NGUYEN CAO QUOC

AI ENGINEER

+84 929 791 239 | caoquoc.work@gmail.com | imcaoquoc.github.io/nguyencaoquoc

Summary

Passion in crafting useful AI products with high performance and accuracy. Have 2 years in developing native AI applications, skilled in many modern AI technologies, experienced with concurrency programming, performance optimization. Also have experiences in researching and evaluating AI solutions. Currently pursuing a Master's degree in Computer Science for wider and deeper AI knowledge.

Skills

Core AI Concepts	Deep Learning, Machine Learning, Generative AI, RAG, Computer Vision, NLP
Languages	Python
Frameworks & Libs	PyTorch, TensorFlow, LangChain, LangGraph, LlamaIndex, FastAPI
Cloud & Tools	Azure Cloud, Docker, Model Context Protocol
Databases	PostgresQL, MongoDB, Milvus (Vector DB)

Professional Experience

TMA Solutions *AI Engineer* Jan 2025 – Present

TMA Agentic Platform

- Implemented **AI agents for healthcare** applications to deliver intelligent solutions that improved adaptability to industry needs.
- Took responsibility for developing a **dynamic language module** that enabled adaptive and context-aware for agent behavior.
- Engineered intelligent workflows on **ophthalmology** to enhance process automation and improve clinical decision support.

CadenceAI - Clinical Contracts Extraction and Invoice Matching

- Took responsibility for building the foundational architecture of the AI services.
- Leverage **Azure Cloud**, **LangGraph** and **Pydantic Agent** to create Contract Extraction Agent and Invoice Validation Agent.
- Ensuring the output format of Agents for embedding into **Milvus** database.
- Ensuring correctness and completeness of information before the invoice-matching step.
- Implemented concurrent AI Agent service flows, improve execution speed by **50%**.
- Designed and refined prompt instructions, and used **Pydantic** to enforce strict JSON output formats from LLMs.
- Leveraged **TOON** format in specific cases to reduce token usage.
- Collaborated with the tech lead to design the invoice–contract matching logic, improving overall correctness accuracy by **40%**.

CRM Intelligent Agent

- Developed and optimized the lead scoring module using AI models integrated with the **HubSpot API** through webhook events.
- Built AI-powered tools for generating automated quotations using **FastMCP**, reducing manual effort in sales processes.
- **Fine-tuned LLM prompts** for product-quantity matching, increasing system usability and reducing matching errors.
- Researched and deployed a **recommendation system** based on the **Collaborative Filtering** approach.
- Delivered insights and system enhancements to streamline sales processes and strengthen decision-making across business units.

Customer Service Agent

- Developed a **Customer Navigation Agent** using the **A* pathfinding algorithm** to guide users within a shopping mall environment.
- Built executable tools for a **Customer Service Agent** using **FastMCP** to enable automated task execution.
- Integrated the agent into the **TMA Agentic Platform** to deliver a proof-of-concept (**POC**) demo for clients.

VINAI RESEARCH *AI Engineer Intern* Sep 2024 – Jan 2025

Face Recognition System

- Researched and integrated **ArcFace Loss** into Deep Learning models to improve **Face Identification** performance.
- Evaluated models using **TAR@FAR** metrics, accuracy on **IJBB/IJBC** and compare against baselines.
- Analyzed **Cosine Scheduler**'s impact on learning rate adjustment and model convergence.

- Used **FastAPI** to develop interactive demos to show my solution's capabilities.

VIETVANG CONSULTING *AI Engineer* Jan 2024 – Sep 2024
BigC - Stock Keeping Unit Detection and Counting System

- Created and labeled image datasets using **Labelme**, ensured reliable training data for SKU detection and counting.
- Researched **YOLOv10** architecture, compiled detailed documentation to provide technical reference for project team.
- **Fine-tuned model** on custom datasets and evaluated performance. Increased model accuracy on real-world data.
- Developed interactive **POC** to showcase **YOLOv10** capabilities, helping customers understand AI applications and contexts.
- Used **FastAPI** to implemented APIs to support development teams in building customer-facing application.

Education

2024 – Present **Master of Computer Science**
University of Information Technology, VNU-HCM
2020 – 2024 **Bachelor of Computer Science**
University of Information Technology, VNU-HCM

Achievements

- IELTS 6.0 (English Certification) - View Certificate
- IBM AI Engineer (Coursera Professional Certificate) - View Certificate
- IBM Data Scientist (Coursera Professional Certificate) - View Certificate

Publications

Explainable Intelligence in Digital Twins (EIDT - Scopus-indexed conference)
Le Xuan Tung and **Nguyen Cao Quoc**: *Leveraging Large Language Models as Faithful Explainer for Text Classification* - EIDT 2025.
Link: Springer Lecture Notes in Electrical Engineering

Honors and Awards

FIRST PRIZE OF SOICT HACKATHON 2024 AI POWERED APPLICATION Nov 2024 – Jan 2025
Chatbot for processing administrative documents

- Utilized **Python** and **Llama-Cloud**'s API to process raw data into the required format, ensuring data consistency and quality.
- Leveraged **Llama-Index**'s API to convert data into markdown format and renamed files according to specified naming conventions.
- Employed **OpenAI**'s API to develop an API capable of generating **FAQ** pairs related to file content and file names.
- Conducted quality checks and evaluations of the generated FAQ pairs, **fine-tuning** them to ensure accuracy and relevance.