Tuan Quang

Piscataway, NJ | anhtuanquang2016@gmail.com | +1-(818)-747-9964 | GitHub | LinkedIn | US Citizen

Senior AI Engineer with 7+ years building and optimizing production AI systems in financial services, including LLM inference, multi-agent workflows, vector search, guardrails, and evaluation on AWS and GCP. Proven record of cutting inference cost and latency through quantization and model optimization while leading engineers through architecture and design reviews.

SKILLS

- **Programming:** Python, Java, Go, C#
- **ML & LLM Frameworks:** PyTorch, TensorFlow, Hugging Face Transformers, LoRA/PEFT, LangGraph, LangChain
- **LLM Inference & Optimization:** LLM Inference, FP8/INT4 Quantization, Model Compression, Latency/Throughput/Cost Optimization, Resource Right-Sizing
- **Agentic AI:** Multi-Agent Systems, Agentic RAG, Function Calling, Memory & Context Management, Amazon Bedrock Guardrails, Human-in-the-Loop
- **Vector Search & Retrieval:** Similarity Search, Pinecone, Weaviate, FAISS, Neo4j (GraphRAG)
- **Evaluation & Observability:** Model Evaluation, RAGAS, LangSmith, A/B Testing, Hallucination Rate, Grafana, CloudWatch
- **Cloud & MLOps:** AWS (Bedrock, SageMaker, ECS), GCP, Kubernetes, Docker, Terraform, CI/CD

PROFESSIONAL EXPERIENCE

AI VIET NAM | Remote, Viet Nam

Software Engineer, Applied AI **Sep 2025 – Present**

- Designed retrieval and LLM evaluation experiments comparing semantic chunking strategies on Recall@5, latency, and per-query inference cost, improving retrieval Recall@5 by 20%.
- Benchmarked FP8 and INT4 quantization for production LLM inference workloads, profiling latency, throughput, and output-quality tradeoffs, reducing inference cost by 30% while maintaining application quality.
- Applied a state-of-the-art ensemble technique to a production AI system, improving reasoning accuracy from 56.16% to 79.18%.
- Architected 10 production-grade multi-agent AI systems (RAG, vector databases, FastAPI, Docker, CI/CD, AWS) and mentored 100+ members, with 90% earning AWS Cloud Practitioner certification.
- Authored and maintained core AI engineering documentation for 100+ members, standardizing best practices across projects.

Galileo Financial Technologies (Contract via Altimetrik) | Remote, US

Software Engineer, Applied AI **Apr 2025 – Sep 2025**

- Trained and evaluated PyTorch models on AWS SageMaker, applying LoRA-based fine-tuning to financial fraud-detection workflows and improving precision by 40%.
- Deployed and orchestrated multi-model Python ML services on AWS SageMaker, ECR, and ECS with CI/CD, experiment tracking, model versioning, and monitoring, cutting operational cost by 25% and release cycles from weeks to days.
- Defined the technical roadmap and architecture from business requirements, guiding engineers to build a scalable system with reusable components and right-sized resources, cutting development time by 40% and infrastructure cost by 60%.

LPL Financial (Contract via Altimetrik) | Remote, US

Software Engineer, Applied AI **Jan 2024 – Jan 2025**

- Architected an agentic RAG platform over 500K+ financial documents with similarity search using LangGraph, LangChain, Neo4j, and Pinecone, reducing analyst research time by 60%.
- Engineered a multi-agent orchestration pipeline unifying rule-based, retrieval-augmented, and generative components with guardrails and human-in-the-loop review, reducing workflow failures by 70%.
- Built a document ingestion pipeline with AWS Glue, Amazon Textract, and PaddleOCR that converts unstructured financial documents into validated LLM-ready datasets, reducing end-to-end processing time by 50%.
- Led engineers on agent architecture, running design and code reviews and establishing standards for automated testing, documentation, and design patterns, reducing feature-delivery time by 50%.

Ford Credit (Contract via Altimetrik) | Remote, US

Software Engineer **Feb 2022 – Dec 2023**

- Built personalized recommendation models (XGBoost, PyTorch) from user-behavior data, deployed on AWS, and validated variants through A/B testing, lifting CTR by 50% over the previous system.
- Built distributed PySpark pipelines to clean, transform, and aggregate large financial datasets, improving data-processing efficiency by 50%.
- Automated AWS and GCP infrastructure and CI/CD with Terraform and Tekton, adding automated deployment validation that reduced deployment effort by 60% and sustained 99.9% availability.

AeroVironment | Simi Valley, CA

Software Engineer **Jan 2020 – Feb 2022**

- Developed and deployed a TensorFlow object detection model on embedded aircraft hardware with OpenCV preprocessing, sustaining 60 FPS real-time inference within onboard compute constraints.
- Monitored model performance with Grafana dashboards and analyzed detection failures in MATLAB, informing retraining that raised mAP@0.5 from 40% to 50%.
- Architected an autonomous UAV flight-control system in cross-platform C# (mobile and desktop) with optimized routing and real-time decision logic, reducing mission execution time by 25%.
- Translated operational requirements into scalable system designs with cross-functional teams, streamlining release workflows and reducing application deployment time by 60%.

California State University, Northridge | Los Angeles, CA

Research Assistant **Jan 2018 – May 2019**

- Deployed optimized NLP, ML, and computer vision models on NVIDIA Jetson AGX Xavier (16 GB RAM) with real-time metrics streaming to AWS CloudWatch, enabling on-device inference at 15 FPS and remote performance tracking.
- Integrated camera and sensor inputs with rule-based control logic and ML perception models into a unified decision pipeline, improving navigation accuracy by 60%.

PROJECTS

E-Mind: Slides to Interactive Lectures | Startup Project

Backend and AI Engineer **2024 – 2026**

- Architected an AI education platform on AWS orchestrating LangChain, LangGraph, Claude SDK, and OpenAI SDK across content extraction, script generation, narration, and validation, reducing script-generation time by 75%.
- Fine-tuned Hugging Face models on Vietnamese educational datasets and built a RAGAS/LangSmith evaluation pipeline tracking faithfulness, relevancy, context precision/recall, latency, and token cost, improving generation quality by 20% (BLEU/ROUGE, human eval) and reducing hallucinations by 15%.

EDUCATION

California State University, Northridge **2015 – 2019**

B.S. Computer Science/Engineering

Cum Laude | GPA: 3.67 | Outstanding Graduating Senior Award | University Scholarship