TAN WEI HAO

AI Engineer | Full-Stack Developer

CONTACT

- +60-11-11199174
- weihaotan.77@gmail.com
- Kajang, Selangor, 43000
- linkedin.com/in/tan-wei-hao-147b69149

PORTFOLIO

tanweihao0418.github.io

More Details

AI & ML SKILLS

- Deep Learning
- NLP & LLMs
- RAG Systems
- Computer Vision
- PyTorch & TensorFlow
- Fine-Tuning
- OCR Technology

PROGRAMMING

- Python
- C# / .NET
- Java
- JavaScript
- PHP
- SQL / NoSQL

Professional Summary

An innovative AI Engineer with hands-on experience in the complete lifecycle of AI solutions, from development to deployment. Expertise in fine-tuning and implementing Large Language Models (LLMs), developing advanced chatbots with Retrieval-Augmented Generation (RAG), and building high-accuracy computer vision systems. A skilled full-stack developer who leverages cutting-edge tools like GitHub Copilot to accelerate development and deliver robust, AI-driven applications.

Experience

AI Solution Engineer (Fullstack)

Mar 2024 - Oct 2025

TimeTec Cloud

- Fine-tuned Deepseek large language models using Python with Unsloth and QLoRA for advanced NLP applications
- Developed an AI-powered chatbot using the OpenAI Agent SDK, leveraging Retrieval-Augmented Generation (RAG) and a custom knowledge base with Qdrant for vector storage
- Enhanced HR solutions by implementing OCR technology to automate data extraction processes and improve workflow efficiency
- Utilized GitHub Copilot and MCP tools, including Playwright for web testing and GitHub MCP, to accelerate development workflows
- Developed and implemented new features for a large-scale HR application using frameworks like Vue.js, improving system functionality and user experience
- Designed and implemented RESTful APIs to support internal modules and facilitate third-party integrations
- Optimized database queries and stored procedures, significantly improving application performance and reliability

Software Engineering Intern

Oct 2023 - Feb 2024

Resort World Tech Lab - Genting Berhad

- Developed and trained a high-accuracy poker card detection model using YOLOv5 and PyTorch for real-time casino games
- Conducted comprehensive benchmarking and performance testing of the AI model, analyzing FPS, latency, and precision/recall metrics
- Created a WPF-based Baccarat game in .NET Core, integrating the AI model for automated gameplay
- Built a reusable .NET Class Library for card detection and data collection, and published it as a NuGet package for production use
- Handled multithreading and UI events in WPF with Dispatcher, async/await, and Task for smooth real-time camera processing

AI Intern

Feb 2023 - May 2023

Footfallcam - Meta Research

- Trained CNN models (ResNet50, MobileNet) with Keras & TensorFlow to detect store open/close status from surveillance images
- Applied PCA for dimensionality reduction and used SVD and regression analysis with Scikit-learn to predict sales outcomes
- Created an image data collection pipeline using FastAPI for efficient ingestion and processing of visual datasets
- Built automation scripts using Selenium to extract geographic coordinates and integrate OpenAI, Google Geocoding, and Bing Maps APIs
- Conducted testing on ESP32-CAM, analyzing image blurriness and estimating age/gender using cloud-based face recognition APIs

FRAMEWORKS

- ASP.NET
- Vue.js
- Angular
- FastAPI
- WPF
- Playwright

LANGUAGES

English	Fluent
Chinese	Native
Bahasa Melayu	Fluent

Education

Bachelor of Software Engineering
New Era University College
CGPA: 3.84
2021

Projects

AI-Powered ChatBot Platform

Developed a comprehensive full-stack chatbot platform with Vue.js 3, Python FastAPI, OpenAI API, and Qdrant vector database. Built universal Web Component for easy website integration with one-line script implementation. Implemented Retrieval-Augmented Generation (RAG) pipeline with semantic search capabilities, multi-format document processing system, and RESTful API with JWT authentication. Built custom web crawler with intelligent content parsing and advanced LLM-based document chunking for enhanced RAG system accuracy.

Real-Time Card Detection System

Developed a sophisticated real-time card detection system using C#/.NET 8, WPF, and multiple YOLO models (v5/v7/v8). Implemented advanced computer vision pipeline with frame buffering for 30 FPS processing, multithreading for non-blocking video capture, and CUDA GPU acceleration for optimized inference performance. Achieved 80% reduction in false positives through ensemble confidence threshold optimization and dynamic model switching. Built WPF application with live preview interface and integrated comprehensive logging for performance monitoring.

Certifications

Microsoft Azure Fundamentals (AZ-900)
View Credential