TUAN QUANG

Senior AI Engineer | LLM Systems & Multi-Agent Architecture | MLOps
(818) 747-9964 | anhtuanquang2016@gmail.com | tuanquang.com | New York, NY | US Citizen

SUMMARY

Senior AI Engineer with 7+ years in software engineering and production LLM systems, multi-agent pipelines, and cloud-native MLOps infrastructure. Architected agentic RAG platforms processing 500K+ documents that cut research time by 70%, designed inference pipelines serving 10K+ daily financial API requests at sub-second latency, and published two peer-reviewed papers in multimodal AI. Proven track record of owning system architecture end-to-end, mentoring engineering teams, and driving build-vs-buy decisions across LangGraph, LangSmith, AWS Bedrock/SageMaker, and modern orchestration tooling.

EXPERIENCE

AI Engineer (Contract) | Galileo Financial Technologies – Remote *Apr 2025 – Sep 2025*

- Owned end-to-end MLOps architecture on AWS (SageMaker, ECR, ECS), defining deployment topology and CI/CD strategy for a 6-person engineering team - reduced operational cost by 25% through automated experiment tracking via Weights and Biases and DVC.
- **Built and evaluated ML services** on Amazon Bedrock and SageMaker, fine-tuning foundation models on proprietary financial data with evaluation pipelines that improved fraud detection precision by 40%.

Software Engineer – (Applied AI Engineer) | LPL Financial – Remote *Jan 2024 – Jan 2025*

- **Designed and led system architecture** for an enterprise agentic RAG platform (LangGraph, LangChain, CrewAI) with short/long-term memory, making tradeoff decisions across latency, retrieval quality, and cost that cut manual document research time by 70% across 200+ financial advisors.
- **Mentored 2 engineers** on LLM integration patterns, prompt engineering, and evaluation frameworks establishing the team's first structured code review process for AI-specific pull requests.
- **Engineered production inference pipelines** on SageMaker and Bedrock, fine-tuning Claude 3.5, Llama 3 (70B), and Gemma 3 with quantization and structured output, reduced end-to-end latency by 25%.
- **Deployed cloud-native AI microservices** on AWS (Lambda, S3, DynamoDB, ECR, ECS) with Terraform/Docker and full observability via OpenTelemetry, Grafana, and Prometheus—achieving 99.9% uptime on production inference endpoints.

AI Researcher & Educator (Part-time) | AI VIET NAM – Remote, Vietnam *Jan 2024 – Present*

- **Led curriculum design** for end-to-end MLOps program (50+ 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 optimization** (FP8/INT4 quantization, speculative decoding), measuring impact on inference cost and perplexity; implemented semantic chunking strategies to maximize retrieval quality.

Software Engineer | Ford Credit – Remote *Feb 2022 – Dec 2023*

- **Designed high-availability backend microservices** on GCP with Java, Spring Boot, and PostgreSQL serving 50K+ daily transactions; modernized Angular frontend 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.

Software Engineer | AeroVironment – Simi Valley, CA *Jan 2020 – Feb 2022*

- **Architected autonomous UAV flight control system** with optimized routing and real-time decision logic, reduced mission execution time by 25%—presenting design to cross-functional stakeholders for production approval.
- **Built cross-platform control interface** (C#/.NET backend, React/JS UI) with low-latency telemetry for real-time situational awareness.

Research Assistant | California State University, Northridge – Los Angeles, CA *Jan 2018 – May 2019*

- **Built modular automation framework** on Raspberry Pi (Linux/Python) for autonomous robotics; engineered deployment pipelines reduced manual configuration time by 40%.
- **Optimized lightweight computer vision algorithms** for 1–4GB RAM edge hardware, enabling real-time robot tracking and object recognition.

PROJECTS

Legal AI Pipeline | Personal Project *2024 – Present*

- **Architected multi-agent legal AI system** with LangGraph/LangChain for employment law: hybrid retrieval (BM25/FAISS with Reciprocal Rank Fusion) over 100K+ legal documents, Neo4j graph store for entity relationships, LangSmith observability, and Pydantic structured output.
- **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.
- **Deploying on AWS** (Bedrock, ECR, ECS, Lambda) with FastAPI and Airflow orchestration for automated document ingestion and processing pipelines.

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

- **Developed AI platform** using LlamaIndex/LLMs to transform slides into interactive narrated videos (+75% script generation efficiency); fine-tuned Hugging Face models on Vietnamese educational data and deployed on AWS.

AI Health Counselor App | Personal Project 2023 – 2024

- Built RAG-powered health counselor using LangChain with ChromaDB retrieval over curated medical Q&A datasets, structured output via Pydantic, and safety guardrails (topic boundary filtering, hallucination detection, mandatory disclaimers) to ensure responsible AI responses.
- Implemented context-aware multi-turn dialogue with function calling for symptom follow-up routing and sliding-window conversation memory, improving response coherence across extended sessions.

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 end-to-end CLIP-based multimodal pipeline from PyTorch fine-tuning to FastAPI backend for video retrieval accuracy, presented at the 12th International Symposium on Information and Communication Technology.

TECHNICAL SKILLS

Core: LangGraph, LangChain, PyTorch, AWS Bedrock/SageMaker, FAISS, Neo4j, FastAPI, Docker, Kubernetes, Terraform

LLM & AI Systems: CrewAI, AutoGen, LlamaIndex, RAG (Agentic/Corrective/Self-RAG), Prompt Engineering, Function Calling/Tool Use, Structured Output, Guardrails, LLM Evaluation, MCP Server

ML & Deep Learning: TensorFlow, JAX, Hugging Face (Transformers, PEFT, Accelerate), Scikit-learn, OpenCV, Fine-tuning (QLoRA/LoRA), Quantization (GPTQ/AWQ/GGUF), ONNX, W&B, MLflow

Data & Infrastructure: Pinecone, Chroma, Weaviate, PostgreSQL, MongoDB, DynamoDB, Kafka, PySpark, Airflow, DVC, Feast, OpenTelemetry, Grafana, Prometheus, GitHub Actions, CI/CD

Languages: Python (Proficient), Java, JavaScript/TypeScript, SQL, C++, C#

EDUCATION

B.S. Computer Science/Engineering – California State University, Northridge *2015 – 2019*

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