TEETAT KARUHAWANIT

Bachelor of Electrical Engineering, Chulalongkorn University (2020-2024)

📱 +66 625545554

✉ ttwanit@hotmail.com

in Teetat Karuhawanit

SKILLS

- Python (4 years of experience), Quart microframework
- Object-Oriented-Programming (C++)
- SQL (Azure Data Studio, Google BigQuery)
- Golang
- Microsoft Azure Portal, Amazon Web Services (AWS)
- Docker
- Unit testing and integration testing framework (Pytest, Selenium)
- Git version control
- Natural Language Processing, Large Language Model
- Infrastructure as code tools (Terraform, Azure CLI, Azure resources manager)
- Neural Network and Machine Learning (Tensorflow, Keras, PyTorch, Sklearn)
- Database administration (Azure Data Lake Gen 2, PostgreSQL, Apache Spark)

WORK EXPERIENCES

AI Engineer at HDmall (March 2024 - now)

- Led SQL-based data cleaning and deployed a **Retrieval-Augmented-Generation (RAG) Large Language Model (LLM)** on **Microsoft Azure Portal**, integrating **OpenAI GPT-4o's** capabilities. Developed a **multimodal LLM with OCR** to read images of PDP packages sent by customers, enhancing customer satisfaction and achieving a 125% faster response time compared to the original CX process. **Contributing to the company's ability to raise $5.6M in funding in Southeast Asia.** (Read more: https://techcrunch.com/2024/04/02/hd-mall-southeast-asia-ai-healthcare/)
- **Fine-tuned a Large Language Model (LLM)** using the **QLoRA method** to optimize performance and reduce hallucinations. Integrated **semantic ranking through Azure** and adjusted hyperparameters to maximize LLM performance.
- Integrated **Elasticsearch** and **text-to-SQL function calling** to enhance the efficiency of the Large Language Model (LLM). Implemented and managed various databases, currently using **PostgreSQL** with **pgAdmin** for monitoring, to develop a **hybrid search algorithm with Google VertexAI Custom Search Engine**. Employed **integrated vectorization techniques in Azure** to enable real-time data processing and updates every two days.
- Implemented a **knowledge graph database with Retrieval-Augmented-Generation (RAG)** to reduce hallucinations and enhance the precision and response speed of a Large Language Model, achieving a 75% improvement of speed.
- Utilized **Postman, Flask framework, Redis memorystore by GCP, and Python Requests library** to call APIs and integrate with LINE chat. This integration significantly supported the CX team by reducing their workload by 55%, handling approximately 100 customer chats per minute.
- Deploying a **Retrieval-Augmented-Generation (RAG) Large Language Model (LLM)** on **Amazon Web Services (AWS) using AWS Bedrock, Lambda, Elasticache, EC2 and Elastic Beanstalk** for the HDcare team, leveraging the capabilities of **Claude 3.5 Sonnet**. Implemented a **Tool Use (similar to OpenAI's function calling)** to enable **full-text search using text-to-SQL** on an Excel file corpus. Incorporated **OCR** to read product information from the PDP website, classifying customer priority and identifying potential buyers. This system effectively helped the HDcare team to filter out low-priority chats and summarize the chat history, reducing their workload significantly.

Enterprise Solutions Specialist at Thinking Machines Data Science (June 2024 - now)

- Assisted in deploying an **Agentic-workflow type Retrieval-Augmented-Generation (RAG) Large Language Model (LLM)** through **Microsoft Azure Portal and Langgraph** to create a **State Graph**. Utilized **Terraform** for resource provisioning and developed a streamlined CI/CD pipeline using **GitHub Actions**. Integrated the LLM with LINE OA using **FastAPI** to support product sales for clients including Sabina, Tidlor, and GoWabi.
- Estimate the Total Cost of Ownership (TCO) including cloud costs and labor expenses for the project. Calculated appropriate customer charges to ensure they were adequate yet competitive, aiming to maximize the company's profit.
- Collaborated with LINE Corporation to implement a Large Language Model solution integrated with LINE and LINE Shopping, targeting SMEs across Thailand to enhance their sales capabilities.

Internship (2023)

- Completed an internship at NCKU in Taiwan, where I developed a **computer vision** adaptive control system using the **PyTorch** library. Utilized the advanced deep learning model **YOLOv8** for real-time **video segmentation** and object control.

Relevant experience (2023)

- Implemented a **Generative Adversarial Network (GAN)** using the **TensorFlow** library to translate multimodal brain images from MR to CT scans. This project, in collaboration with Chulalongkorn University Hospital, aimed to reduce radiation exposure for brain cancer and tumor patients, minimize the time required for duplicate brain scans, and lower costs by eliminating the need for external image translation software by 2,000,000 THB.
- Responsible for cleaning the data pipeline and correlating it with partner data to analyze customer service utilization using **Microsoft Azure SQL databases**. This project reduced HDmall's advance payment costs for redeemed coupons from an external company by 400,000 THB.

LANGUAGES

- Thai (Native speaker)
- English (Professional)

References available upon request