HIEU NGUYEN

AI Engineer | AI Researcher

hieunguyen1053 | LinkedIn
hieunguyen1053 | github.com

Email: hieunguyen1053@outlook.com
Mobile: 091 132 0620
Address: Nha Be District, Ho Chi Minh City

AI Engineer and AI Researcher specializing in LLM agents, Vietnamese language models, OCR, and production AI systems. Experienced in training and deploying transformer-based models, hosting LLMs with vLLM and SGLang, building knowledge-enhanced QA workflows, and leading AI-assisted software development practices with specification-driven development and test-driven development.

SKILLS SUMMARY

- **Languages:** Python, Java, TypeScript, Rust, Swift, C++, SQL
- **AI / ML:** PyTorch, Transformers, LangChain, LangGraph, NumPy, Pandas, Matplotlib
- **Backend / Infra:** FastAPI, Django, Docker, Kubernetes, MySQL, Redis, Elasticsearch, MinIO, vLLM, SGLang
- **Development Workflow:** Specification-driven development, test-driven development

WORK EXPERIENCE

AI Engineer | Private Company — Sep 2024 - Now

- Led the adoption of AI-assisted coding at the company by proposing and guiding the implementation of Claude Code within the engineering workflow.
- Applied specification-driven development and test-driven development practices to accelerate implementation while keeping review, validation, and delivery quality under control.
- Built LLM-powered Q&A agents for public service platforms used by provincial administrative agencies and the Ministry of Public Security of Vietnam, improving accessibility through natural language interactions.
- Hosted and served LLM workloads with vLLM and SGLang to improve inference throughput and simplify production deployment.
- Architected a multi-agent workflow where a supervising agent coordinates domain-specific agents to handle complex requests with clearer delegation and maintainability.
- Designed a knowledge-enhanced QA system backed by an automatically constructed knowledge graph to support explainable search over political figures and related information.
- Deployed the full stack on-premise with local LLMs to meet data security and privacy requirements while avoiding dependency on third-party APIs.

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

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

WORK EXPERIENCE

AI Engineer | ADEMAX JSC **Sep 2021 - Aug 2024**

- Developed a TrOCR-based OCR system for English and Vietnamese text, improving accuracy over Tesseract and ABBYY OCR.
- Built a Transformer-based Vietnamese spell-checking model with strong detection and correction performance on the VSEC benchmark.
- Applied few-shot prompting and extraction guidance to convert unstructured documents into structured data formats.

EDUCATION

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

Bachelor of Science - Computer Science; GPA: 8.20

Completed the Computer Science program

PROJECTS

Legal AI | LINK **Sep 2024 - Now**

AI researcher & AI Engineer

- **Technologies:** Python, LangGraph, Neo4j, Amazon S3 Vectors, FastAPI, Next.js
- **Team size:** 1
- Designed and built a legal knowledge graph that links articles, amendments, references, and regulatory documents for explainable retrieval.
- Used Amazon S3 Vectors as the vector storage layer to support semantic retrieval and multi-step reasoning over legal documents.
- Combined symbolic retrieval with LLM-based reasoning to improve answer accuracy and transparency for legal question answering.

Lightsum | LINK **Sep 2023 - Dec 2023**

Freelance - AI Developer & AI Engineer

- **Technologies:** Python, PyTorch, Transformers, FastAPI
- **Team size:** 3
- Fine-tuned machine translation and summarization models for the technology domain in English and Vietnamese.
- Built the API layer for model serving with load balancing and dynamic batching to support reliable inference.

Ademax OCR | LINK **Sep 2021 - July 2024**

AI Developer & AI Engineer

- **Technologies:** Python, PyTorch, Transformers, Vision Transformers, LangChain, OpenCV, FastAPI, Django, MySQL, MinIO, Redis, Elasticsearch, Prometheus, Grafana
- **Team size:** 6
- Trained an OCR model from scratch with the TrOCR architecture, improving CER by 2% and WER by 9% over Tesseract and ABBYY.
- Deployed the model through a scalable API with load balancing, dynamic batching, caching, 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 prior transformer-based approaches.

PROJECTS

Ademax Spelling | LINK **Nov 2021 - July 2024**

AI Developer & AI Engineer

- **Technologies:** Python, PyTorch, Transformers, FastAPI, Django, MySQL, MinIO, Redis, Prometheus, Grafana
- **Team size:** 6
- Developed a Transformer-based Vietnamese error-correction model with strong detection and correction scores on the VSEC benchmark.
- Deployed the model through a scalable API with load balancing, dynamic batching, caching, and post-training optimization.

CERTIFICATIONS

Build a NLP solution with Azure AI Language (Microsoft) | CERTIFICATE **Jul 2024**

- Deploy a language resource, and use prebuilt models
- Create a custom text classification solution
- Create a custom named entity recognition (NER) solution

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

- TOEIC 640

HONORS & AWARDS

Awarded scholarships (TDTU) | AWARD **2019 - 2021**

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