Duc Tai Dinh

+84 344 643 850 | AI Engineer, HCMC | ductai.dt05@gmail.com | github.com/ductai05 | linkedin.com/in/ductai05

Backend Engineer & AI Researcher with strong foundations in NLP/Multimodal Retrieval and production-level system design. Experienced in building high-throughput microservices for ML model orchestration and co-authoring AI research. Seeking a AI/ML Engineer / MLOps role to bridge the gap between algorithmic research and scalable production.

Work Experience

Backend Engineer (AI Platform) Nov 2025 — Present
Acamar AI Vietnam Company Limited *Ho Chi Minh City, Vietnam*

- Engineered, maintained robust RESTful APIs in Go and Python and built resilient, event-driven data pipelines using Kafka, PostgreSQL, and Cassandra to ensure real-time data ingestion and synchronization for distributed AI services.
- Containerized of microservices using Docker and deployed them across multi-cloud environments (GCP, AWS, Azure), establishing efficient CI/CD workflows.

Publications & Competitions

KPTER: K-Pointer for Temporal Event Retrieval Jul 2024 — Nov 2025

- **Co-author** | Role: **AI Researcher**, **Leader**, Team WATLERE (AI Challenge HCMC 2025 & SOICT 2025)
- Won the **Best Presentation Award** at SOICT 2025 and ranked Top 12+/80+ in AI Challenge HCMC.
- Proposed and implemented the "K-pointer" sequential re-ranking algorithm and WRRF mechanism for precise temporal video segment alignment. Integrated semantic embeddings, OCR, ASR, and Open-vocabulary object detection into a comprehensive zero-shot retrieval framework.

ZSE-Cap: A Zero-Shot Ensemble for Image Retrieval and Prompt-Guided Captioning May 2025 — Jul 2025

- **First Author**: arXiv:2507.20564 | Role: **Lead R&D**, Team ITxTK9 (EVENTA Challenge @ ACM Multimedia 2025)
- Ranked **Top-4** on the private test set (mAP: 0.966, R@1: 0.955, Caption CLIPScore: 0.828) without task-specific fine-tuning on the 400K+ OpenEvents V1 dataset.
- Engineered a multimodal retrieval architecture extracting embeddings via CLIP, SigLIP, and DINOv2 with a custom L2-distance weighted ensemble. Developed advanced prompt-engineering strategies for Gemma 3, injecting retrieved contexts to generate highly accurate, event-aware captions.

Projects

AI Engineer, Vietnamese Legal GraphRAG (github.com/ductai05/VietLegalGraphRAG) Apr 2026 — Present

- Engineered an advanced **RAG** pipeline utilizing **LightRAG** to map complex Vietnamese legal text knowledge graph. Explaned legal jargon for user by integrating vietlegal-harrier embeddings and LLMs for role-adaptive QA.
- Implemented a contract analysis module leveraging **PaddleOCR** to extract and evaluate legal risks from document images. Deployed robust FastAPI microservices to deliver real-time insights accompanied by transparent legal citations.

AI Engineer, TradingClaw (github.com/ductai05/TradingClaw) Apr 2026 — Present

- Engineered an LLM-driven ecosystem featuring specialized **LangGraph agents** capable of real-time web search and personalized behavioral analysis. Bridged black-box predictions with actionable insights by implementing **XAI**.
- Deployed scalable **FastAPI** microservices integrated with **PostgreSQL** to manage high-frequency data streams, user states, and API cost metrics effectively.

AI/Data Engineer, Sino-Nom/Chinese OCR (github.com/ductai05/NLP-ChineseOCR) Nov 2025 — Dec 2025

- Engineered a high-concurrency data scraping pipeline using **Goroutines** to optimally crawl and curate the CWKB historical dataset, assembling a robust 3.4K+ image corpus alongside NomNaOCR.
- Fine-tuned **PP-OCRv5** detection architecture (PP-HGNetV2_B4 backbone) with DB Algorithm using PaddlePaddle GPU. Configured a distributed training pipeline (2x NVIDIA T4) utilizing Cosine Annealing and optimized DBLoss.
- Achieved a massive **30.2%** increase in H-mean score (from 0.731 baseline to 0.952) and **0.966 precision** evaluated on the NomNaOCR test set, substantially resolving detection failures for highly degraded Sino-Nom characters.

Education

University of Science, VNU-HCM Ho Chi Minh City
Bachelor of Science, Artificial Intelligence *Aug 2023 — May 2027*

- **Cumulative GPA**: 8.79/10.0 | Vu A Dinh Scholarship 2024 & 2025 | University - Local Government Scholarship.
- **Core Coursework**: DSA, OOP, Networks, Databases, OS, NLP, Machine Learning, Data Mining & Analysis, SE4AI.

Skills

- **Languages**: Python, Go, C/C++, SQL, TypeScript/JavaScript, R.
- **AI & ML Frameworks**: PyTorch, TensorFlow, Keras, Numpy, Pandas, scikit-learn, Unsloth, LAVIS.
- **Backend & Infrastructure**: FastAPI, NestJS, Docker, Kafka, Git, Linux (NixOS), CMake.
- **Databases**: PostgreSQL, MySQL, Supabase, Milvus, ElasticSearch, Cassandra.
- **Research Areas**: LLMs, VLMs, OCR, ASR, Object Detection, Multimodal Retrieval.

Certifications & Awards

- **Machine Learning & Math for ML Specializations:** Issued by Stanford University & DeepLearning.AI (Mar 2024).
- **Google Advanced Data Analytics Professional Certificate** (Mar 2024).
- **First Prize in Mathematics**: Provincial-level Excellent Students Contest Grade 12 (Mar 2023).
- **V-ACT 2023 (VNUHCM Academic Competency Test)**: Score: **1012/1200**, top 0.244%. (Mar 2023).
- **VSTEP level 4, English B2 CEFR**: Issued by University of Social Sciences and Humanities - VNUHCM (Sep 2023).