

Yen-Hao Huang

hyenhao1@gmail.com | 0928305017 | linkedin.com/in/yenhaohuang/ | github.com/yenhao-huang

Profile

Hello, I am Yen-Hao Huang, a Master's graduate in Computer Science from National Taiwan University. I have two years of experience in machine learning development, working on projects involving chatbot development, RAG systems, and performance optimization tools. Through these projects, I have developed a strong understanding of the end-to-end AI development pipeline, including building ETL pipelines, designing models, and deploying them into production systems.

Skills

Data Processing NumPy, Pandas, OpenCV	Model Development PyTorch, HuggingFace, OpenAI API	Model Evaluation Scikit-learn, Matplotlib
Deployment FastAPI, Docker, Node.js	Programming Languages Python, C++	Tools & Platforms Git, Linux, VSCode

Education

National Taiwan University , M.S. in Computer Science	Sep. 2022 – Dec. 2024
<u>Master's Thesis: Online Tuning of DRAM Controllers Using Continual Reinforcement Learning</u>	
- Selected for presentation at the Work-in-Progress Session of the top-tier DAC conference - Applied reinforcement learning to dynamically tune DRAM parameters for rapidly changing workloads - Improved efficiency by 35% with a multi-agent continual learning approach	
National Chung Cheng University , B.S. in Computer Science	Sep. 2018 – Jun. 2022

Experience

NLP Software Engineer , Wingene	Oct. 2025 – Present
<u>Large-Scale Retrieval</u>	
Designed and implemented a large-scale retrieval system for structured and unstructured data	
Research Assistant , National Taiwan University	Sep. 2022 – Dec. 2024
<u>Learning-Based RAID Controller Tuning Tool</u>	
- Collaborated with HPE to develop a machine learning-based RAID tuning tool, achieving 30% performance gain - Captured system behaviors using profiler to track cache, memory, disk, and process metrics - Introduced a rejection mechanism to handle unknown workloads, enhancing model generalization	

Projects

RAG-Based Study Abroad Chatbot	May. 2025 – Aug. 2025
RAG, Causal LLM, Chatbot	
- Developed a chatbot system for FAQs, user experiences, and legal/military consultations - Integrated LLMs with external data via LangChain - Optimized chunking and embeddings, reaching 85.5% Precision@5	
LLM Sentiment Analysis Tool	Apr. 2025 – Jun. 2025
DeepSeekR1, Memory Optimization, Sentiment Analysis, Deployment	
- Applied LLM compression (LoRA, quantization, gradient checkpointing) to cut memory usage by 16× - Fine-tuned model on sentiment analysis task, achieving over 95% accuracy on tens of thousands of reviews	
Pothole Detection System	Jan. 2025 – Mar. 2025
YOLOv11, Object Detection, Model Fine-tuning	
- Developed a detection pipeline for road potholes, achieving mAP@0.5 of 0.82 - Deployed model with FastAPI to enable inference via Web API	
Zero-Shot Multimodal Classification System	Sep. 2023 – Jan. 2024
Multimodal Learning, Vision-Language Model, Zero-shot Classification	
- Designed a zero-shot learning framework using SigLIP for classifying unseen categories without labeled examples - Improved ImageNet-1K accuracy from 60% to 73%	