Arthur Pogosian

arthurmpogosian@gmail.com | LinkedIn | Github

Education

University of Pennsylvania — Philadelphia, PA

Masters of Science in Computer Science (GPA 4.00) — *Dec 2025*

Bachelor of Arts in Mathematics, Computer Science — *May 2025*

Experience

Axiom Math — Jan 2026

Research Engineer Intern — *San Francisco, CA*

- Working with ex-FAIR/MSL researchers to create autoformalization systems that push the SOTA in math.

Cambridge AI Safety Hub — Dec 2025 – Feb 2026

Research Fellow — *Remote*

- Selected for MARS 4.0, researching with Peter Hase on the phenomenon of backtracking in reasoning models.

Pinterest — May 2025 – Aug 2025

Machine Learning Engineer Intern — *San Francisco, CA*

- Designed and implemented novel transformer-based user multi-embedding models for rich multi-interest signals.
- Improved recall@10 by 2x from single embedding baseline and 25% over previous SOTA PinnerSage baseline.
- Researched novel methods for compressing super-long user history sequences for more accurate personalization.
- Deployed model in production and integrated in Homefeed for 500 million MAU, led ablation studies on features.
- Collaborated to finetune LLMs for recommender systems to predict user interest on low activity users.

UPenn Brachio Lab — July 2024 – Oct 2025

Graduate Machine Learning Research Assistant — *Philadelphia, PA*

- Working with Eric Wong at the Brachio Lab on feature extraction benchmarks in LLMs and transformers.
- Researching jailbreaking in LLM systems using theoretical bases, extending the LogicBreaks paper.

Affirm — May 2024 – Aug 2024

Software Engineer Intern — *New York, NY*

- Developed scoped user sessions to help prevent millions in fraud, improving security and checkout retention.
- Built new features to extend login session infrastructure and support the Apple Pay integration at Affirm.

UPenn Distributed Systems Lab — Jan 2024 – June 2024

Research Assistant — *Philadelphia, PA*

- Worked on Fast Succinct Non-Interactive Zero-Knowledge Regex Proofs. Publication under Sebastian Angel.

Reunion LLC — Mar 2022 – May 2024

Founding Engineer — *Remote*

- Successfully launched pixel-art MMORPG alongside two cofounders with over 1000 players and 100k revenue.

Projects

Trajectory Steering Vectors | *PyTorch, TransformerLens*

- Investigated process-level steering method, revealing model biology and achieving 54% jailbreak success.

Mitigating Adversarial Prompt Perturbations with COT in LLMS | *PyTorch, Jupyter Notebook*

- Developed gradient-free adaptive CoT defense restoring 97% baseline accuracy on Nova Pro across 48k prompts.

Restricted GeoCLIP | *PyTorch, Jupyter Notebook, Matplotlib, Pandas, Numpy*

- Built novel dataset of **117,000** UK street images, finetuned CLIP with **27%** accuracy for the Top 50 UK cities.

GoogLite | *Java*

- Full search engine completely from scratch; Web server, distributed KVS, Spark, Crawler, Indexer, PageRank.

Technical Skills

Languages: Python, Java, Kotlin, Rust, C, C++, SQL
Technologies: TensorFlow, PyTorch, Pandas, Matplotlib, PyTest, NumPy, Scikit-learn, Git, Docker
Concepts: Machine Learning, Deep Learning, NLP, Bayesian Optimization, Recommendation Systems, LLMs