<Your name>

<Your address> | <Your email> | +84 <Your phone number> | linkedin.com/in/<Your linkedin username>/
github.com/<Your github username>/

Summary

AI Engineer with a strong focus on Generative AI, including LLM training/inference, diffusion pipelines, and agent systems. Proven ability to develop and implement AI applications in image recognition and generative models. Experienced in optimizing training pipelines, model evaluation, and deployment using technologies like VLLM and Triton Inference Server. Seeking to leverage expertise in AI/ML to drive innovative solutions.

Education

FPT University Ho Chi Minh **Oct 2021 – Jun 2025**
Bachelor's Degree in Artificial Intelligence *Ho Chi Minh City, Vietnam*

- **GPA:** 7.43/10
- **Relevant Coursework:** Machine Learning, Deep Learning, Natural Language Processing, Computer Vision

Experience

AI Engineer, BEQ Holding **Feb 2025 – Oct 2025** *Ho Chi Minh City, Vietnam*

- Developed Large Language Model (LLM) applications for medical reasoning, collecting 22k dataset samples
- Fine-tuned LoRA model for Medical Reasoning using DeepSeek R1 0528 distill Qwen 8B architecture
- Implemented quantization and deployed models on VLLM and TensorRT LLM for production use
- Built intelligent multi-agent chatbot using Langchain and Langgraph for context-aware responses
- Researched and developed workflow systems, implemented function calling tools, and built workflows on Windmill

AI Engineer Intern, GRADIENTS TECHNOLOGY **Aug 2024 – Feb 2025** *Ho Chi Minh City, Vietnam*

- Conducted research and development in generative AI stable diffusion for image processing
- Trained stable diffusion inpainting 1.5 models for object insertion (8K images) achieving SSIM: 0.833, LPIPS: 0.060, FID: 14.996
- Developed white balance correction models using 10K image dataset
- Optimized training pipeline for stable diffusion inpainting on low-end GPU (RTX 2080 TI with 12GB VRAM)
- Converted stable diffusion inpainting 1.5 models to ONNX format and deployed on Triton Inference Server

AI Engineer Intern, QAI FPT Software **Jan 2024 – Apr 2024** *Quy Nhon, Vietnam*

- Participated in AI application development for image recognition in Personal Protective Equipment (PPE) detection
- Collected and processed 20k images, performed data labeling with bounding box annotations
- Trained YOLO detection model achieving 0.95 mAP50 for PPE detection
- Converted trained models to TensorRT format for optimized inference performance

Skills

Programming Languages: Python
AI Frameworks: PyTorch, Transformers, (Sentence)Transformers, PEFT, Langchain, Langgraph, pydantic-ai, Unsloth, TRL, HuggingFace, VLLM, TensorRT LLM, Triton Inference Server, Ultralytics, ONNX
AI IDEs: Trae, Claude Code
API Frameworks: FastAPI
Database: PostgreSQL, Qdrant
Workflow: Windmill
Tools: Git, Docker, Linux
Languages: Vietnamese (Native), English (Independent)