PHO HIEU ANH

AI Engineer · LLM Systems · Multi-Agent

discordhieuanh@gmail.com | 0335 511 504 | 41 Pho Vong, Hai Ba Trung, Ha Noi

ABOUT ME

Final-year Computer Science student with **hands-on production experience** building a complete, self-hosted AI ecosystem from the ground up. I designed and implemented every layer of a multi-agent system — from **LangGraph ReAct orchestration** and **Model Context Protocol (MCP)** microservices, to **Hybrid Search (BM25 + Vector)**, **GPU-accelerated OCR pipelines**, and **hallucination detection middleware** — with no reliance on managed cloud AI services.

My strongest competency is translating cutting-edge AI research into working, containerized, deployment-ready systems that run entirely on private infrastructure.

TECHNICAL SKILLS

Agent & Reasoning	LangGraph (ReAct Pattern), Model Context Protocol (MCP), Multi-step Tool Calling, Prompt Engineering (CoT, Few-shot), Loop Detection & Budget Management
LLM & Inference	HuggingFace Transformers, 4-bit Quantization (bitsandbytes/NF4), BitsAndBytesConfig, Mistral-7B, Llama-3, Qwen, bfloat16 mixed precision
RAG & Search	Hybrid Search (Elasticsearch BM25 + KNN Vector), ChromaDB / Pinecone, LangChain RecursiveTextSplitter, Sentence Transformers (Embeddings), Reranking
Computer Vision / OCR	GOT-OCR 2.0 (AutoModelForImageTextToText), PyMuPDF (fitz), pdf2image (Tesseract), PIL preprocessing (contrast/sharpness enhancement), Hallucination loop detection
Quality & Guardrails	BERTScore Hallucination Detection (F1 scoring), Semantic Filter (LLaMA-1B Inspector), Fuzzy Duplicate Detection (Unicode normalization)
Web Intelligence	Firecrawl (Playwright Headless), SearXNG, LLM-Ready Markdown Extraction, RabbitMQ message queue scraping
Backend & API	FastAPI, Async Python (httpx, asyncio), RESTful microservices, LangGraph state machines
Databases & Caching	MongoDB (chat history), Redis (multi-layer caching), Elasticsearch 8.x
Infra & DevOps	Docker, Docker Compose (multi-service), Nginx Reverse Proxy, NVIDIA GPU pinning, Arize Phoenix (observability)
Languages & OS	Python, Bash, Git, Linux (Ubuntu)

PROJECTS

■ AIchat — Self-Hosted Multi-Agent AI Ecosystem

Personal Project | 2024 – 2025 | Python, LangGraph, MCP, Docker, CUDA

Agent Orchestration (ReAct Loop)

- Implemented a **multi-node LangGraph StateGraph** driving a full ReAct loop with structured output parsing.
- Engineered **3-layer loop prevention**: tool budget, exact duplicate, and **fuzzy duplicate detection** (Unicode normalization + sorted word sets).
- Built **auto web-search intent detection** with 35+ keyword patterns enabling zero-configuration search routing.
- Designed **force-Final-Answer injection** and server-side query sanitization to protect internal metadata.

Model Context Protocol (MCP) Architecture

- Decoupled reasoning from tool execution via 3 independent **MCP services**: Local Data (:8011), Web Agent (:8012), OCR Vision (:8013).
- Orchestrator dynamically discovers tools at runtime via **/tools/list**, building a live routing map with graceful degradation.
- Enforced hard security blocks at the tool-node level to strictly control access to internet search.

Hybrid Search & RAG Pipeline

- Designed **Elasticsearch 8.x Hybrid Search**: dense vector KNN combined with BM25 text match in a single query.
- Built ingestion pipeline supporting digital extraction and **OCR fallback** for scanned documents with adaptive DPI selection.
- Stored data simultaneously in Pinecone and Elasticsearch with thread-level metadata filtering.

OCR Vision & Hallucination Guardrails

- Deployed **GOT-OCR 2.0** on GPU containers with **aggressive hallucination cleaners** filtering repetition loops and junk sequences.
- Integrated **BERTScore** middleware: scores Final Answers against observations, intercepting results below F1 < 0.15.
- Integrated **SearXNG + Firecrawl** for multi-threaded web scraping into clean Markdown with RabbitMQ queue management.

Infrastructure & Performance

- Optimized LLMs with **4-bit NF4 quantization** and bfloat16, reducing VRAM usage ≈75% while preserving quality.
- Deployed dual-model architecture: Mistral-7B for generation + LLaMA-3.2-1B as a semantic content filter.
- Orchestrated full stack with **Docker Compose**, Nginx proxy, GPU pinning, and Arize Phoenix for tracing.

■ Transformer Research & Advanced Prompt Engineering

Personal Research | 2023 | Python, PyTorch, Transformers

- Deep-dived into Transformer internals (Self-Attention, Multi-Head Attention) to understand failure modes.
- Benchmarked CoT and Few-shot prompting, applying findings directly to AIchat's ReAct system prompts.

WORK EXPERIENCE

Marketing Collaborator — SEC English Center (SEC)

2023 – 2024

• Executed digital marketing campaigns across social media, effectively increasing brand awareness and student enrollment.

EDUCATION

Bachelor of Computer Science — National Economics University (NEU)

2022 – 2026 | GPA: 3.3 / 4.0

ACHIEVEMENTS & LANGUAGES

• **2nd Place — AI Olympics 2025**, NEU College of Technology (NCT)
• **English**: Advanced (Technical reading/writing) | **Vietnamese**: Native

REFERENCES

Dr. Luu Minh Tuan — Vice Dean, Faculty of IT, National Economics University (NEU)
Email: Lmtneu@gmail.com | Phone: (+84) 904 143 460