Le Duc Anh Tuan (Charles)

LLM Engineer

tuanlda78202

tuanleducanh78202@gmail.com

Education

- **Hanoi University of Science and Technology** — Hanoi, Vietnam
  Bachelor of Science in Data Science and Artificial Intelligence — *Oct. 2020 - Aug. 2024*
  - **Thesis**: Efficient Class Incremental Learning for Object Detection

Experience

- **Moreh, Inc.** — Seoul, South Korea
  GPU Engineer — *May. 2025 - Oct. 2025*
  - Implement a pure HIP C++ version of OpenAI's GPT-OSS from scratch (without rocBLAS/hipBLAS); optimize model loading, batching, multi-streaming, multi-GPU communication, CPU–GPU–SRAM memory access, FlashAttention, matrix-core GEMM, MoE load balancing, etc. Achieved 30k TPS (20B) and 10k TPS (120B) on a single node with 8× AMD MI250x GPUs.
- **Menlo Research** — Singapore, Singapore
  LLM Researcher — *Nov. 2024 - May. 2025*
  - Developed a lightweight Speech Tokenizer (22M) using Residual Vector Quantization, achieving SOTA results on viVoice and LibriSpeech by extending the codebook size to 2048, training a two-phase training process (KL+CE and CE loss), contributing to the open-source WhisperSpeech codebase; outperformed original Whisper and PhoWhisper
  - Researched the Speechless model, a modified Llama 3.2 1B architecture, to generate synthetic semantic audio representations from multimodal inputs using ASR datasets; paper accepted at Interspeech 2025 (Rank A)
  - Modified the Llama tokenizer and performed continual pre-training on semantic tokens from ASR datasets, followed by post-training with mixed raw text, sound-text, and noise sound datasets (filtered by language identification, deduplication, length, and quality) to align with user preferences
  - Published the package-modularized Ichigo on PyPI, supporting an asynchronous API for platform developers and implementing audio chunking with overlapping to support long audio input for ASR
- **Viettel Group** — Hanoi, Vietnam
  Data Scientist — *Apr. 2024 - Nov. 2024*
  - Developed a multi-agent Conversational Recommendation System with multimodal capabilities, supporting vision input and speech-to-speech interaction with end-users
  - Implemented AdaptiveICL to align with pre-defined expertise plans and designed a synthetic data pipeline for fine-tuning reasoning, SQL query generation, and function calling
  - Built Retrieval, Ranking, and Query tools for database interaction; implemented a Candidate Bus for item candidate storage and Web Search for external resource integration
  - Engineered an end-to-end system, including a Docker-wrapped API to bridge Application and Infrastructure layers

Projects

- **Leo (2025)**
  - Architected an LLMOps system for a personal AI assistant; encompassing Data, Feature, Training, Inference, and Observation components, following clean architecture principles
  - Implemented an offline pipeline that retrieves data from data services and stores on S3; designed an ETL pipeline to crawl links and perform quality filtering; set up a feature generation pipeline for fine-tuning datasets and creating vector embeddings indexed in MongoDB for Hybrid Contextual Retrieval; and established a training pipeline with evaluation and serving model on HF/AWS endpoints, all orchestrated by ZenML
  - Designed an online pipeline featuring an agentic RAG system, served via API using LiteLLM; utilizes summarization and retrieval tools (powered by fine-tuned LLM endpoints and a vector index database), supports Search MCP server, and incorporates observability components through prompt monitoring and RAG evaluation
- **Gemini Omni (2024)**: Developed a real-time web application showcased at Google I/O Extended Hanoi, featuring speech-to-speech functionality, multimodal integration, and RAG for event updates
- **Detect Cheating in Examination (2022)**: Researched and deployed real-time cheating detection solution using Pose3D and VideoMAE for 50-person exam rooms; featured on VTV24, DanTri, HUST, etc

Publications

- **Speechless (Interspeech, 2025)**: Speech Instruction Training Without Speech for Low Resource Languages
- **Poseless (arXiv, 2025)**: Depth-Free Vision-to-Joint Control via Direct Image Mapping with Vision Language Model

Skills

- **Areas of Interest**: Multimodal LLMs, Multi-agent Systems, LLM Systems, High Performance Computing
- **Programming Language**: Python, C++, CUDA, HIP, Triton, TypeScripts, Java, Javascripts, SQL
- **LLMOps**: Docker, Kubernetes, AWS S3/Bedrock/SageMaker, MLflow, Airflow, ZenML, Weaviate, WandB
- **Framework**: PyTorch, TensorFlow, Hugging Face, vLLM, SGLang, LlamaFactory, Langchain

Certificates

- **Vietnamese Standardized Test of English Proficiency** 2024-2026
  - **Level**: B2

Volunteering

- **Google Developer Groups Hanoi**: Spoke at DevFest 2022 and organized Google I/O Extended 2022, 2023, 2024; DevFest 2022, 2023; IWDxFFE 2023 and Build with AI 2024; certified by Google's Global Headquarters
- **SheCodes Vietnam**: AI Mentor of SheCodes Hackathon Hanoi 2023
- **Nestlé**: Ambassador of MT SparkTheNext Leaders Program 2023 and 2024
- **AIESEC**: Representative of Mini Leadership Conference 2022
- **VinAI Research**: Technical Collaborator at AI Day 2022

Achievements

- **Top 1,** Viettel Digital Talent 2024 (Data Science and Artificial Intelligence)
- **Third Prize,** Excellent Students Contest in Math 2019 (Provincial Merit Competition)
- **Winner,** Innovation Lab Asia CrowdPitch
- **Best Incubatee,** TechYouth (VinUniversity)
- **CCMG Grantee,** Cyberport (2nd largest incubator in Hongkong with 5 unicorns)
- **Top 1,** X-Challenge by VCCorp
- **Top 1,** Prometheus in digital transformation by European Union (out of 4000 teams)
- **Top 1,** Business Challenge 6 by Vietnam National University
- **Top 5,** Youth Impact Entrepreneurs by PNJ (VN30 Index)
- **Top 6,** Hult Prize Asia Summit 2022 (out of 1000 teams)
- **Top 20,** University Startup World Cup (out of 5000 teams)
- **Top 30,** Moonshot Global (out of 3000 teams)
- **Top 100,** XPITCH global (out of 4000 teams)