NGUYEN HUYNH HOANG KHA

AI ENGINEER

0769532711
www.khaportfolio.info
Tan Binh District, Ho Chi Minh City
nhhkha.91tn@gmail.com
Github Linkedin

ABOUT ME

AI Engineer with hands-on experience in NLP, Computer Vision, and Generative AI. Skilled in developing, fine-tuning, and deploying models in real-world applications. Participated in international AI competitions and contributed to practical AI solutions across multiple domains.

EDUCATION

University of Information and Technology
- Bachelor in Computer Science

LANGUAGES

- English (IELTS 6.0)
- Vietnamese

SKILLS

AI & Machine Learning Skills

- **Machine Learning & Deep Learning**: TensorFlow, PyTorch, Scikit-learn, XGBoost, LightGBM
- **Natural Language Processing (NLP):** Hugging Face Transformers, SpaCy, NLTK, BART, GPT, LLaMA
- **Computer Vision (CV):** OpenCV, YOLOv8, Detectron2, MMDetection
- **Model Optimization & Training**: Hyperparameter Tuning (Optuna, GridSearchCV), Bayesian Optimization

AI Deployment & MLOps

- **MLOps & Model Deployment:** Docker, Kubernetes, CI/CD, MLflow, FastAPI, Flask
- **Vector Search & Embeddings:** FAISS, Pinecone, ChromaDB, RAG Pipeline.
- **AI Serving & APIs: TensorFlow Serving,** Triton Inference Server, RESTful APIs, GraphQL
- **Data Processing & Feature Engineering:** Pandas, NumPy, Dask, Polars

WORKING EXPERIENCES

CYBERSOFT TECHNOLOGY CO.,LTD — Feb 2024 - Now

AI Engineer

- Optimized model performance through evaluation, tuning, and integration into production systems.
- Built and deployed AI models for real-world use cases such as enterprise chatbots and educational assistants.
- Developed teaching materials and demo projects on agentic AI for corporate training and internal adoption.

Mentor DA *June 2025 - Now*

- Helping and imparting DA knowledges about Power BI and SQL related to DA

AI instructor *June 2025 - Now*

- Helping and imparting DA knowledges about Machine Learning and Deep-Leanrning models about AI

FREELANCE — Nov 2022- Jan 2023

AI Engineer

- Developed custom AI chatbot solutions for small businesses to automate customer support and service handling.
- Integrated AI systems to automate business processes such as data processing, internal communication, and task management.
- Researched and implemented AI applications in financial trading using NLP and Large Language Models (LLMs) for market analysis and news interpretation.

CERTIFICATES

14 Basics Certificates and Badges and 1 Intermediate Certificate about Machine Learning and Generative AI from GOOGLE:

- *Introduction to Responsible AI*
- *Introduction to Large Language Models*
- *Introduction to Generative AI*
- *Responsible AI for Developers: Interpretability & Transparency*
- *Responsible AI for Developers: Fairness & Bias*
- *Machine Learning Operations (MLOps) for Generative AI*
- ***Inspect Rich Documents with Gemini Multimodality and Multimodal RAG***
- *Vector Search and Embeddings*
- *Introduction to Vertex AI Studio*
- *Create Image Captioning Models*
- *Transformer Models and BERT Model*
- *Encoder-Decoder Architecture*
- *Attention Mechanism*
- *Introduction to Image Generation*

Considered as a great top 10 AI Developer in Google Cloud Skill Boost

PERSONAL PROJECTS

MULTIMODAL VIDEO Q&A SYSTEM (RAG-BASED CHATBOT) 03/2024 - 08/2024

Model Researcher & AI Deployment

- Designed and deployed a multimodal RAG-based Q&A system capable of processing video, text, image, and speech inputs for context-aware information retrieval.
- Integrated Whisper (speech-to-text), CLIP (vision-language), and Transformer-based embeddings for multimodal fusion.
- Built an end-to-end pipeline: Inputs → Preprocessing → Embeddings (CLIP/Whisper/OpenAI) → Vectorstores (FAISS/Chroma) → LangChain + LangGraph orchestration → Response generation → Gradio interface.
- Optimized retrieval accuracy (+40%) and latency (-30%), leading to improved user engagement in e-learning and knowledge management applications.
- **Delivered an interactive prototype with natural multimodal Q&A, showcased in the Advanced AI for Business program.**
- Tech Stack: Python, PyTorch, Whisper, CLIP, Transformers, FAISS, Chroma, OpenAI API, Google Colab, LangChain, LangGraph, GPT Models, RAG, Gradio, Embedding Models

MULTI-HOP REASONING QA CHATBOT 05/2025 - 08/2025

AI Researcher

- Built an MVP QA system with multi-hop reasoning, tracing reasoning chains over DBLP/Wikidata knowledge graphs.
- Applied LLMs (LLaMA-3/Qwen-7B) for question decomposition (CoT) and GNNs (GraphSAGE/GAT) for multi-step relation encoding.
- Orchestrated workflows with LangGraph, merging LLM + GNN outputs into final answers with step-by-step explanations.
- **Delivered an interactive prototype (FastAPI + Streamlit) with caching (SQLite/FAISS) for faster, auditable queries.**
- Tech Stack: LLaMA-3/Qwen-7B, PyTorch Geometric/DGL, LangGraph, DBLP/Wikidata, FastAPI, FAISS.

MHQ-REACTRAG — RESEARCH-GRADE MULTI-HOP QA ASSISTANT 03/2024 - 06/2024

Model Researcher & AI Deployment

- Developed a research-grade multi-hop QA system combining Chain-of-Thought (CoT), ReAct loops, and RAG to provide verifiable answers with reasoning traces.
- Integrated tool-use capabilities (web search, vector retrievers, Neo4j KG, Python sandbox) and a FEVER-style verifier to ground claims and reduce hallucinations.
- Designed a science-grade pipeline with logging, benchmarks, ablation studies, and reproducible experiments for evaluation across QA datasets (HotpotQA, StrategyQA, GSM8K).
- Delivered an MVP with FastAPI backend + Streamlit/Gradio UI, supporting FAISS (dev) and scalable vector DBs (Milvus/Pinecone) in production.
- Tech Stack: LLaMA-3/Qwen-7B, LangChain/LangGraph, PyTorch, FAISS/Milvus/Pinecone, Neo4j, Gradio.

VIRTUAL - TRY ON 2D 03/2025 - 06/2025

Model Researcher & AI Deployment

- Pipeline: User photo + garment image → Human Parsing (segmentation) → Pose Estimation → U-Net image generation → Image Warping → Final try-on visualization.
- Built a 2D Virtual Try-On (VTON) system for e-commerce, allowing users to try clothes virtually using only a single image.
- Leveraged U-Net for segmentation, OpenPose/MediaPipe for pose alignment, and image warping for realistic rendering.
- **Result: Produced realistic try-on outputs with high alignment quality; deployed as a demo for online fashion platforms.**
- Tech Stack: Python, OpenCV, PyTorch, U-Net, OpenPose/MediaPipe, Image Warping

STUDENT'S BEHAVIOUR DETECTION 03/2025 - 06/2025

Model Researcher & AI Deployment

- Pipeline: Classroom video → YOLOv8 + MediaPipe for face/pose detection → Behavior recognition (drowsiness, yawning, distraction) → Time-series analysis → Prediction of academic performance.
- Designed a computer vision system to monitor student engagement in real-time.
- Combined behavioral cues (eye closure, yawning, head pose) with time-series prediction models to estimate attention levels and correlate with academic outcomes.
- **Result: Generated early-warning alerts for at-risk students, helping educators improve classroom effectiveness.**
- Tech Stack: OpenCV, MediaPipe, YOLOv8, Python, Scikit-learn, CSV Time-series Analysis, Behavior Prediction Models

PUBLICATIONS / PAPERS

Retrieval-Augmented Generation Architectures: Integrating LangChain, LangGraph, Vector Stores, and Open-Source Large Language Models -A Survey and Case Study with GPT-OSS-20B

Author: Kha Nguyen