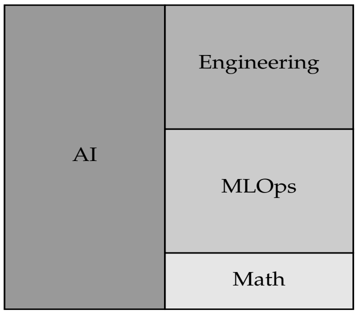
Che Viet Hai

cheviethai123@gmail.com
haicheviet.com
linkedin.com/in/hai-che
github.com/haicheviet



Education

Bachelor of Science | ***Major: Computer Science*** — Aug. 2015 – May 2019
University of Information Technology (GPA: 3.02/4.0) — Ho Chi Minh, Vietnam

Work Experience

Senior AI Engineer — Mar 2023 – Now
Koidra AI — Ho Chi Minh, Vietnam

- Support and implemented a physics-informed predictive model, leveraging environmental parameters and crop data to optimize multiple factors (ex: irrigation, lighting), leading to a 10% increase in overall crop yield.
- Spearheaded the adoption of monorepo architecture, consolidating codebases from various departments into a single repository, resulting in a 80% reduction time in CI-CD pipeline, enhance code stability and collaboration between multiple teams.
- Empower growers with custom reports and AI-driven strategies through a Koidra chatbot powered by LLM. Increases both engagement and delivers precision agriculture solutions directly to growers' needs.
- Expertise in PostgreSQL database optimization and self-hosting, resulting in a 5-fold reduction in backend API latency and linear scaling with time-series data requirements.
- Lead efforts to maintain a well-organized and up-to-date database of research experiments and results, remain code high quality over 80% coverage and facilitating faster feedback loops.

Lead AI Engineer — May 2019 – Feb 2023
Jobhopin Asia — Ho Chi Minh, Vietnam

- Build an AI recruitment platform and neural network search engine that optimize the recruitment process and engagement rate between job seekers and recruiters at scale.
- Communicated findings and progress to management and stakeholders through presentations and technical reports, highlighting the statistical outcome of the models.
- Success delivering a recruitment platform at SLA with SAP partners (Techcombank, TH group, ...) that generated significant revenue and was one of the main revenue streams at Jobhopin.
- Monitoring, A/B testing and secure AI platform in AWS, rate limit, and scale service as website traffic.
- Multi-tenant AI services, event-driven platform, APM distributed tracing, and separate app layer between computing and serving.
- Optimize AI model to low-level coding (Rust) and edge device inference (Web and Mobile).
- Provided guidance, code review, and mentorship to team members to ensure the quality and maintainability of the codebase, over 80% test coverage quality, full-flow CI-CD

AI Engineer — Apr 2018 – May 2019
TMA solution — Ho Chi Minh, Vietnam

- Work as an NLP developer to create intelligent software and provide solutions for NLP problems.
- Collaborate with Project Manager and Director to provide POC model and demo for potential problems from an existing customer.
- Associate with BA to perform data collection of 8 leading RPA companies in Southeast Asia.
- Contribute to building scraping system for Data Scientist team; built a system to scrap 500+ websites in 1 week.

AI Engineer — Dec 2018 – Feb 2022
Personal Team — Ho Chi Minh, Vietnam

- Work as pattern architecture for AI programs and provide insight for technological solutions.
- Engaging in product and technology research, design, and coding to develop new product-market fit.
- Work closely with many experts in our team to actualize the idea and prepare for the production step.
- Getting a robust accuracy with 97.27% on OCR and 99.23% overall performance in industry cameras.

Open-Source Contribution

Qdrant - High-performance, massive-scale Vector Database Github
Main contributor

Fullstack machine learning inference template Github
Author

Extensive and production-grade for a Python based Monorepo Github
Author

Academic and Research

Applying Transfer Learning in Stock Prediction Based on Financial News | ***Journal Paper*** Fall 2020
Seventh International Conference on Computational Science and Technology, ICCST2020 — First author

University Lab | ***Leader of team NLP*** Fall 2017
University of Information Technology

Skills

Coding: FastAPI, Django, Asyncio, Pytorch, Python, Rust
LLM: RAG, Self-Prompting, Function Calling, Fine-Tuning
Database: Neo4j cluster, Redis, Mysql, Elasticsearch, Qdrant vector database, TimescaleDB, Influxdb
Devops AI: Sagemarker, Opentelemetry, ECS-Fargate, Fluent-bit, CloudFormation, SQS-SNS-Lambda, VPC, TailScale, Monorepo
English: Toiec 845