Ha Minh Quan – AI Engineer

Phone: +84 376316144 | Tan Binh District
github.com/quanmofii | linkedin.com/in/ha-minh-quan-b10717294 | haminhquan12c7@gmail.com

SKILLS & PROFICIENCIES

- **Languages:** Python, JavaScript, TypeScript, SQL.
- **Frameworks:** FastAPI, Flask, Django, Langchain, Autogen, Next.js, Nest.js, Node.js.
- **Libraries:** TensorFlow, Keras, Scikit-Learn, PyTorch, Spacy, PaddleOCR, OpenCV, BeautifulSoup, Playwright, Numpy, Pandas, Matplotlib, Seaborn, React.
- **Database:** ChromaDB, Qdrant, Weaviate, MySQL, PostgreSQL, MongoDB, SQLite, Redis.
- **AI Skills:**
  - Data Processing (Collection, Normalization, Cleaning, Visualization & Evaluation).
  - NLP Pipelines, Vector Embedding & Retrieval, LLM Integration, RAG, Agentic Chatbots, Supervised / Unsupervised Learning Training, Fine-tuning, Evaluation, Prompt Engineering.
  - Modeling (Classification, Regression, Clustering, Sequence Modeling, LLM).
  - ChatGPT, GPT-3.5/4, Transformers, Bert, LSTM, RNN, CNN, Yolo-OCR, Arima.
- **Principles & Architecture:** Database Schema Design (SQL/NoSQL/Vector), Monolithic / Layered / Modular Design, Clean Architecture, OOP, RESTful API, SOLID / DRY / KISS / YAGNI.
- **Other:** OpenAI API / Playground, HuggingFace, Git, Github, Gitlab, Docker, Postman.

WORK EXPERIENCE

C-UNIT SQUARE CO., LTD — **Tan Binh District, Ho Chi Minh City**
AI Engineer — *6/2024 – 2/2025*

- **C- Unit Chatbot: Enterprise Chatbot for Japanese Market** (LLM, RAG, Prompt Customization)

Developed a customizable enterprise chatbot powered by LLMs to support internal teams and clients, with a strong focus on tailored prompts and document-based RAG retrieval.
Tech: FastAPI, LangChain, OpenAI GPT, MySQL, ChromaDB, Next.js, React, TypeScript, Docker.
Role: Full-stack development, system architecture and AI workflow deployment.

  - Designed and implemented backend using FastAPI with secure authentication, chat history management, dynamic prompt storage, and real-time LLM streaming.
  - Built preprocessing pipeline for both uploaded documents and crawled web content: Format conversion, text extraction, cleaning, filtering, normalization, chunking, and metadata tagging.
  - Embedded processed content into ChromaDB using OpenAI embeddings and implemented LangChain-based retrieval workflows for contextual responses.
  - Integrated ChatGPT as the primary LLM, incorporating custom prompt templates, document retrieval via RAG, and chat history injection to ensure continuity and relevance in conversations.
  - Engineered business-specific prompts tailored to departmental workflows (e.g., finance, marketing, operations), enhancing chatbot relevance and contextual accuracy.
  - Implemented responsive chat interfaces using Next.js, React, and TypeScript, supporting real-time LLM streaming responses.

- **Store Management Assistant GPTs** (Custom GPTs, Google Apps Script, Prompt Customization)

Custom GPTs for business data querying and automated reporting.
Tech: Custom GPTs, Google Apps Script, Google API.
Role: Custom AI workflow deployment.

  - Developed a lightweight GPT assistant integrated with Google Sheets to automate sales reporting and data lookups via natural language queries.
  - Utilized Google Apps Script to retrieve and preprocess financial data.

TMA SOLUTIONS **District 12, Ho Chi Minh City**
AI Engineer Intern *11/2023 – 1/2024*

- **Prescription OCR & Post-processing with LLM** (Computer Vision + GPT + Prompt Engineering)

Built a system to extract structured data from prescription images using PaddleOCR and refined it using GPT-3.5-turbo with optimized prompts and fuzzy logic post-processing.
Tech: Fuzzy matching algorithms, PaddleOCR, OpenAI API, Streamlit, NumPy, multi-threading.
Role: OCR post-processing, LLM integration, prompt optimization

- Integrated PaddleOCR for prescription image text extraction and GPT-3.5 for structured formatting.
- Designed fuzzy logic rules to refine and standardize GPT outputs, improving reliability across edge cases.
- Optimized prompts and parallelized GPT API calls, reducing token usage and improving processing speed.
- Accuracy improved from 58% to 76.3%; processing latency reduced by ~9% under typical load.

PERSONAL PROJECTS

- **PAPERY: Academic Document Chatbot (FastAPI, Next.js, LLM, Agentic RAG, OCR)**

Team member: 3 *1/2025 – now*
AI-powered chatbot for querying complex academic documents (math, law, pharmacology) with multi-language translation and citation tracking.
Link: *https://github.com/quanmofii/papery*
Tech: FastAPI, LangChain, OpenAI GPT, BabelOCR, Qdrant, Next.js, PostgreSQL, Docker.
Role: Team Leader, Full-stack development, AI pipeline design, agent orchestration.

- Built an end-to-end AI chatbot system capable of querying and interacting with complex academic documents (math, legal, pharmaceutical, etc.), supporting multilingual translation while preserving original layout (tables, formulas, images, and domain-specific terms).
- Designed and implemented a document ingestion pipeline with OCR-based structure-aware extraction (YOLO-Template OCR, BabelOCR), chunking, embedding, and metadata tagging.
- Implemented Retrieval-Augmented Generation (RAG) pipeline with reranking, scoring, and source-based traceability (highlight page number and section of cited content).
- Integrated multiple LLMs, including OpenAI GPT (ChatGPT API) and open-source models to optimize performance across RAG tasks and domain-specific reasoning.
- Designed agent-based architecture using LangChain to support task decomposition, tool calling, and autonomous reasoning over documents.
- Developed responsive frontend with Next.js and Tailwind, including user auth, chat history, document upload/viewing, and streaming LLM responses.

- **More Projects at:** *github.com/quanmofii*
- **Personal Portfolio:** *quanmofii.github.io*

EDUCATION

NGUYEN TAT THANH UNIVERSITY **Ho Chi Minh City, Viet Nam**
Engineer's Degree in Artificial Intelligence *2020 – 2024*

ACHIEVEMENTS

NTTU Software Hackathon – Fourth Prize (Diabetic Retinopathy Image Prediction)
Role: Team Manager, Model development. *4/2023*
NTTU AI Hackathon – Encouragement Award (COVID-19 Prediction)
Role: Team leader *8/2022*