

LÊ CHÍ ĐẠI

dai.lechi@hcmut.edu.vn | +84-363-259-454 | GitHub: caoTayTang | Portfolio: kaiosthefox.dpdns.org | HCMC

AI engineer with production experience deploying vision models with TensorRT and NVIDIA Triton on Jetson Orin and building evaluated LLM applications. Seeking to specialize in LLM inference serving.

EDUCATION

Ho Chi Minh City University of Technology | *B.S. Computer Science (Honors)* Expected 2027
GPA 3.9/4.0 | TOEIC 905 | Merit Scholarship, 5 consecutive semesters

EXPERIENCE

AI Engineer | *Rainscales – Real-time AI Detection* Jun 2026 – Present

- Deployed detectors as **FP16 TensorRT** engines (ONNX, trtexec) on **NVIDIA Triton** with dynamic batching.
- Run DeepStream on live RTSP feeds across 15 rig fleets, up to **30 concurrent RTSP streams per Jetson Orin**.
- Built a second-stage verifier: rule triggers send cropped multi-frame evidence to a self-hosted **Qwen3-VL**.
- Raised alert **precision 6.6% → 18.7%** with good recall on test set.
- Extended red-zone rules to 360° fisheye cameras, compensating for camera ego-motion.

AI Intern | *AKTech – Marker-free Multi-Camera Calibration* Apr 2026 – Jun 2026

- Co-developed a service that calibrates multi-camera rigs from a person’s T-pose, replacing checkerboards.
- Implemented pose-based RANSAC initialization and confidence-weighted Ceres bundle adjustment.

Research Assistant (part-time) | *ML4U Lab & AISIN* 2025 – Present

- Added camera-intrinsic embeddings to monocular 3D detection: KITTI Hard AP_{3D} 20.80 → 21.13 (MonoBEN).

PUBLICATIONS

Q. H. Nguyen, **C. D. Le**, et al. “MonoBEN: Monocular 3D Object Detection Beyond Fixed Camera iNtrinsics.” *RIVF 2026* (IEEE), accepted for presentation.

PROJECTS

CoinSight | *Multi-agent SRE copilot for a streaming market-data platform* [GitHub](#)
LangGraph · A2A · MCP · FastAPI · PostgreSQL · Kafka/Spark · Airflow · OpenTelemetry

- Built Analyst and Operations LangGraph agents that delegate over **A2A** behind an **MCP** gateway; the Operations agent triages incidents (lineage, Airflow, Jaeger, service health) with 7 runbook skills.
- Built a fault-injection eval (seeded outages, stale data, prompt injection) scored against SQL answer keys: **92.9% held-out** passes, every tool choice correct, all injections resisted.
- Kept hallucinated numbers at **≤1.1%** via a grounding validator that rejects figures absent from SQL or tool output.
- Benchmarked 5 models, including self-served **Qwen3-8B on Modal**, on pass rate, cost, and latency;

VoiceReaderAI | *NAVER Hackathon National Finalist, team of 4* [Demo](#)
Gemini · LangChain · Pinecone · CLOVA OCR & Voice · VieNeu-TTS · Android

- Led the architecture and AI backend of an Android reading assistant for ADHD users.
- Built the photo-to-answer pipeline: CLOVA OCR, Pinecone retrieval, and Gemini Q&A over the document.
- Added Vietnamese narration with voice cloning via VieNeu-TTS alongside CLOVA Voice.

SKILLS

Inference & deployment	NVIDIA Triton Inference Server, TensorRT, ONNX, FP16, DeepStream, Jetson
ML & vision	PyTorch, YOLO, RT-DETR, detector fine-tuning, Qwen3-VL, camera calibration
LLM apps & backend	LangGraph, FastAPI, eval harnesses, RAG, OpenTelemetry, Kafka, PostgreSQL
Languages & systems	Python, C/C++, SQL, Linux, Docker, GitHub Actions