Khoi Dinh Tran

Ho Chi Minh City, Vietnam • koitran.work@gmail.com • (+84) 819701764 • Linkedin • Github • Portfolio

WORK EXPERIENCE

AI Engineer — **May 2024 - Apr 2025**
Inspire Lab Technology — **Ho Chi Minh City, Vietnam**

- Architected and built from scratch the core infrastructure of an enterprise-level **Agentic AI** system for automated **SEO content generation**. The system was regularly validated by SEO professionals for quality, reducing costs and cutting content production time by **80%**.
- **Fine-tuned** various **LLMs** (Llama 3.1, Qwen 2.5, ...) resulting in **35% improvement** in content quality based on professional feedback.
- Developed pipelines for automated **image generation and editing** to complement textual content.
- Implemented **CI/CD workflows** that reduced deployment time from 3 hours to 20 minutes, enabling bi-weekly feature releases.
- Presented complex analytical findings and optimization recommendations to management in clear, actionable terms.
- **Technologies:** Python, Pydantic, LangChain, LangGraph, CrewAI, ComfyUI, LoRA, Ollama, OpenAI, FastAPI, Docker, AWS, ...

AI Engineer intern - Healthcare Data Analysis — **Mar 2024 - Apr 2024**
TMA Solutions — **TMA lab 6, Ho Chi Minh City, Vietnam**

- Proposed and developed innovative **OCR** approaches to extract critical data from prescription across hundreds of hospitals nationwide.
- **Technologies:** Python, OpenCV, Yolo, PaddlePaddle

AI Engineer trainee - AI Proctoring — **Oct 2023 - Dec 2023**
ISODS George Washington Institute of Data Science & Artificial Intelligence — **Remote (New York)**

- Developed specialized approach to detect abnormal head movements and positioning with high precision of **~83%** on synthesis data.
- Created innovative ear-related behavior detection capabilities to identify prohibited actions such as wearing earbuds or headphones, achieving high precision of **~91%** on synthesis data and **~89%** on collected real-world data.
- **Technologies:** Python, OpenCV, Yolo, 6DRepNet, Pytorch

PROJECTS

AI Assistant Platform for Japanese Retail — **Nov 2024 - Present**

- Architected and implemented a custom **multilingual chatbot system** for Japanese commercial retail industry, incorporating proprietary tools that execute contextually based on conversation scenarios, resulting in **95% reduction** in customer service response times.
- Engineered specialized **prompt systems** for multiple LLMs that improved response accuracy by **80%** across **8** distinct **NLP tasks** including intent recognition and context-aware responses.
- Developed **10 proprietary tools** that operate within the Chatbot framework, dynamically adapting to user context.
- Collaborated with cross-functional teams to integrate the solution with existing business systems.
- **Technologies:** Python, Prompt Engineering, Dify, Docker, FastAPI, PostgreSQL, Redis

ChatBot That See and Hear Your Content — **Jul 2024 - Jul 2024**

- Built an intelligent conversation system that goes beyond text-based interactions, and enhancing **LLMs/vLLMs/STT** models with **Retrieval-Augmented Generation (RAG)** techniques to create a **multimodal chatbot** capable of understanding your images and videos.
- **Technologies:** Python, RAG, LangChain, Ollama, Hugging Face, LLaVA, Whisper, Chroma, Gradio

Store AI Assistant — **Jul 2024 - Jul 2024**

- Designed and developed an intelligent customer support system for a food retailer by integrating **self-hosted LLMs (Ollama)** with **RAG**, enabling personalized product recommendations and automated responses to variations of customer inquiries.
- **Technologies:** Python, RAG, LangChain, Unstructured, Ollama, Hugging Face, Chroma, Gradio

SKILLS

AI: Deep Learning, NLP, Computer Vision, RAG, Generative AI, Agentic AI

AI Frameworks: LangChain, LangGraph, CrewAI, MCP, Transformers, PyTorch, TensorFlow, Ollama, Dify, ComfyUI, ...

Vector Databases: Chroma, Weaviate, Qdrant, ...

Programming Languages: Python, SQL, C++, C, Java

MLOps & DevOps: Docker, CI/CD, Git

Web Development: FastAPI, Streamlit, Gradio

Cloud Technologies: AWS (S3, EC2, RDS)

Databases: MySQL, PostgreSQL, Redis

AWARDS & PUBLICATIONS

Top 2 of Surgical Tool Detection (SurgVU 2024) by **MICCAI 2024**

Honored to receive **2nd place** recognition for our work on Surgical Tool Detection at MICCAI 2024.

Top 5 of Surgical Task Recognition (SurgVU 2024) by **MICCAI 2024**

Honored to receive **5th place** recognition for our work on Surgical Task Recognition at MICCAI 2024.

Evolving Prompts for Synthetic Image Generation with Genetic Algorithm on **MAPR 2023**

Published research on improving text-to-image prompt generation using an enhanced genetic algorithm with elitism mechanisms.

EDUCATION

B.A. in Computer Science — **Oct 2020 - Jun 2024**
The VNUHCM-University of Information Technology (UIT) — **GPA:** 8.41/10

CERTIFICATIONS

IELTS (6.5) by **IDP Education Ltd in Viet Nam** — **Feb 2024**