Pham Thanh Trung

(+84)-845604327 phmtrung2@gmail.com Linkedin Portfolio

Professional Summary

A product-oriented engineer with a *strong ability to research AI (ML/DL), robust model training skills, and a visionary architecture design mindset*. Excelled at designing, orchestrating, and implementing best-fit AI approaches to *deliver production-ready solutions*. *Practical and highly experienced* at conducting technical research to overcome development bottlenecks without trading off the delivery schedule.

Education

Kyunghee University Global Campus **Feb 2025 - Present**

PhD combined program in Artificial Intelligence

FPT University Ho Chi Minh City **Aug 2020 - Dec 2023**

Bachelor of Artificial Intelligence (GPA: 8.87/10); Excellent graduate thesis (graded 9.2/10)

Works Experience

Kyunghee University **Feb 2025 - Present**

AI Researcher - PhD student *Yongin City, Korea*

- AI Researcher at the university laboratory, specializing in quantum computing and Multi-Modal Large Language Models (MLLMs). Conducted research to leverage next-generation quantum computing capabilities for advanced AI applications, including LLMs. In the initial stage, reduced the **3 million training parameters** of the VGG19 classification layer to just **150 quantum training parameters**, achieving a **0.025 improvement** in image classification accuracy.

IMT Solutions **Feb 2024 - March 2025**

Artificial Intelligence Engineer (Mid-level) *HCM City, Vietnam*

- AI Researcher and Engineer within the AI Department. Mainly worked on projects involving Large Language Models (LLMs), MLOps, computer vision, and multi-agent AI systems.
- Architected, implemented, and **delivered** an agentic KIE system for a multi national F&B corporation, achieving a **peak accuracy of 94%** and maintaining an operational accuracy **above 90%**. Reverse engineered enterprise-grade KIE cloud services to reduce the overall agentic KIE system's inference cost by more than **40%**.
- Designed and delivered an automated agentic system to convert scanned PDF appliance manuals into the DITA structured format, reducing **weeks of manual labor** for digitization tasks to just **6-16 minutes** per document (averaging **70 pages** per document).
- Implemented and trained a multi-agent, domain-driven machine translation system, improving the **BLEU score** from **0.74 and 0.68** (for GPT-4+RAG and **Google Translate**, respectively) to **0.93** using an **NLLB + Wikidata** multi-agent NMT pipeline.
- Designed scalable AI architectures tailored to customer requirements and coordinated implementation tasks among team members.
- Skills: Agentic AI · Python (Programming Language) · Natural Language Processing · LLMs · Computer Vision · AI/ML systems development · Generative AI · Team management

GMO-Z.com RUNSYSTEM Joint Stock Company **Nov 2022 - Feb 2023**

Artificial Intelligence Research Intern *HCM City, Vietnam*

- AI Researcher within the R&D Department, primarily focusing on computer vision applications for eKYC. Proposed Fourier transform approaches for deepfake and facial anti-spoofing classification, improving the overall classification **F1 score** from **0.86** to **0.91**.

Selected Projects

Key Information Extraction from Documents **May 2024 - Feb 2025**

- **Role**: Team leader, main contributor.
- **Description**:
  - ◇ Researched and applied multi-modal LLMs to extract key information (invoice numbers, prices, item names,...) from scanned and digital documents.
  - ◇ Developed an extraction system leveraging Azure Document Intelligence services, featuring a CI/CD pipeline and a semi-automated system for recognizing new document layouts and configuring labeling workflows.
  - ◇ Analyzed and reverse-engineered Azure Document Intelligence custom model services to enhance internal system capabilities.
  - ◇ Implemented multi-modal, few-shot document layout classification services to support layout-oriented key information extraction system.

- ◇ Enhanced extraction performance by designing and integrating a multi-agent processing pipeline.

- **Stack used:** Azure Document Intelligence cloud services, Langchain, Fastapi, Pytorch, Huggingface transformers, Huggingface TGI, Qdrant, Gemini API, OpenAI API, azure pipeline and docker.

Pali ancient language of the Buddhism Tipitaka machine translator **Feb 2023 - Sep 2025**

- **Role**: Personal project
- **Description**:
  - ◇ Researched and implemented a parallel sentence mining process for an extremely low-resource ancient language using LaBSE and Meta's LASER framework. Mined and curated comprehensive parallel Pali-English datasets. Resulted in **120K parallel sentence pairs** curated dataset of Pali-English.
  - ◇ Modified and retrained NLLB-200 (1.3B and 3B parameter versions) on curated Pali-English datasets. Fine-tuned the BGE-M3 multi-lingual embedding model on the Pali-Vietnamese-English language cluster to be used in future data mining.
  - ◇ Conducted extensive experiments to optimize and improve the performance of NLLB models. Improved the performance of baseline models from **BLEU score of 0.5542 to 0.7837**.
  - ◇ Deployed and served optimized NLLB models. Implemented the front-end web interface for the translation platform.
- **Stack used:** Pytorch, FastAPI, Huggingface TGI, Huggingface transformer, wandb, docker, LASER (embedding), bge-m3, langchain and Faiss.

Research on Large Language Models (LLMs) to enhance domain-driven neural machine translation **Jan 2024 - Feb 2025**

- **Role**: Team leader, main contributor.
- **Description**:
  - ◇ Integrated multiple approaches (RAG, Chain-of-Thought, multi-agent systems, and end-to-end models,...) to boost translation performance without requiring post-deployment retraining.
  - ◇ Fine-tuned the NLLB model on specialized, domain-driven machine translation datasets.
  - ◇ Conducted parallel corpus data mining to enrich and refine the fine-tuning datasets. Expanded **60%** of the initial domain driven dataset with automated data-mining process.
- **Stack used:** Fastapi, gradio, NLLB model, Huggingface transformers, langchain

Medical vision Q&A LLM system **Feb 2025 - Mar 2025**

- **Role**: Personal project
- **Description**: Fine-tuned InternVL2.5 on the VQA-RAD dataset to demonstrate the medical visual question-answering capabilities of open-source vision-language models. Developed a ChatGPT-style user interface supporting registration, authentication, and file attachments. Built and deployed LLM serving APIs, web backend servers, microservice interfaces, and database.
- **Stack used:** FastAPI, Huggingface TGI, Huggingface transformer, SQLAlchemy, Pytorch and Streamlit.

Technical Skills

MLOps & Agentic AI: LangChain, FastAPI, vLLM, Hugging Face TGI, Docker, Azure Pipelines, Weights & Biases. Exp. fine-tuning/serving open-source LLMs/VLMs (LLaMA, QwenVL, InternVL) and deploying closed-source APIs.
ML/DL Framework: Highly proficient in **PyTorch**, **TensorFlow/Keras**, and **ONNX**.
Core AI Architectures: Transformers (BERT, GPT, T5, ViT, Swin), CNNs (ResNet, UNet), RNNs (LSTM, GRU), Generative Models (Diffusion, GANs, VAEs), Reinforcement Learning, Evolutionary Algorithms (NEAT, GA).
Quantum Computing: Pennylane, TorchQuantum.
Cloud and HPC management: Managing, monitoring, and troubleshooting **high-performance GPU servers (NVIDIA H200, H100, A100, A6000)**. Allocating HPC resources for scalable team workloads. Azure Cloud services, especially **Azure AI services**.
Programming languages: Python, C & C++, Java, SQL, Javascript, ReactJS, HTML/CSS.

Publications

- Journal (Q1): Tran, D. T., Nguyen, N. D. H., **Pham, T. T.**, Tran, P. N., Vu, T. D. T., Nguyen, C. T., ... & Dang, D. N. M. (2025). SwinTExCo: Exemplar-based video colorization using Swin Transformer. Expert Systems with Applications, 260, 125437.

Achievements

- 100% Scholarship at Kyunghee University for PhD combined program.
- Champion prize of FPT Edu Research Festival-ResFes finale 2023
- 100% Scholarship at FPT University HCM.
- Third prize of FPT Education Mathematics Olympic 2021.
- Silver medal in the informatics category at the Traditional Olympic 30/4 Contest 2018.