

LE HOANG VIET

📍 Ho Chi Minh City, Vietnam · 📞 +(84) 36-3314-108 · ✉ lehoangviet2k@gmail.com
🐙 github.com/Mikyx-1 · 🌐 mikyx-1.github.io · 🎓 Google Scholar

EDUCATION

B.Eng. in Automation & Control Engineering

Ho Chi Minh University of Technology – Vietnam National University

📅 Aug. 2018 – Nov. 2023

📍 HCMC, Vietnam

WORK EXPERIENCE

AI Engineer

@Monorevo Inc.

📅 Apr. 2025 – Present

📍 HCMC, Vietnam

- Applied a VLM with Structured Output Prompting & Chain-of-Thought reasoning to engineering drawing analysis, replacing PaddleOCR and achieving a **35% accuracy improvement**.
- Fused MatMul-Activation layers in LLM serving pipeline, delivering **16% faster inference** at production scale.
- Implemented Rocchio algorithm for human-based feedback for 2D/3D similarity search.

Computer Vision Engineer

@Ftech Co. Ltd

📅 Nov. 2023 – Apr. 2025

📍 Da Nang, Vietnam

- Fine-tuned SegFormer-B0 with motion-aware augmentation (**+13% mIoU**); deployed as a C++ OBS plug-in.
- GPU-ported OpenCV's `guidedFilter` & `alpha-matting` in PyTorch, eliminating CPU bottlenecks on high-res video.
- Built a fault-tolerant distributed video-processing pipeline for VTV Cab, fully automating broadcast production.

AI Engineer

@Bosch Global Software Vietnam

📅 May 2023 – Nov. 2023

📍 HCMC, Vietnam

- Delivered driver-behaviour analysis system (gaze, drowsiness, occupancy) showcased at **ETAS 2023**.
- Knowledge-distilled YOLOv7-E6E → YOLOv5-S for Jetson Nano deployment with **only 8% mAP-0.5 loss**.

SELECTED PROJECTS

Reinforcement Learning Zoo [GitHub]

PyTorch, Gymnasium, MuJoCo, Atari/ALE, Weights & Biases

- Built an **RL benchmarking hub** of **9 algorithms** spanning MLP, CNN, and GNN policies (PPO, SAC, TD3, DDPG, DQN, A2C, REINFORCE, SARSA, GNN-PPO) under one unified interface.
- Verified results with seeded, reproducible evaluation, matching **paper-level scores** — **5512 ± 25** on MuJoCo **Humanoid-v5** and **100/100** on ALE/Boxing-v5.

Tiny LLM [GitHub]

PyTorch

- Built a **60M-param GPT-style LLM from scratch**—BPE tokenizer, transformer with hand-rolled KV cache, and the full training stack (pretraining → instruction SFT → reasoning fine-tuning) from scratch.
- Implemented tool-augmented reasoning via ChatML `<CALC>` tokens spliced with a safe AST calculator at decode time, plus bfloat16 mixed-precision **multi-GPU DDP** training and a stdlib-only token-streaming web chat.

torchinfo (Contributor) [GitHub]

PyTorch

- Implemented accurate parameter and computation counting for modern architectures (LLMs, VLMs, YOLO detectors), correcting counts for tied embeddings, weight-shared submodules, and parameters held on non-leaf layers.
- Fixed kernel-size handling for 0-d tensors and expanded supported inputs to dictionaries with non-tensor values, scalar tensors, and `torch.distributions` outputs

ACHIEVEMENTS

OpenAI Codex for Open Source grant (\$1320)

📅 Aug. 2026

Kaggle Expert

📅 Oct. 2024

High Achiever Scholarship

Awarded to the **top 1%** of students by GPA, academic year 2022–2023

📅 2023

SKILLS

Languages: Python, C/C++, CUDA, Rust, Bash **Frameworks:** PyTorch, TensorFlow, Keras

CV / NLP / RL: OpenCV, Ultralytics, Detectron2, Transformers, PEFT, Stable Baselines3, Gymnasium

Inference & Opt.: ONNX-Runtime, TensorRT, Triton, bitsandbytes, Numba

MLOps & Infra: Docker, Git, AWS, FastAPI, Kafka, GitHub Actions, W&B (+300 hours of training), Elasticsearch, Faiss

Languages: Vietnamese (Native), English (Full Professional — IELTS 8.0), Chinese (Intermediate - HSK4)