ARMIN GERAMI

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

SKILLS

Coding:

Python (20k+ lines) | C++ (20k+ lines) | CUDA (5k lines) | Verilog (2k lines)

Expertise:

Training and Implementing **LLMs** | **Transformer** Architecture | Deep Learning | **Machine Learning** | **Linear Regression** | Optimization | High Performance Computing | Differentiable Programming | Algorithms & Data Structures | Calculus | Probability Theory | **Signal Processing** | **Statistics** | Information Theory

ML Libraries:

Pytorch | **JAX** | SciPy | Numpy | Scikit | Huggingface

EXPERIENCE

Research Assistant

University of Maryland, CS Department

June 2023 – present Maryland, United States

- Applications of **Transformers** and **LLMs**.
- Computational efficiency of **Transformers**: **Linear Attention**, KV Caching, Pruning.
- Partial Information Decomposition in **Transformers** and **RAG**.
- Spatial Audio Rendering.
- 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

High Performance Computing, Intern

Iran's National Telecommunication Research Center

Summer 2019 Tehran, Iran

- High Performance Computing, Verilog

PERSONAL PROJECTS

- Deployed a Python model to identify high-probability calendar call spread options by quantifying favorable volatility conditions, including elevated IV relative to RV, sufficient liquidity, and a steep or inverted term structure.

INVENTION DISCLOSURES

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

VOLUNTARY

- Peer reviewed 6 papers; Neurips, ICLR

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, HPC, LLMs, Spatial Audio **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

FIRST AUTHOR PUBLICATIONS

- On The Application of Linear Attention in Multimodal Transformers
  Transformer, Multimodal, CUDA, Python
  Preprint
- Transformer Based Linear Attention with Optimized GPU Kernel Implementation
  Transformer, High Performance Computing, CUDA, Python
  TMLR 2025 (Submitted)
- Room Impulse Response Synthesis via Differentiable Feedback Delay Networks
  Signal Processing, Spatial Audio, Differentiable Programming, Python
  ICASSP 2026 (Submitted)
- Auditing Algorithmic Bias in Transformer-Based Trading
  Transformer, Multimodal, Information Theory, Python
  Neurips 2025
- Quantifying Document Impact in RAG-LLMs
  Transformer, Information Theory, LLM, Python
  TMLR 2025 (Submitted)
- Efficient Spatial Audio Rendering Via Differentiable FIR To IIR Estimation
  Signal Processing, Spatial Audio, Differentiable Programming, C++
  ICASSP 2025
- Graph Edge-Coloring Utilization for Accelerating Sparse Matrix Vector Multiplication
  High Performance Computing, Hardware Design, Verilog, C++
  ASPLOS 2024