Luong Quang Dung

0367790670 luongquangdung00@gmail.com dungluongquang quangdungluong

Education

Ho Chi Minh City University of Technology **August 2018 – August 2022**
Bachelor of Engineering in Control Engineering and Automation *GPA: 8.06*

Experience

Success Software Service **May 2025 – Now**
AI Engineer

- Designed and deployed a Retrieval-Augmented Generation (RAG) system using FastAPI, vLLM for high-performance LLM serving, and Milvus for vector storage and semantic search, delivering precise, context-aware answers across multiple domains.
- Created data pipelines for PDF ingestion, chunking, embedding generation, and indexing with Airflow and Docker-based microservices. Built and maintained domain-specific knowledge bases with semantic search and embedding-based retrieval.
- Optimized system performance with model quantization, streaming inference, and observability tools to improve response time and reliability.
- Architected and implemented a real-time Voice Agent for healthcare and wellness, using WebRTC for low-latency, bi-directional audio streaming.
- Integrated STT, LLM-based reasoning, and TTS modules to enable natural, conversational experiences for end-users.
- Designed a scalable microservice architecture supporting conversational state tracking and contextual memory.

FPT Software **June 2022 – May 2025**
AI Engineer

- Designed, implemented, and optimized scalable real-time video analytics systems for edge devices (Jetson, Qualcomm) and servers environments.
- Utilized NVIDIA DeepStream SDK, Triton Inference Server, Redis, Kafka, FastAPI, Flask, MinIO, and Docker within a microservices architecture to ensure system scalability and efficiency.
- Collaborated with cross-functional teams to deliver high-quality solutions within tight deadlines.
- Conducted performance optimizations and model fine-tuning to improve system accuracy and efficiency.
- Built LLM-based chatbot and context-extraction systems, integrating multi-document PDF ingestion, semantic retrieval, and contextual memory for open-domain Q&A.
- Implemented an OCR system for customer bill parsing, extracting key details and exporting data to CSV, elevating the accuracy from 85% to 95% while optimizing 40% inference time through optimization, post-processing, and fine-tuning.
- Designed and constructed a system that utilizes object detection for identifying patterns in technical drawings and incorporates an OCR system for text recognition, achieving an exceptional accuracy rate of 98%.

Projects

LLMOps - End-to-End RAG Platform | *NLP, RAG, LLMOps* **May - September 2025**

- Built a production RAG pipeline for PDF ingestion → chunking → embeddings → vector retrieval with Airflow orchestration and MinIO/Postgres storage.
- Implemented LLM Gateway to route requests across multiple LLMs, supporting dynamic model selection, load balancing, and guardrails.
- Exposed retrieval and LLM services via FastAPI with centralized routing, access control, and observability dashboards, containerized with Docker.
- Improved answer quality and latency through evaluation, observability, iterative prompt/guardrail tuning and streaming inference patterns.

Review Analytics | *Pytorch, NLP* **August 2022**

- Construct an end-to-end baseline using PhoBERT architectures with various model heads.
- Use the KFolds Cross Validation training technique and out-of-fold for error analysis.
- Use various techniques to enhance performance: hyper-parameters tuning, Label Smoothing, Pseudo-labeling, Voting ensemble.

Skills

Languages: Python, C++
Framework: PyTorch, ONNX, TensorRt, DeepStream, LangChain, LangGraph, Transformers
Cloud: AWS, Azure
Database: SQLite, MySQL, PostgreSQL
Soft Skills: Critical Thinking, Problem Solving
Others: Docker, Git, Prometheus, Grafana, MLflow

Certifications

- Machine Learning DevOps Engineer Nanodegree Udacity
- Data Scientist Nanodegree Udacity
- Cloud DevOps Engineer Nanodegree Udacity
- TOEIC Score 895

Awards

Best Performer of FPT Software Quy Nhon 2024 **January 2025**
Recognized for dedication and contributions to team success

1st runner-up Vietnam MLOps Marathon **September 2023**
A competition focused on real-world business challenges, and honing AI project deployment skills

Top 1 HuggingFace Competition **April 2023**
Detecting Ships in Ports to Avoid Congestion and Manage Traffic

Top 4 Final Quy Nhon AI Hackathon **September 2022**
Placed second in Qualifying, advanced to the Finals, and finished in the top 4