

PHAM PHU NGOC TRAI

Vinhomes Grand Park, Ho Chi Minh City, Viet Nam
+84 903 027 947 - phamphungoctrail@gmail.com
GitHub: jayllfpt - LinkedIn: Pham Phu Ngoc Trai

EDUCATION

FPT University, Ho Chi Minh City - GPA: 8.2/10

- Major: Artificial Intelligence (100% Scholarship)

EXPERIENCE

Senior AI Engineer

FPT Software

Jan 2022 - April 2025

Quy Nhon - Vietnam

Led the architecture and end-to-end development of **enterprise-grade chatbot systems** leveraging **LangChain** and Large Language Models (LLMs), enabling context-aware dialogue, multi-turn reasoning, and robust API/tool integrations for a range of business domains.

Designed, optimized, and deployed advanced NLP and document AI solutions, including multilingual document extraction, invoice understanding, and MRZ recognition, using SOTA models (YOLOv8/v10, Transformer-based OCR) in large-scale production environments.

Built and maintained **modular conversational pipelines**—intent detection, retrieval-augmented generation (RAG), tool-calling, and knowledge grounding—integrated tightly with web, API, and cloud-native infrastructures (AWS, GCP, Docker, FastAPI). Applied deep prompt engineering, retrieval strategies, and conversation memory to maximize chatbot accuracy, coherence, and user satisfaction in real-world settings.

Collaborated cross-functionally with Product, Data Engineering, and DevOps teams to design CI/CD, observability, and auto-scaling for reliable, secure, and cost-effective AI chatbot deployment.

Mentored and upskilled junior engineers in **LangChain orchestration**, LLM evaluation, cloud-native MLOps, and conversational AI best practices.

Delivered production-ready AI tools and chatbot solutions for industries including HR, enterprise automation, education, and digital transformation—demonstrating measurable gains in workflow efficiency and user engagement.

Published and maintained reusable AI components, including the **Table2HTML** Python package, adopted by internal teams and external clients for automated table extraction workflows.

PROJECTS

LLM-Driven Calendar Chatbot on Google Cloud Platform

- Architected and implemented a **multi-agent chatbot system** leveraging advanced Large Language Models (LLMs) to automate calendar scheduling and complex task management for enterprise use cases.
- Led the **end-to-end system lifecycle**: requirements analysis, scalable design, CI/CD deployment, monitoring, and optimization on Google Cloud Platform (GCP), ensuring high availability and security.
- Built **robust context-aware conversational pipelines** using LangChain, enabling accurate user intent classification, dynamic dialogue management, and autonomous information retrieval.
- **Integrated with Google Workspace APIs** (Calendar, Gmail, Drive), ensuring seamless data flow and secure OAuth2-based authentication.
- Established **best practices in prompt engineering** and LLM orchestration, optimizing latency, cost, and quality of responses for real-time automation.
- Applied **DevOps practices**: automated deployment with Cloud Functions and Infrastructure as Code (IaC) on GCP, with centralized logging and error tracking.
- Designed and executed **unit/integration tests** for both API and conversational flows, ensuring reliability and maintainability in production.
- Mentored junior engineers in advanced LangChain patterns, LLM evaluation, and secure cloud deployment.

Resume Ranking - JobFit

- Developed and optimized a **resume parsing and ranking pipeline** using LLMs via OpenAI API, automating extraction of structured data from unstructured CVs.
- Applied advanced **prompt engineering** and LangChain chains/tools to increase extraction accuracy and robustness across diverse CV formats.
- Automated tool chaining for **multi-step information retrieval and enrichment**, combining LLM output with classical NLP and rule-based filtering.
- Designed **RESTful APIs** and web services with Flask for scalable, modular deployment; packaged as Docker images for cloud-native infrastructure.
- Enhanced HR workflow by integrating with third-party HR systems, enabling **end-to-end screening and filtering** at scale.
- Ensured **data security and compliance** (GDPR, PII masking), and implemented logging/monitoring for real-time system health and feedback loop.
- **Benchmarked model performance** and conducted A/B tests for continuous improvement; maintained documentation and onboarding guides for cross-team adoption.
- Collaborated cross-functionally with HR, Data Engineering, and Product teams to iterate requirements, deploy to production, and monitor impact.

TECHNICAL SKILLS

Programming Languages	Python (4 years), C++ (4 years)
Frameworks/Libraries	LangChain, PyTorch, Flask, OpenCV, YOLO, Transformers, Hugging Face
Cloud Platforms	AWS, GCP (Google Cloud Functions, Vertex AI), Azure
Tools	Docker, Git, CI/CD (GitHub Actions)
AI/ML Techniques	Multi-Agent Systems, Prompt Engineering, Tokenization, Semantic Search

ACHIEVEMENTS

Research Competition

- **First Prize** Research Festival 2023 at FPTU Ho Chi Minh (2023)

Programming Contests

- **Silver Medal** Trai He Phuong Nam (2018)
- **Silver Medal** "30th April" Olympic contest (2019)
- **First Prize** Provincial Excellent Student Award (2020)

University Achievements

- **100% Scholarship** FPT University (2020 - 2024)
- Honorable Student status in 4 Semesters (Summer 2022, Fall 2022, Spring 2023, Summer 2023)

Certificates

- Microsoft Certified: Azure Fundamentals (2024)
- NVIDIA Getting Started with AI on Jetson Nano (2024)
- Coursera Big Data Certificate (2023)
- Coursera IBM Full Stack Software Developer Certificate (2023)

Other Achievements

- **Best Mentor Award** FPT University Scholarship Hunter Association (2022)
- Microsoft Office Specialist Microsoft PowerPoint (2019)