Huynh Cao Tuan Kiet

Ho Chi Minh City, Vietnam hctuankiet243@gmail.com (+84) 905171069 Tuan Kiet Huynh Cao
kiethuynh.me KHONGPHAITKID

Computer Science student building LLM applications, AI agents, retrieval pipelines, and real-time deep learning systems. Strong foundation in machine learning, statistics, algorithms, and software engineering.

Education

University of Science, Ho Chi Minh City, Vietnam *Oct 2022 – Aug 2026*

- Bachelor of Science in Computer Science – Fourth-year student
- GPA: 3.7/4.0 (or 8.97/10)
- **Computer Science Coursework:** Machine Learning, Deep Learning, Natural Language Processing, Data Mining, Graph Neural Networks, Big Data, Algorithms & Data Structures, Object-Oriented Programming, Databases, Computer Networking, Computer Architecture, Operating Systems
- **Mathematics Coursework:** Discrete Mathematics, Calculus I, Calculus II, Linear Algebra, Probability & Statistics, Combinatorics, Applied Statistics, Applied Mathematics

Experience

AI Engineer Intern *Neurond Technology JSC* *Jun 2025 – Aug 2025*
AI Engineer (Probation) *Neurond Technology JSC* *Sep 2025 – Nov 2025*

- Designed and developed a full-stack internal AI chatbot platform, handling frontend and backend implementation with Next.js.
- Built LLM-powered conversational features using AI SDK and OpenAI APIs, enabling internal knowledge access and task automation.
- Designed AI agents with tool execution workflows, state management, and iterative reasoning patterns for internal task automation.
- Implemented backend services and data storage with PostgreSQL to support chat history, agent state, and evaluation data.
- Prototyped speech AI workflows for speech-to-text and real-time speech-to-speech interaction in web-based AI systems.
- Benchmarked LLM quality, latency, and cost using structured evaluation metrics to support model selection.
- Deployed and operated AI services on Microsoft Azure, using Azure AI Foundry and Azure Blob Storage for model workflow and application data integration.

Research Experience

Autonomous Driving – Real-Time LiDAR Semantic Segmentation *Nov 2025 – Present*

- Researching real-time semantic segmentation of LiDAR point clouds for autonomous driving perception.
- Trained and evaluated deep learning models on SemanticKITTI and nuScenes for large-scale 3D scene understanding.
- Benchmarked segmentation quality and efficiency using mIoU, per-class IoU, inference latency, and FPS.
- Benchmarked and optimized accuracy-latency trade-offs for real-time inference on resource-constrained systems.

Publication

"RangeViM: Full-Resolution Range-View LiDAR Segmentation with a Lightweight Hybrid Vision Mamba Backbone,"
First Author • International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), 2026. **(CORE B)**

Projects

Custom Chatbot Platform

- Architected an end-to-end LLM platform integrating multiple model backends through a unified API gateway, enabling routing between base, fine-tuned, and RAG-enhanced models.
- Fine-tuned Gemma 2-9B with LoRA/QLoRA on a 26K synthetic Q&A dataset using Kaggle and Google

Colab, optimizing for response quality under constrained compute.
- Architected a RAG pipeline with PostgreSQL vector storage and MongoDB conversation memory for context-aware multi-turn interactions.
- Developed FastAPI services for model inference, retrieval orchestration, and frontend integration.
- Designed an LLM-as-judge evaluation framework on 100 held-out queries, with the fine-tuned model scoring higher than the RAG baseline on composite response quality (80 vs 71).

Accessible Reading Assistant
- Developed an assistive document reader for visually impaired users, enabling voice-controlled interaction, content summarization, and question answering using LLMs.
- Integrated speech, document processing, summarization, and Q&A components into an end-to-end assistive reading workflow.
- Used Pydantic schemas to validate and structure LLM outputs for reliable downstream processing.
- Engineered real-time multimodal voice features, including text-to-speech, speech-to-text, and visual processing pipelines, to improve STEM accessibility for visually impaired users.

Personal Technical Blog Platform
- Built and shipped a production personal blog/CMS using FastAPI, Jinja2, and SQLite with admin workflows for post editing, publishing, archiving, and media management.
- Implemented security controls including signed session cookies, CSRF protection, login throttling, IP-aware lockout handling behind Nginx reverse proxy, and hardened cookie/CSP/security headers.
- Designed a protected "Life Deck" media feature with gated access, per-IP rate limiting, attempt tracking, and admin security visibility.
- Developed analytics dashboards for time-based traffic views (24h/30d/12m), path-level insights, and interactive chart visualizations.
- Deployed on VPS with Docker + Nginx and automated CI/CD using GitHub Actions.

Technical Skills

Programming Languages: Python (primary), C++, JavaScript

AI / Machine Learning: PyTorch, Scikit-learn, NumPy, Pandas; LLMs: RAG, LoRA/QLoRA, prompt engineering, model evaluation, OpenAI, Azure OpenAI

Backend: FastAPI, Pydantic; Databases: PostgreSQL, MySQL, MongoDB, SQLite

Cloud & DevOps: Microsoft Azure (Azure AI Foundry, Blob Storage), Docker, Nginx, DigitalOcean, GitHub Actions CI/CD, Git, Gitflow

Frontend: React, Next.js, Tailwind CSS

Tools & Platforms: Kaggle, Google Colab

Awards and Competitive Achievements

- Recipient of three merit-based academic scholarships, awarded to top-performing students each term based on GPA.
- 1st Place, *Mastering IT* Academic Competition, HCMUTE (2025) – team-based CS problem-solving contest.
- 2nd Place, *Mastering IT* Academic Competition, HCMUTE (2024).
- 3rd Place among 90 teams, *Thach Thuc* Academic Challenge, University of Science (2025).
- 3rd Place among 101 teams, *Thach Thuc* Academic Challenge, University of Science (2024).
- 2nd Place, University Coding Challenge, HCMUS (2023).