Graduate Student Senator**, Northeastern GSG.