TAI NGUYEN PHU

Binh Duong, Ho Chi Minh • tainguyenphu2502@gmail.com • (+84) 945 409 269
github.com/YuITC • yuitc.github.io/Tai-Nguyen-Phu-Portfolio

Final-year Computer Science student with a strong background in Machine Learning and Deep Learning, currently specializing in Large Language Models (LLMs) and generative AI. Seeking an AI Engineer internship or entry-level position focused on the development, optimization, and deployment of LLM- and generative AI-powered applications.

EDUCATION

Bachelor of Computer Science **Sep 2022 - Jun 2025**
University of Information Technology - VNUHCM, Vietnam **GPA:** 8.24/10

CERTIFICATIONS

IELTS Academic (Overall: 7.0) **Feb 2022**

PROJECTS

Vietnamese Legal Document Retrieval **Apr 2025 - Apr 2025**

Technologies: Python, PyTorch, SentenceTransformers, FAISS, Gradio, Docker, Pandas

- Fine-tuned a multilingual Sentence-BERT model on a curated Vietnamese legal corpus, achieving domain-specific semantic representation for accurate document retrieval.
- Boosted retrieval performance dramatically, achieving NDCG@10: 60.4% and MAP@10: 53.6% on the MTEB benchmark (BKAI Legal Retrieval dataset), compared to only 0.023% and 0.014% with the pretrained model.
- Designed a full semantic retrieval pipeline including data preprocessing, model fine-tuning, evaluation with MTEB benchmark, and real-world deployment.
- Implemented GPU-accelerated FAISS index (approximately 100,000 documents) for ANN search with sub-second latency.
- Built end-user Gradio interface, containerized with Docker.

Semantic Book Recommender **Mar 2025 - Apr 2025**

Technologies: Python, PyTorch, HuggingFace Transformers, SentenceTransformers, Prompt Engineering, Gradio, Pandas, NumPy

- Developed a semantic-based book recommendation system leveraging sentence embeddings and large language models (LLMs) for context-aware, personalized suggestions.
- Implemented vector search using pre-trained sentence embeddings to find semantically similar books beyond keyword-based methods.
- Applied zero-shot classification via pre-trained LLMs to categorize books into genres without requiring labeled datasets.
- Conducted sentiment analysis on user reviews to enhance recommendation quality by factoring reader opinions.
- Built an interactive Gradio dashboard enabling users to input preferences and receive tailored recommendations in real time.

Job Application AI Assistant **Mar 2025 - Mar 2025**

Technologies: Python, PyTorch, Llama, LangChain, Prompt Engineering, Groq API, FAISS, Streamlit

- Developed an AI-powered assistant utilizing Llama 3.3-70B via Groq API and LangChain to generate personalized cold emails and tailored job application strategies.
- Integrated Groq API for ultra-fast LLM inference, enabling low-latency, high-throughput real-time interactions.
- Implemented a semantic search system using FAISS to match user profiles with relevant job postings and enhance recommendation accuracy.
- Built and deployed an interactive Streamlit web application supporting real-time, context-aware job application assistance.

Scene Text Recognition **Feb 2025 - Mar 2025**

Technologies: Python, PyTorch, YOLOv11m, OpenCV, CRNN, FastAPI, Ray Serve, Streamlit

- Built an end-to-end OCR pipeline combining YOLOv11m for scene text detection and CRNN (ResNet34 backbone) for sequence-based recognition.
- Trained and fine-tuned models on the ICDAR2003 dataset, achieving approximately 88% precision in real-world text detection tasks.
- Deployed scalable, high-performance API services using FastAPI and Ray Serve, optimized for GPU acceleration.
- Developed a modular Streamlit web application for real-time text detection and easy integration into external systems.

Data Science Job Salary Prediction **Dec 2024 - Jan 2025**

Technologies: Python, Scikit-learn, XGBoost, Pandas, NumPy, Matplotlib, Seaborn, Streamlit

- Developed a machine learning-powered salary prediction model for Data Science jobs, trained on real-world Glassdoor job market data.
- Engineered and optimized features through exploratory data analysis (EDA), encoding, and feature importance evaluation.
- Trained and compared multiple models (XGBoost, Random Forest, Decision Tree, Gradient Boosting, Linear Regression, SVM), achieving $R^2 = 0.815$ with XGBoost. Maximized model performance via hyperparameter tuning with GridSearchCV.
- Built and deployed an interactive Streamlit web application, enabling users to input job attributes and receive real-time salary predictions.

SKILLS

Programming Languages: Python, C++, SQL, HTML

Frameworks: PyTorch, Transformers, LangChain, LLaMA, LoRA, FAISS, OpenCV, Scikit-learn, XGBoost, NumPy, Pandas

Deployment: Git, GitHub Actions, Docker, AWS, FastAPI, Ray Serve, Weights & Bias, Streamlit, Gradio

Soft Skills: English, Presentation, Critical Thinking, Teamwork, Problem Solving