HIEU NGUYEN

AI Engineer | AI Researcher

LinkedIn
GitHub

Email: hieunguyen1053@outlook.com
Mobile: 091 132 0620
Address: Binh Hung Ward, HCMC

AI engineer and researcher focused on practical NLP and LLM systems for Vietnamese and enterprise workflows. I build production-ready models, retrieval systems, and AI tooling with a strong emphasis on accuracy, reliability, and measurable impact.

SKILLS

- **Languages:** Python, TypeScript, Java, SQL
- **AI / ML:** PyTorch, Transformers, LangChain, LangGraph, NumPy, Pandas
- **Backend / Infrastructure:** FastAPI, Django, Docker, Kubernetes, MySQL, Redis, Elasticsearch, vLLM, SGLang
- **Engineering Practices:** Specification-driven development, test-driven development, evaluation

WORK EXPERIENCE

AI Engineer | Private Company — Sep 2024 - Present

- Introduced on-premises LLM-assisted coding with Claude Code and Codex CLI across selected engineering workflows, reducing repetitive implementation and review work in internal development tasks.
- Built GitLab automation on a self-hosted stack to review merge requests, inspect legacy repositories, and generate follow-up issues or merge requests, shortening manual triage for recurring maintenance work.
- Developed LLM-powered Q&A agents for public-service platforms used by provincial agencies and the Ministry of Public Security of Vietnam, improving response consistency for domain-specific administrative queries across multiple knowledge sources.
- Designed a supervisor-and-subagent architecture for domain-specific request handling, reducing routing complexity and making new agent capabilities easier to add without changing the core workflow.

AI Researcher | NLP-KD Lab - TDTU — Jul 2021 - Sep 2024

- Trained Dama 2 7B from scratch for Vietnamese on the Llama 2 architecture and placed 2nd on the VLSP 2023 LLM benchmark.
- Developed Phi-3 Vietnamese and Mistral 7B Vietnamese variants to improve math reasoning, code generation, multitask performance, and structured outputs.
- Fine-tuned and evaluated large language models across instruction following, function calling, and JSON generation tasks.
- Worked with high-performance GPU infrastructure to train and iterate on large-scale language models efficiently.

AI Engineer | ADEMAX JSC — Sep 2021 - Aug 2024

- Owned the development and productionization of Vietnamese OCR, spell-checking, and document extraction systems from training and evaluation to deployment.
- Turned research prototypes into reliable inference services with optimized latency, memory usage, and throughput for real-world document workloads.
- Improved end-to-end document processing quality across OCR, text correction, and structured extraction pipelines used in production.

PROJECTS

Lumen **Mar 2026 - Present**

AI Engineer

Technologies: Electron, TypeScript, claude-agent-sdk, Tectonic, Zotero

Team size: 1

- Built Lumen as a personal, local-first Mac desktop app for AI-assisted scientific writing, keeping the workflow off the browser and closer to the user's files.
- Designed the app for agent-friendly desktop automation, so tools like Claude Code can work with documents, citations, and editing actions more directly than in web-based systems.
- Integrated Tectonic-based LaTeX compilation and Zotero for end-to-end drafting, citation management, and bibliography generation.

Legal AI **Sep 2024 - Present**

AI Researcher | AI Engineer

Technologies: Python, LangGraph, Neo4j, Amazon S3 Vectors, FastAPI, Next.js

Team size: 1

- Built a legal Q&A system around a knowledge graph that links articles, amendments, references, and regulatory documents, reducing manual legal research and document review time for end users.
- Fine-tuned Gemma 3 27B and gpt-oss-20B for legal-domain tasks, delivering 2-3x improvements on several VLegal-Bench subtasks.
- Served more than 20 paying users, including law students and legal lecturers, who used the system for legal research and question answering.

Ademax OCR **Sep 2021 - Jul 2024**

AI Developer | AI Engineer

Technologies: Python, PyTorch, Transformers, Vision Transformers, LangChain, OpenCV, FastAPI, Django, MySQL, MinIO, Redis, Elasticsearch, Prometheus, Grafana

Team size: 6

- Researched and trained a TrOCR-based OCR model from scratch for Vietnamese text, achieving an F1 score of 0.929 for error detection and 0.908 for error correction on the VSEC benchmark.
- Improved Character Error Rate (CER) by over 2% and Word Error Rate (WER) by over 9% compared with Tesseract and ABBYY.
- Built the production API for the OCR system with load balancing, dynamic batching, caching, monitoring, and 8-bit quantization, reducing inference time by 50% and memory usage by 4x while preserving 98% accuracy.
- Applied few-shot prompting and extraction guidance to convert documents into structured outputs, improving accuracy by 10% over previous encoder-decoder transformer models.

Ademax Spelling **Nov 2021 - Jul 2024**

AI Developer | AI Engineer

Technologies: Python, PyTorch, Transformers, FastAPI, Django, MySQL, MinIO, Redis, Prometheus, Grafana

Team size: 6

- Researched and trained a Transformer-based Vietnamese Error Correction (VEC) model from scratch, achieving an F1 score of 0.929 for error detection and 0.908 for error correction on the VSEC benchmark.
- Reduced spelling errors by 20% compared with previous solutions such as ViSpell and Google Docs, and deployed the model through a scalable API with load balancing, dynamic batching, caching, and post-training optimization.

EDUCATION

Ton Duc Thang University **Sep 2018 - Nov 2024**

Bachelor of Science in Computer Science, GPA 8.20

Completed the Computer Science program.

CERTIFICATIONS

TOEIC Certificate (IIG) | CERTIFICATE **Nov 2023**

- TOEIC 640

HONORS & AWARDS

TDTU Scholarship Recipient | AWARD **2019 - 2021**

- Received scholarships for the 2019-2020 and 2020-2021 academic years