

Le Van Hoang

(+84) 368178627 | lehoang2797@gmail.com | github.com/henrylee97 | linkedin.com/in/henrylee-dev/

EDUCATION

Hanoi University of Science and Technology

Aug. 2016 – Jan. 2021

Bachelor of Engineering in Computer Science.

CPA: 3.4/4

EXPERIENCE

Senior AI Engineer

Apr 2025 – Present

Torilab

Ha Noi, Vietnam

- Architected and shipped a **multi-persona conversational AI platform** using **LangChain, LangGraph, and the OpenAI Agents SDK**, powering **10+ production avatars** across company products, such as companion and lifestyle coaching applications, serving **2K+ monthly active users**.
- Designed a **multi-layer long-term memory system** for persona agents, combining a **persistent user profile, episodic semantic recall, and rolling session summaries**. Enabled cross-session personalization and measurably increased average session length.

Senior AI Engineer

May 2022 – Apr 2025

Samsung SDS R&D Center

Ha Noi, Vietnam

- Led the end-to-end design of a **legal document intelligence platform**, architecting a multimodal pipeline to process diverse document types (PDFs, images) using vision-language models (VLMs). Deployed a secure on-premise RAG stack with a self-hosted LLM, reranker, and vector search, improving retrieval relevance and reducing processing latency.
- Built an **Intelligent Document Processing platform** to extract key information and tables from diverse documents. Utilized Key-Value Understanding, Table Recognition, and Document Reconstruction models with a robust Django, Celery, and RabbitMQ backend for scalable performance. Applied to digitalize data for several large companies, including a **top-3 bank in Vietnam**.
- Developed an **end-to-end MLOps pipeline** using Kubeflow on Kubernetes, **reducing model deployment time by 50%** and enabling automated training for 10+ production models.
- Developed a **Sales Boosting Tool** using Key-Value Understanding techniques and TensorRT inference, supporting **50K monthly transactions with 98% accuracy** and less than 1s per request, reducing processing time by 80% for finance operations.

Applied Scientist

Nov. 2020 – May. 2022

VinBrain

Ha Noi, Vietnam

- Built an advanced system for **reviewing medical device registration applications** for the Ministry of Health, utilizing document alignment, text detection, text recognition, and question-answering models. Enhanced deployment efficiency by optimizing model performance through knowledge distillation. **Achieved 95% accuracy and reduced processing time by 80%** compared to manual methods.
- Developed a **facial stroke alert system** utilizing face landmark detection and a classification model, **reached an F1 score of 90%**, enabling early intervention and treatment.
- Built an **eKYC system** for the AIViCare app, **achieved 95% accuracy** and reduced user verify time to under 5 minutes.

Computer Vision Engineer

Jul. 2019 - Nov. 2020

VCCorp

Ha Noi, Vietnam

- Contributed to the development of **content moderation features for the Lotus social network**:
 - * Built a **video recognition model using 3D CNN** for detecting violent actions with **more than 90% accuracy**. Used for filtering violent content in videos on the Lotus social network.
 - * Developed and maintained a **MobileNetV3-based model** for sensitive content image classification, **sustaining 98% accuracy** for user-uploaded content.

PROJECTS

AI Interior Design <i>Generative AI, Django, Celery, Redis</i>	Jan. 2024 – Jul. 2024
• Developed a website providing interior design ideas for users. • Created a generative pipeline for interior design utilizing SDXL, custom LORA, ControlNet, and IP-Adapter. • Built a backend with Django, serving a REST API using Celery and Redis for asynchronous tasks.	
AoDai LORA <i>LORA, Pytorch</i>	May. 2023 – Jun. 2023
• Finetune a Stable Diffusion model using LoRA technique on a large dataset containing thousands of images of traditional long Vietnamese ao dai	

TECHNICAL SKILLS

Languages: Python, C/C++, SQL
Frameworks: Pytorch, Tensorflow, Django, FastAPI
Libraries: OpenCV, Pandas, NumPy, Matplotlib, TensorRT, ONNX, Celery, RabbitMQ
Developer Tools: Git, Docker, VS Code

AWARDS

Math: Awarded First Prize in the Provincial Mathematics Competition for Grade 12

LANGUAGE

English: TOEIC 820