# ETHAN LEW

+60169098820 | ernest07.el@gmail.com

## Professional Summary

AI Full-Stack Engineer with 4+ years of experience designing and shipping production-grade agentic AI systems. Specialized in multi-agent orchestration using LangGraph, Retrieval-Augmented Generation (RAG) workflows, LLM evaluation frameworks, and scalable backend architectures across Amazon Web Services (AWS) and Microsoft Azure. Hands-on with model optimization (LoRA/QLoRA, GPTQ/AWQ) and full-stack delivery using NextJS, Go, Python, Redis, and PostgreSQL in regulated financial environments.

## Work Experience

**AI Full-Stack Engineer** | Maybank 2024 – Present

- Architected multi-agent loan automation system orchestrating specialized LangGraph agents for concurrent financial data retrieval (Go backend), financial modeling, and risk assessment, reducing processing from 3 days to 1 day across 500+ monthly applications
- Designed multi-agent ML evaluation framework using LangGraph where specialized AI agents act as domain-expert evaluators across distinct quality metrics, with NextJS frontend, Redis caching, PostgreSQL, and RESTful APIs, adopted company-wide for standardized model validation
- Developed 2 reference architectures for AI solutions across AWS and Azure, establishing standardized blueprints with security best practices and cloud component integration that reduced Enterprise Architecture (EA) approval time by 70%
- Developed enterprise applications using Microsoft Power Platform integrating Azure OpenAI, Claude, and Gemini for financial use cases
- Translated AI research papers into production solutions, collaborating with cross-functional teams using Git for version control

**AI Engineer** | Saratix AI Sdn. Bhd. 2022 – 2024

- Engineered end-to-end LLM optimization pipeline from CUDA kernels to Kubernetes, improving GPU utilization from 60% to 85% and reducing inference latency by 35%
- Developed production transcription system processing 600+ meetings monthly with 95% accuracy, reducing manual effort by 90% through WhisperX, Pyannote, and LLM summarization on AWS EC2 with FastAPI and RabbitMQ
- Trained and optimized Large Language Models (LLMs) through LoRA/QLoRA fine-tuning and hyperparameter tuning, developing RAG pipelines for intelligent chatbots
- Optimized models through custom CUDA kernels and quantization (GPTQ/AWQ), reducing model size by 50% and inference latency by 35%, enabling 40% cost savings
- Led cross-functional AI project teams through Agile methodology, delivering OCR systems and containerized deployments across cloud platforms

## Skills

**Programming Languages:** Python, SQL, Go, TypeScript, C++
**AI and LLMs:** LangGraph, CrewAI, Multi-Agent Systems, RAG Pipelines, GPT-4, Claude, LLaMA, Mistral, Fine-tuning (LoRA/QLoRA)
**Machine Learning (ML):** Model Training and Evaluation, Hyperparameter Tuning, Feature Engineering, CUDA, GPTQ, AWQ, Quantization
**Databases:** PostgreSQL, MongoDB, CosmosDB, Redis, Elasticsearch, Vector Databases
**Web and Backend:** NextJS, React, Tailwind CSS, FastAPI, RESTful API Development, Microservices Architecture
**Cloud and DevOps:** AWS, Azure, Google Cloud Platform (GCP), Kubernetes, Docker, Terraform, Git, CI/CD Pipelines

## Education

**Master of Artificial Intelligence (AI)** | University of Malaya 2023 – 2024
**Bachelor of Computer Science (CS)** | University Tunku Abdul Rahman 2019 – 2022