NGUYEN CAO QUOC

AI ENGINEER

+ (84) 929791239 | imcaoquoc.github.io/nguyencaoquoc | caoquoc.job@gmail.com

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.

CORE SKILLS

- Deep Learning
- Machine Learning
- Generative AI
- Retrieval-Augmented Generation
- Computer Vision
- Natural Language Processing

TECHNICAL SKILLS

- Pytorch
- Tensorflow
- LangChain
- LangGraph
- Llama-index
- PydanticAI
- Agno
- Model Context Protocol
- FastAPI
- Milvus
- Weaviate
- LightRAG
- FAISS
- Python
- PostgresQL
- MongoDB

PROFESSIONAL EXPERIENCE

TMA Solutions *Jan 2025 - Present*

AI Engineer

TMA Agentic Platform

- Designed and implemented AI agents tailored for healthcare and marketing applications to deliver domain-specific intelligent solutions that improved adaptability to industry needs.
- Took responsibility for developing a dynamic language module that enabled adaptive and context-aware agent behavior across diverse use cases.
- Engineered intelligent workflows for hospital use cases (ophthalmology) to enhance process automation and improve clinical decision support.
- Documented research findings and proposed improvements to enhance platform scalability and reliability through well-founded insights.

Clinic Contracts Extraction and Invoice Matching

- Took responsibility for building the foundational architecture of the AI services
- Took ownership of the Contract Extraction Agent, ensuring the output format is consistent and optimized for embedding into the vector database
- Took ownership of the Invoice Validation Agent, 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. Also 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/ERP Platform

- Developed and optimized the lead scoring module using AI models integrated with the HubSpot API, enabling more accurate assessment and prioritization of sales opportunities.
- Built AI-powered tools for generating automated quotations using FastMCP, reducing manual effort and accelerating response times in sales processes.
- Fine-tuned LLM prompts for product-quantity matching from natural language input, increasing system usability and reducing matching errors.
- Researched and deployed a recommendation system based on the Collaborative Filtering approach to deliver personalized product suggestions, boosting conversion rates.
- Delivered insights and system enhancements to streamline sales processes and strengthen decision-making across business units.

VINAI RESEARCH *Sep 2024 - Jan 2025*

AI Engineer Intern

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.
- Documented implementation, experiments and proposed enhancements.
- Analyzed Cosine Scheduler's impact on learning rate adjustment and model convergence.
- Integrated it into training pipelines and provided detail reports on accuracy and efficiency improvements.
- Used FastAPI to develop interactive demos to show my solution's capabilities.
- Created visualizations and presentations for both technical and non-technical audiences.
- Collected feedback and iteratively improved demo quality.

VIETVANG CONSULTING AND SOLUTIONS PROVIDING JOINT STOCK COMPANY *Jun 2023 - Sep 2024*

AI Engineer

Stock Keeping Unit Detection and Counting System

- Created and labeled image datasets using advanced tools, ensured reliable training data for SKU detection and counting
- Tailored data processing to meet customer and model requirements which improved compatibility and training efficiency
- Created and improved SKU evaluation criteria to standardized quality checks for consistent model assessment
- Researched YOLOv10 architecture, compiled detailed documentation to provide technical reference for project team
- Fine-tuned model on custom datasets and evaluated performance using precision and recall. Increased model accuracy on real-world data
- Refined model to improve accuracy to deliver better detection results for customers.
- Diagnosed, researched solutions and resolved customer's issues by collaborating with customer, project manager and CEO.
- Implemented solution to solve problem and enhance customer's satisfaction.
- Developed interactive demos to showcase YOLOv10 capabilities, helping customers understand AI applications and effectively communicating technical value in business contexts
- Designed and implemented APIs to support development teams in building customer-facing application.
- Developed comprehensive documentation and interactive API specs.
- Provided technical support and guidance to internal teams in developing backend and accelerating product deployment.

EDUCATION

UNIVERSITY OF INFORMATION TECHNOLOGY, VIETNAM NATIONAL UNIVERSITY, HCMC

2024-Present

- **Master of Computer Science**

UNIVERSITY OF INFORMATION TECHNOLOGY, VIETNAM NATIONAL UNIVERSITY, HCMC

2020-2024

- **Bachelor of Computer Science**
- GPA: 7.27

PUBLICATIONS

Explainable Intelligence in Digital Twins (EIDT - Scopus-indexed conference)

Leveraging Large Language Models as Faithful Explainers for Text Classification

Developed a prompt-based, model-agnostic framework that turns GPT-3.5/4o into post-hoc explainers for BERT classifiers on 20 Newsgroups and IMDB benchmarks. The system ranks influential tokens and generates natural-language rationales, improving trust and debugging for high-stakes NLP workloads.

Key Contributions

- Designed persona- and chain-of-thought-driven prompts that force LLMs to output top-k salient words plus faithful rationales.
- Ran faithfulness studies by deleting LLM-identified tokens to measure probability drops, benchmarking against LIME and SHAP.
- Showed GPT-4o rivaled or surpassed SHAP at higher k while providing human-readable narratives for classifier behavior.

Impact

- 20 Newsgroups accuracy: 75.08% (BERT) with GPT-4o explanations matching SHAP fidelity at $k \geq 7$.
- IMDB sentiment accuracy: 92.89%, where GPT-4o explanations consistently outperformed LIME and SHAP for $k > 1$.
- Earned EIDT recognition and certification for advancing transparent NLP workflows.

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.
- Handled table data within files to ensure critical content was preserved during processing.
- Employed OpenAI's API to develop an API capable of generating FAQ pairs related to file content and file names, automating and streamlining the process.
- Conducted quality checks and evaluations of the generated FAQ pairs, fine-tuning them to ensure accuracy and relevance.
- Collaborated with team members to design, build, and test the chatbot flow, ensuring seamless user interactions and functionality.

ACHIEVEMENTS

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