

ARMIN GERAMI

@ agerami@umd.edu ☎ (667)-200-9253 📍 United States in www.linkedin.com/in/armingerami/

EXPERIENCE

High-Frequency Trading ML Research Intern

Summit Securities Group

📅 Aug 2026 – Nov 2026 📍 Illinois, United States

- Feature extraction from high-frequency market data for signal generation.
- Training Gradient Boosting and Transformer models to predict short-term price movements.
- Developing backtesting simulations and trading strategies.

AI/ML Engineer Intern

Apple

📅 Feb 2026 – Aug 2026 📍 California, United States

- Worked on Developing an agentic multimodal RAG-LLM for the RF Systems team as an internal tool.
- Automated evaluation using Human-In-The-Loop reinforcement learning.
- Configured vector database indexing and retrieval pipelines across Milvus and S3.
- Integrated REST and MCP endpoints to deliver data pipeline.

Research Assistant

University of Maryland, CS Department

📅 June 2023 – present 📍 Maryland, United States

- Computational efficiency of **Transformers: Subquadratic Attention, KV Reduction**, Quantization.
- Applications of **Transformers** and **LLMs**.
- Python, C++, CUDA, Deep Learning, High Performance Computing, Differentiable Programming

Software Developer

University of Maryland

📅 Jan 2022 – June 2023 📍 Maryland, United States

- Designed and developed a server-based tool to assist the state of Maryland with monitoring their buildings.
- Python, Full-Stack, Data Visualization, Data Mining

SKILLS

Languages:

C++ Python CUDA

Quantitative & ML Domains:

Deep Learning Machine Learning Signal Processing
Transformer Architecture High Performance Computing
Differentiable Programming Gradient Boosting
Linear Regression

Libraries & Frameworks:

PyTorch JAX NumPy SciPy Scikit-Learn vLLM Ollama
HuggingFace

Infrastructure & Tools:

Milvus AWS (S3) Linux Git SQL

HONORS & AWARDS

- NSF NeuroPAC Fellowship Award (2025)
- Outstanding Graduate Research Assistant Award (2024)
- Ranked 21st in Iran's National University Entrance Exam (2016, among 250,000 students).
- Qualified for national Math and Informatics Olympiad (2014, 2015).

EDUCATION

PhD in Computer Science

University of Maryland, College Park

📅 2023 - 2027 (expected) 📍 United States

Major: Computer Science

Focus: Transformers, Machine Learning, HPC **GPA:** 3.7

MSc in Electrical Engineering

University of Maryland, College Park

📅 2022 - 2023 📍 United States

Major: Telecommunications

Focus: Signal Processing, Communication Systems **GPA:** 3.8

BSc in Electrical Engineering

Sharif University of Technology

📅 2016 - 2020 📍 Tehran, Iran

Focus: Signal Processing, Communication Systems **GPA:** 3.5

SELECT PUBLICATIONS

- Mixture-of-Gaussians Attention For Sublinear Long-Context Decoding
Transformer, vLLM, Sparse Attention, CUDA, Python
ICLR 2026 (Submitted)
- Two-Level Generative Evolutionary Prompt Adaptation for Non-Parametric Continual Learning
Continual Learning, Agentic Reasoning, Python
Preprint
- On The Application of Linear Attention in Multimodal Transformers
Transformer, Linear Attention, Multimodal, CUDA, Python
CVPR 2026
- Auditing Algorithmic Bias in Transformer-Based Trading
Transformer, Multimodal, Information Theory, Python
Neurips 2025
- Graph Edge-Coloring Utilization for Accelerating Sparse Matrix Vector Multiplication
High Performance Computing, Hardware Design, Verilog, C++
ASPLOS 2024
- Efficient Spatial Audio Rendering Via Differentiable FIR To IIR Estimation
Signal Processing, Spatial Audio, Differentiable Programming, C++
ICASSP 2025

INVENTION DISCLOSURES

- Differentiable FIR To IIR Filter Estimation
- Rapid Energy and Emission Auditor