TUAN QUANG

New York, NY | anhtuanquang2016@gmail.com | (818) 747-9964 | https://tuanquang.com | US Citizen

SUMMARY

Senior AI Engineer with 7+ years of software engineering experience across production LLM systems, agentic RAG platforms, multi-agent pipelines, and cloud-native MLOps infrastructure. Architected RAG systems processing 500K+ documents, reducing research time by 70%; designed low-latency inference pipelines serving 10K+ daily financial API requests; and published peer-reviewed research in multimodal AI. Proven track record owning end-to-end system architecture, mentoring engineers, and driving technical decisions across LangGraph, LangSmith, AWS Bedrock, SageMaker, and modern orchestration tooling.

TECHNICAL SKILLS

LLM & RAG: LangGraph, LangChain, LlamaIndex, CrewAI, AutoGen, RAG, Function Calling, Guardrails, LangSmith
ML & Deep Learning: PyTorch, TensorFlow, JAX, Hugging Face, PEFT, LoRA/QLoRA, ONNX, Scikit-learn
Cloud/MLOps: AWS, SageMaker, Bedrock, ECR, ECS, Kubernetes, Docker, Terraform, MLflow, W&B, DVC
Data/Retrieval: FAISS, Pinecone, Chroma, Weaviate, Neo4j, PostgreSQL, MongoDB, Kafka, PySpark, Airflow
Languages: Python (Proficient), Java, JavaScript/TypeScript, SQL, C++, C#

PROFESSIONAL EXPERIENCE

Galileo Financial Technologies | Remote

AI Engineer (Contract) *Apr 2025 – Sep 2025*

- Owned end-to-end MLOps architecture on AWS SageMaker, ECR, and ECS, defining deployment topology and CI/CD strategy; reduced operational costs by 25% through automated experiment tracking, model versioning, and deployment workflows with Weights & Biases and DVC.
- Built and evaluated ML services on Amazon Bedrock and SageMaker, fine-tuning foundation models on proprietary financial data and improving fraud detection precision by 40%.
- Led cross-functional collaboration among data scientists, software engineers, and business stakeholders to translate AI research into scalable production systems, accelerating model deployment cycles from weeks to days.

LPL Financial | Remote

Software Engineer – (Applied AI Engineer) *Jan 2024 – Jan 2025*

- Designed enterprise agentic RAG architecture using LangGraph, LangChain, and CrewAI with short- and long-term memory; reduced manual document research time by 70% across 200+ financial advisors.
- Mentored 2 engineers in building a financial AI chatbot using Amazon Bedrock, RAG, and fine-tuning techniques; established evaluation workflows that improved response accuracy by 40%.
- Engineered production inference pipelines on SageMaker and Bedrock for Claude 3.5, Llama 3 70B, and Gemma 3; implemented quantization, structured outputs, and observability with OpenTelemetry, Grafana, and Prometheus, reducing end-to-end latency by 25%.

AI VIET NAM | Remote, Viet Nam

AI Researcher & MLOps Instructor (Part-time) *Jan 2024 – Present*

- Led curriculum design for end-to-end MLOps program (100+ students), covering data versioning, experiment tracking, SageMaker Pipelines, RAG with LangChain, vector databases, Bedrock, FastAPI, and Docker CI/CD.
- Researched agentic RAG architectures using LangGraph, investigating multi-agent collaboration and tool-use patterns for complex reasoning in financial and technical domains.
- Evaluated LLM serving optimizations, including FP8/INT4 quantization and speculative decoding, measuring their impact on inference cost and perplexity; implemented semantic chunking strategies to improve retrieval quality.

Ford Credit | Remote

Software Engineer *Feb 2022 – Dec 2023*

- Designed high-availability credit backend microservices on GCP using Java, Spring Boot, and PostgreSQL, serving 50K+ daily transactions; modernized Angular dashboards, improving data accessibility by 75%.

- Automated multi-environment infrastructure with Terraform and Tekton CI/CD; built SQL-driven dashboards for real-time KPI visibility, achieving 99.9% measured uptime across production services.

AeroVironment | Simi Valley, CA

Software Engineer *Jan 2020 – Feb 2022*

- Architected an autonomous UAV flight control system with optimized routing and real-time decision logic, reducing mission execution time by 25%; presented system design to cross-functional teams for production approval.
- Built a cross-platform UAV control interface using C#/.NET and React, integrating low-latency telemetry streams to support real-time situational awareness and mission monitoring.

California State University, Northridge | Los Angeles, CA

Research Assistant *Jan 2018 – May 2019*

- Built a modular automation framework on Raspberry Pi using Linux and Python for autonomous robotics; engineered deployment pipelines that reduced manual configuration time by 40%.
- Optimized lightweight computer vision algorithms for edge devices with 1 to 4GB RAM, enabling real-time robot tracking and object recognition.

PROJECTS

Legal AI Pipeline | Personal Project *2024 – Present*

- Architected a multi-agent legal AI system with LangGraph and LangChain for employment law, combining BM25, FAISS, Reciprocal Rank Fusion, Neo4j, LangSmith, and Pydantic structured outputs across 100K+ legal documents.
- Designed evaluation framework using Recall@K and MRR metrics via LangSmith to benchmark retrieval quality; chose BM25+FAISS with RRF over pure vector search after testing showed 18% improvement in legal citation recall.
- Deployed AWS-based legal document ingestion pipelines using Bedrock, ECR, ECS, Lambda, FastAPI, and Airflow to automate document processing workflows.

E-Mind: Slides to Interactive Lectures | Startup Project *2024 – 2025*

- Architected an AI education platform using LlamaIndex and LLMs to transform slide decks into interactive narrated videos, improving script generation efficiency by 75%; fine-tuned Hugging Face models on Vietnamese educational data and deployed the platform on AWS.
- Led collaboration with educators, product stakeholders, and engineers to define AI-powered lecture generation requirements; translated user feedback into LlamaIndex pipeline improvements that reduced content creation time by 75% and improved learner engagement by 50%.

EDUCATION

B.S. Computer Science/Engineering – California State University, Northridge *2015 – 2019*
Cum Laude | GPA: 3.67 | Outstanding Graduating Senior Award | University Scholarship

PUBLICATIONS

Improving Generalization in Visual Reasoning via Self-Ensemble

Computer Vision Foundation / arXiv, Oct 2024 • Tien Huy Nguyen / Tuan Quang

- Co-authored training-free self-ensemble framework enhancing reasoning in Large Vision-Language Models, achieving state-of-the-art on SketchyVQA, Outside Knowledge VQA, and OOD-VQA benchmarks.

Enhancing Video Retrieval with Robust CLIP-Based Multimodal System

SOICT'23 Conference, Dec 2023 • Dung Le / Tuan Quang

- Engineered an end-to-end CLIP-based multimodal pipeline, from PyTorch fine-tuning to FastAPI backend deployment, to improve video retrieval accuracy; presented the system at SOICT'23.