




Tran Quoc Huy

Software Developer | 1 YoE

 github.com/huy-trn  +84378258645  tranquochuy645@gmail.com

SKILLS

Programming: JavaScript/TypeScript, Python, C, Java, Go

Web development: RESTful API, Websocket, Webhooks, NodeJS, ReactJS, HTML, CSS

LLMs integration: RAG, LLMs finetuning

CI/CD: Jenkins, Gitlab-CI, GitHub Actions

Cloud services: AWS S3, AWS EC2, Azure Databricks AI Inference

English: Equivalent of 915/990 Reading & Listening TOEIC (Non-official test at Ban Vien Company)

Others: Docker, Docker compose, Git, Linux shell scripting, RegEx

WORKING EXPERIENCE

Software Engineer at Ban Vien Co., Ltd

April 2024 - Feb 2025

- Developed internal tools and web applications integrating Large Language Models (LLMs), including RAG systems and fine-tuning.
- Worked on automated testing and deployment pipelines (CI/CD) for embedded software projects.
- Deployed and managed applications on on-premise infrastructure.
- English-speaking environment.

HIGHLIGHTED PROJECTS

Repomix - Tool for providing AI chatbot codebase context | Open source

#2 Contributor - Developer

References: [Npm package](#) | [Github](#) | repomix.com

Technologies: Typescript, NodeJS

- Developed a code parser to extract key information from the Abstract Syntax Tree (AST) using the tree-sitter library.
- Utilized the Vitest framework to implement unit tests (90% test coverage).
- Contributed innovative ideas and pull requests, driving the project's growth from 6k to 12k+ GitHub stars as of February 2025.

VSCode extension - Coding copilot | *Ban Vien & Renesas Design Vietnam (2024)*

Developer

Technologies: Typescript, Python, NodeJS, LangChain, ElasticSearch, Azure Databricks

- Designed and developed a Retrieval-Augmented Generation (RAG) system, including document preprocessing pipelines, document search using ElasticSearch (vector search, BM25), and LLM inference proxy.
- Customized VSCode extension [Continue](#) to utilize the RAG system.
- Trained a text classification model to categorize topics in user queries, improving the search engine's relevance and efficiency.
- The final chatbot achieved up to 90% accuracy on the Q&A task (AUTOSAR specifications and other domain-specific knowledge), with a response delay of less than 300ms before the first token.
- Fine-tuned the Qwen-2.5-coder-3B model for code autocompletion using a private codebase, ensuring compliance with internal coding guidelines for auto-generated code.

Gitlab bot - AI code reviewer | *Ban Vien & Renesas Design Vietnam (2024)*

Developer

Technologies: Python, FastAPI, Langchain

- Developed a custom fork of PR-Agent for internal use, integrating it with the RAG system to enhance domain knowledge and ensure compliance with internal coding guidelines.
- Implemented a codebase search engine (BM25, tree-sitter).

QI tool - Code/documents analysis tool | *Ban Vien & Renesas Design Vietnam (2024)*

Developer

Technologies: Go, Python, Jenkins, Gitlab-CI

- Developed features for automated document processing (PDF, Excel, Word, etc.).
- Monitored and ensured the reliability of automated jobs.

Video meeting and messaging web app | *Education project (2023)*

Fullstack developer

Reference: Github

Technologies: Typescript, NodeJS, ReactJS, MongoDB

- Implemented RESTful API endpoints for CRUD operations with MongoDB and user authentication with JWT.
- Gained hands-on experience with front-end optimization techniques, e.g., code bundling, code splitting, lazy loading.
- Utilized the video meeting and screen sharing features on the web browser through WebRTC.
- Used the Socket.io library and MongoDB events watcher for the implementation of real-time UI update.
- Implemented programs for user-generated media content transferring, access controlling with JWT, and AWS S3 pre-signed URLs
- Automated code-build and deployment processes with GitHub Actions, Docker, and Bash scripts.

Smart home with ChatGPT | *Education project (2023)*

Developer

Reference: Github

Technologies: Android (Java), ESP32-IDF, Firebase, ChatGPT

- Utilized ChatGPT's function calling feature to control edge devices, exploring the potential of LLMs like ChatGPT in its early days (2023).
- Developed an Android application to provide an interactive UI, specially an conversational-like controlling mode by the use of Google Speech.
- Implemented a wireless configuring method for configuration of WiFi credential for ESP32 board kit by utilizing the SmartConfig protocol.
- Worked with low-level C code for utilization of Firebase RESTful endpoints and real-time streaming data (Server push events handling).

EDUCATION

HCMC University of Technology and Education

Undergraduate

Computer Engineering