Darius Kianersi

703-457-0042 | dariuskianersi@gmail.com | linkedin.com/in/dariuskianersi | github.com/dariuskia

Education

University of Maryland — College Park, MD
Bachelor of Science (BS) in Computer Science and Mathematics; GPA: 3.90/4.0 — *Expected Winter 2025*

Thomas Jefferson High School for Science and Technology — Alexandria, VA
High School Diploma; GPA: 3.97/4.0 — *May 2022*

Experience

Cartesia — Sep. 2024 – May 2025
Research Engineer — *San Francisco, CA*

- Trained alternative text-to-speech (TTS) model architectures with hybrid SSM-transformer backbones.
- Implemented a novel autoregressive voice cloning model to improve speaker similarity by 23%.
- Optimized memory-efficient CUDA graph capture, improving effective batch size by 16x.
- Led synthetic data efforts – distributed data processing on k8s, data-annealing finetuning method with warmup-stable-decay (WSD) scheduler to improve word error rate (WER) by 31%.

NVIDIA — May 2024 – Aug. 2024
Software Engineer Intern - Deep Learning Compilers Team — *Santa Clara, CA*

- Finetuned Llama-3.1 70B and implemented inference-time search methods for XLA compiler test generation, improving C++ code coverage by 32%.
- Built distributed PyTorch training loop with JIT compilation and pipeline/tensor parallelism, improving Model FLOPs utilization (MFU) by 18%.
- Wrote custom GEMM and FlashAttention kernels in Triton with reduced memory footprint and 82% of CuBLAS performance.

Microsoft — May 2023 – Aug. 2023
Research Intern — *Redmond, WA*

- Finetuned Large Language Models (LLMs) like DeBERTa using Low-Rank Adaptation (LoRA), improving retrieval accuracy@3 over tf-idf by 8%.
- Implemented Retrieval Augmented Generation (RAG) in PyTorch and Azure ML for downstream QA.
- Leveraged chain-of-thought and evolutionary (Evol-Instruct) prompting for a synthetic data generation pipeline.

GAMMA Research Lab — Aug. 2023 – Present
Autonomous Driving Researcher — *College Park, MD*

- Leveraged VQ-VAE and Vision Transformer models for autoregressive video generation as a world model.
- Spearheading a novel meta-learning method to enhance sample efficiency in autonomous steering.

Capital One — Jan. 2023 – Apr. 2023
Software Engineer Intern — *College Park, MD*

- Constructed embedding space of merchant accounts with Graph Representation Learning in node2vec.
- Benchmarked similarity searches such as Faiss and ScaNN to compute nearest neighbors on transactional data.
- Deployed machine learning models to Apache Spark enabling fraud detection at scale.

Projects

Real-Time Phishing Detection | github.com/dariuskia/shascam

- Inferenced Mixtral-8x7B with chain-of-thought (CoT) for real-time phone call scam detection over Twilio.
- Deployed a Flask backend with a minimal React Native mobile app for low-latency push notifications.

Impact Investing Platform | github.com/dariuskia/alignly

- AI-powered impact investing platform with OpenAI Whisper transcriptions on FastAPI backend.
- Retrieved S&P500 companies from MongoDB Atlas vector database and executed market orders on Alpaca API.

Technical Skills

Languages: Python, C++, JavaScript, TypeScript, Java, Bash, HTML/CSS, C, Go, MySQL
Frameworks/Libraries: PyTorch, JAX, React [Native], Node.js, Flask, pandas, NumPy
Developer Tools: Git, Azure, Bazel, Jenkins, AWS, CUDA