Hung Tran

+1 (813) 204-0037 | hunglongtran2004@gmail.com | github.com/hunglongtrangithub | linkedin.com/in/hunglongtran

Education

University of South Florida Tampa, FL

Bachelor of Science in Computer Science *Expected May 2026*

- Major GPA: 4.0/4.0, Cumulative GPA: 3.95/4.0
- Relevant Coursework: Computer Architecture, Operating Systems, CUDA Programming, Data Storage & Analysis with Hadoop; Natural Language Processing, Computer Vision, Software Engineering; Statistics, Linear Algebra, Discrete Mathematics, Analysis of Algorithms, Data Structures, Automata Theory & Formal Languages

Technical Skills

- **Platforms**: NVIDIA DGX, Amazon Web Services
- **Languages**: Rust, Python, Golang, TypeScript, SQL (Postgres), HTML/CSS, C/C++, CUDA
- **ML Libraries & Frameworks**: PyTorch, HuggingFace, vLLM, scikit-learn, polars, NumPy, SciPy, Marimo
- **Full-Stack Technologies**: Next.js, React, FastAPI, PostgreSQL, MongoDB, OpenSearch, AWS X-Ray, AWS CDK

Work Experience

Amazon Sunnyvale, CA

Software Development Engineer Intern *May 2025 — August 2025*

- Contributed to Roger - an AI-powered security camera chatbot, implementing **2 critical end-to-end features** (event notification and event monitoring systems) **within 4 weeks** to meet executive demo deadline with Ring director and founder
- Drove prototype success by shipping **80+ pull requests** in 3 months to secure product investment from leadership
- Developed camera event monitoring system, displaying event timeline of **2000+ events**, video playback, and VLM-generated (vision language model) captions for AI response verification and debugging
- Implemented LLM-powered notification system to detect event anomalies, enabling users to browse **1000+ notifications** by camera, date, and notification type
- Authored high-level design document for Roger with **12 observability features**, defining team's development roadmap
- Built contextual user feedback system for **200+ beta users** with **AWS OpenSearch** integration for analytics & improvement
- Developed Roger trace observability feature with **AWS X-Ray** for real-time AI response debugging through Roger web UI
- Proactively built internal Roger chat CLI, adopted by **20+ engineers** and **accelerated local testing & development by 2x**

Moffitt Cancer Center Tampa, FL

Machine Learning Research Intern *October 2022 — May 2025*

- Developed multimodal RAG oncology chatbot using Llama 3 8B, Whisper V3, XTTS-V2, and SadTalker on **NVIDIA DGX**, reducing SadTalker inference time by **83%** through **CUDA memory optimization**, **model quantization**, and **data parallelism**; presented at USF AI+X Symposium
- Architected Flask ETL pipeline on **AWS ECS**, implementing custom inference to process clinical records with fine-tuned clinical NLP models and create NLP medical annotations in Electronic Medical Record Search Engine (EMERSE)
- Benchmarked Llama 2 7B, BioGPT, and GPT-Neo using **PubMedQA** and **MedQA** datasets; applied **LoRA fine-tuning** on Llama 2 7B using proprietary medical data; built **Gradio** demo platform for integrated clinical NLP tasks
- Fine-tuned ModernBERT on **2,054 patient records**, achieving **83% accuracy** for bone marrow transplant survival prediction with Polars-based data processing and feature engineering

Projects

Core Contributor, MuopDB (github.com/hicder/muopdb) March 2025 — Present

- Core contributor to MuopDB - open-source **Rust vector database** designed for multi-user AI memory systems, supporting HNSW, IVF, and SPANN indexing algorithms with on-disk storage via memory mapping
- Implemented HTTP metrics endpoint using **Prometheus** for MuopDB server observability, enabling performance monitoring and system health tracking
- Built the async WAL write groups for MuopDB, achieving **5x throughput improvement** while ensuring data integrity and consistency