NGUYEN VAN TU

AI Engineer

0964511270 nvtu2305@gmail.com github.com/Tuprott991 linkedin.com/in/nguyen-van-tu

Education

VNUHCM - University of Science **2022 – 2026**

Bachelor of Information Technology — Major: Computer Science *Ho Chi Minh City*

- **Coursework**: Machine Learning, Natural Language Processing, Computer Vision, Multivariate Data Analysis, Data Structures & Algorithms, OOP, Databases, Calculus, Linear Algebra, Probability & Statistics
- **GPA:** 3.38/4.0 - **Foreign language:** IELTS 6.5

Experience

Prudential Vietnam **July 2025 – Present**

AI Engineer Intern *Ho Chi Minh City*

- Built **Agentic AI applications** using **LangChain, ADK, Vertex AI, vLLMs, MCP** to automate knowledge inquiry and underwriting workflows.
- Reduced information search time for internal agencies and claim processes, contributing to a **17% improvement in key business KPIs**.
- Contributed to **voice digital signature** and **text-to-speech** modules for secure and automated internal processes.

Vnemex Co Ltd **April 2025 – July 2025**

AI Intern *Ho Chi Minh City*

- Worked on an **eKYC** pipeline with face verification and liveness detection.
- Applied **PyTorch**, **ONNX**, **ResNet**, and **FASNet** for **anti-spoofing**, and used **modified U-Net** for **human cell segmentation**.

Honors & Awards

- **Champions** of Web3 & AI Ideathon (2025 - National Hackathon, among 450+ teams) — More details
- **Finalist** International Data For Life 2025 (among 2600+ teams - Explainable AI in Medical Project)
- **Third Prize** in Intel AI training program 2025
- **Consolation Prize (4th place)** in AI Challenge HCMC 2025 (among 900+ teams)
- **Top 1** track AI-powered process redesign & **Finalist** in VPBank Technology Hackathon 2025
- **GStar Bootcamp** – NTI Global Talent 2025 (17% acceptance rate; Advanced NLP Course) — More details
- **Outstanding in Scientific Research** - University Award 2024 - 2025
- **Champions** of Line Follower Robot competition HCMUS (F-RACE) 2024
- **Awarded** for outstanding contributions to Youth Union and Student Association activities in 2024
  - *Role: **Vice Head** of M.A.T Communications Committee, University of Science – VNUHCM*

Publication & Research

Tu et al. **An Interactive System For Visual Data Retrieval From Multimodal Input**. *The International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM 2025).*

- Work under Dr.Dang Bui from Sep to Nov 2024 to develop a conversational and multimodal video event search event.
- Research and apply AI models like **CLIP**, Whisper, PaddleOCR, TransnetV2, and **GPT-4o API**, allowing users to **semantically retrieve** visual data using natural language, image, and voice.

Tu et al. **AIthena-Vision: Adaptive Temporal Multimodal Event Retrieval with LLM-generated Multiperspective Fusion** *The 14th International Symposium on Information and Communication Technology (SOICT 2025). - An update of our previous AIthena system*

- Researched **large multimodal models**, combining moment-aware **Perception Encoders**, scene text & audio extraction, temporal alignment, and LLM-driven multiperspective generation for precise multimodal retrieval.

Projects

Educhain | *github.com/Tuprott991/Educhain-AI* **Jan – Mar 2025**

- **Role**: Team Leader / AI Engineer
- **Description:** Led team & built a personalized learning platform integrating **LLMs** and **prompt engineering** to build agents for **lightRAG-chatbot**, **RAG-based** quiz & study guide generation, **user knowledge profiling**, and **file processing**.
- Fine-tuned **Qwen2.5-7B** using the **LoRA** method and deployed it with **vLLM** for optimized inference.
- Utilized **FastAPI** for fast and efficient request handling backend. Leveraged **LangChain** to build agents wix th optimized retrieval and LLM-database interactions.
- Implemented file understanding with **Azure Document Intelligence** to extract structured insights, and used **Azure Speech** for voice-based interaction and transcription.
- Integrated **lightRAG** to embed and retrieve personalized content from user-uploaded documents.
- Managed all data with **PostgreSQL**, leveraging **pgvector** for vector storage and **Apache AGE** for knowledge graphs.
- **Techs:** Python, FastAPI, LLMs, LangChain, Azure, PostgreSQL, LoRA, vLLM, lightRAG, Generative AI, ReactJS

BoneDiseaseVQA | *github.com/Tuprott991/BoneDiseaseVQA-2* **March – April 2025**

- **Description:** Developed and trained a **multimodal transformer** (natural language & image) for bone disease visual question answering, achieving **90.50%** accuracy on the validation dataset.
- Implemented a novel semi-open architecture leveraging **viHealthBERT** (medical BERT-style model) and **Vision Transformer** for feature extraction; applied a **cross-multimodal attention** mechanism for feature fusion.
- Incorporated a **learnable answer embedding** for the Transformer decoder, increasing precision by approximately **12%** compared to closed architectures.
- Leveraged **Gradio** for the web interface and **GPT-4o** to enhance diagnostic information.
- **Techs:** Python, PyTorch, BERT, Deep Learning, Transformers, HuggingFace, Torchvision, Gradio, OpenAI

Multimodal Video Event Retrieval | *github.com/Tuprott991/AIthena-Vision* **Aug – Nov 2025**

- **Role:** Team Leader / AI Researcher
- **Description:** Led team to develop an **AI-driven multimodal event retrieval** system based on natural language, scene, voice, OCR, and other metadata.
- Reduced search latency by **54%** through optimized keyframe extraction using **OpenCV** and the **AutoShot** model.
- Applied **PE-Core-BigG-448** and **BEiT-3** with ensemble paradigm for text and image embedding generation, enabling efficient vector search with **FAISS**. Leveraged **Elasticsearch** and **PaddleOCR, VietOCR** for text-in-image retrieval.
- Enhanced retrieval performance with **multimodal inputs** (text, voice, objects); integrated **Gemini-2.5-flash** for query refinement and visual question answering, and employed **WhisperX** for accurate real-time speech-to-text conversion.
- **Techs:** Python, Transformer, FastAPI, Timm, HuggingFace, Numpy, Multimodal, Semantic Search

Technical Skills

- **Languages**: Python (Proficient), Javascript (Proficient), C++ (Good)
- **Libraries/Frameworks**: Pytorch, Tensorflow, vLLMs, FastAPI, LangChain, Google ADK, LLMs, RAG, PyTorch, Transformers, Scikit-learn, Pandas, Milvus
- **Tools**: Azure, GCP, AWS, Git, Docker, Jira, CI/CD