Alexander Kranias

U.S. Citizen | 727-457-4433 | alexander.kranias@gatech.edu
alexkranias.com | linkedin.com/in/alexanderkranias | github.com/alexkranias

Education

Georgia Institute of Technology | Atlanta, GA *August 2022 – Present*
Bachelor of Science in Computer Science, GPA 4.00 Expected Graduation, May 2026

Concentrations: Intelligence (AI/ML), Systems and Architecture
Organizations: GT Create-X Incubator Program, GT Start Up Exchange, Georgia Tech Varsity Rowing, GT Student Tour Guides
Coursework: Data Structures**,** Algorithms**,** Computer Architecture, Systems and Networks, Intro to AI, Robotics, Linear Algebra

Skills

Languages: Python, C, C++, CUDA, Java, React, HTML/CSS, JavaScript, Node.js, MATLAB
Tools/Frameworks: Docker, CMake, Supabase, AWS, ROS2, Jupyter Notebook, Kubernetes, Flask**,** PostgreSQL, SQL, git, Simulink

Experience

Embedded Software Engineer Intern *May 2024 – Present*
Vermeer Corporation Pella, Iowa

- Developed a C++ simulation and ROS2 package of autonomous pile driver ad-hoc networks and cloud environments in Ubuntu.
- Architected a MQTT/DDS Hybrid Ad-hoc Network implemented as ROS2 service and client nodes to scale up to 50+ machines.
- Created custom cost map quadtree (reduced protobuf file size by 96%) and Jupyter Notebook to benchmark serialization formats.
- Incorporated a VIN-data timestamp table distributed across client nodes that minimizes HTTPS cloud-to-machine data transfer.
- Designed UML diagrams in Lucid; followed Agile with Jira; managed git and code reviews for 100+ commits/issues/pull requests.

Software Engineer Intern *May 2023 – July 2023*
Raymond James Financial St. Petersburg, Florida

- Created a full-stack LLM copilot web app for financial advisors and investment bankers that deploys fine-tuned GPT-3 models.
- Developed HTML/CSS frontend and our backend with Azure OpenAI API, Outlook API, and internal CRM software.
- Designed project requirements; interviewed 20+ stakeholders; onboarded team to continue project; created roadmap with COO.

Research

Computer Architecture Research Assistant *August 2023 – Present*
Georgia Tech HPArch Lab (High Performance Architecture Lab) Atlanta, Georgia

- Optimizing matrix multiplication (MatMul) CUDA kernels using Nsight Systems to achieve cuBLAS-like performance.
- Developed materials for GPU Hardware/Software course; designed Tiled MatMul project in CUDA; ran on HPC cluster.
- Formulated last level cache replacement algorithm to mitigate RowHammer attacks by deprioritizing cache blocks from hot rows.
- Built cache block memory access trackers in a remote Linux environment integrated into open-source ChampSim repo (C++).
- Benchmarked cache block accesses and DRAM hot-row utilization for 40+ traces; resulted in new IPCs of 80%-200% of initial.
- Developed memory paging and multithreaded round-robin process scheduling projects in a Linux docker environment.

Machine Learning Research Assistant *October 2022 – October 2023*
DuckAI Research Group Atlanta, Georgia

- Built DuckTrack: Accurate Computer Activity Tracking (tool to build multimodal computer agent datasets, 100+ downloads).
- Created ARB: Advanced Reasoning Benchmark for Large Language Models (accepted to MATH-AI Workshop at NeurIPS ’23).

Extracurriculars

Co-Founder, Sideline (RippleX Finalist) **|** PyTorch, Flask, LangChain, React, Pinecone, PosgreSQL *July 2023 – Present*

- AI video search for sports; demoed to 17 colleges: 7 D1, 3 D2, 5 D3, and 2 NAIA; Led outreach; met with NBA VP of Strategy.
- Developed multimodal vector search of gameplay using LangChain, Pinecone VectorDB, PostgreSQL, and Flask CRUD requests.
- Designed vector embeddings to enable player and video embeddings search using action-timestamp and biography annotations.
- Created a Two-Stream CNN-TSM (Temporal Shift Module) architecture in PyTorch for real-time foul recognition in game footage.
- Built custom foul-labeling GUI in Python to efficiently annotate over 44,000 televised NBA game clips (97 hours of footage).

Projects and Awards

3dReal (Stanford TreeHacks 2024 Winner) | Swift, CUDA, Python, Firebase, instant-ngp *February 2024 - Present*

- Creating custom CUDA library for live 3D video reconstruction by enabling and optimizing real-time fine-tuning of model weights.
- Developed in 48 hours an iOS app that captures social NeRFs using NVIDIA instant-ngp, Firebase, Swift, and CUDA.
- Synchronously generates Neural Radiance Fields (NeRFs) using the cameras of multiple iOS devices, creating a shared 3D selfie.