Praditya Wicaksono

Fullstack AI Engineer

Semarang, Indonesia • aditwicaksono34@gmail.com • praditya wicaksono.com
linkedin.com/in/praditya-wicaksono • github.com/aditwicaksonodinus

PROFESSIONAL SUMMARY

Computational modeling and systems thinking since 2020 — from numerical physics simulation to production AI engineering. Full-stack AI Engineer with a proven track record of building and deploying end-to-end intelligent systems. Experienced in integrating agentic workflows (MCP, LangChain), scalable RAG applications (pgvector, Elasticsearch), and LLM fine-tuning pipelines into production-grade environments. Focused on scalable microservices using FastAPI, robust MLOps practices, and supply chain AI security.

TECHNICAL SKILLS

AI/ML & Research: PyTorch, LLM Fine-Tuning (LoRA, Instruction Tuning), LangChain, LlamaIndex, RAG (pgvector, Elasticsearch), Semantic Search, MCP, OpenAI, Anthropic, GLM, Hugging Face, Evaluation (Ragas, TruLens), XGBoost, SHAP

Backend, Infra & MLOps: Python (FastAPI, Streamlit), Node.js, SQL, PostgreSQL, Docker, Kubernetes, GCP, AWS, Firebase, CI/CD (GitHub Actions), Git-based Model Tracking (DVC), MLOps, Supply Chain AI (NeMo Guardrails, API Security)

Frontend & Mobile: Flutter (Dart/Riverpod), Next.js (TypeScript, Vercel AI SDK), ES Modules (JavaScript), Tailwind CSS

Languages: Indonesian (Native), English (Professional)

PROJECTS

MAESTRO — Intelligent Tutoring System *Flutter, Riverpod, Machine Learning, Firestore, Groq/Gemini APIs*
Adaptive tutoring platform with reinforcement learning 2024 – Present

- Engineered an adaptive learning engine with a modular architecture that dynamically customizes student learning paths in real-time, improving personalized user sessions using reinforcement learning algorithms.
- Implemented a localized, high-performance RAG system utilizing in-memory indexing and LRU caching, significantly reducing LLM response latency and mitigating content hallucination.
- Successfully deployed the application to production, securing official software copyright (HAKI) for the proprietary architecture. Live: maestro-4dcce.web.app

JustExplain — Explainable AI Sentencing Predictor *Python, XGBoost, SHAP, MLflow*
Collaborative legal sentencing prediction with explainable AI (Lead: Jarot) 2025

- Co-developed an ensemble Machine Learning model (XGBoost & Random Forest) trained on 22k+ legal documents from the Indo-Law corpus, achieving 81% F1-score through advanced text embedding and target encoding pipelines.
- Integrated Explainable AI (XAI) using SHAP to provide transparent, interpretable visual breakdowns of model predictions for end-users.
- Deployed an interactive Streamlit dashboard for real-time case simulation, with MLflow tracking for reproducible experiment management.

LLM Fine-Tuning & MLOps Pipeline *Python, Hugging Face, LoRA, MLflow, DVC, Ragas*
Parameter-efficient LLM fine-tuning, automated prompt testing and evaluation 2025

- Fine-tuned LLM backbones (Hugging Face, GLM, LLaMA) via LoRA & instruction tuning, reducing trainable parameters by 95%.
- Built end-to-end MLOps pipelines with prompt testing, automated evaluation (Ragas, TruLens), and Git-based model tracking (DVC).
- Tracked 100+ fine-tuning and prompt engineering runs using MLflow to optimize model performance.

Lectura — Offline-First Presentation Engine *Node.js, SQLite, Reveal.js, Docker, Fly.io*
High-performance Single Page Application with offline database synchronization 2025 – 2026

- Developed a lightweight, 3-mode Single Page Application (SPA) utilizing a zero-dependency native Node.js server to minimize memory footprint.
- Designed a custom offline-first database synchronization system using SQLite, ensuring robust data persistence.
- Live: lectura-edu-demo.fly.dev

Scalable FastAPI AI Service & Gateway *Python, FastAPI, pgvector, Elasticsearch, Docker, Kubernetes, NeMo Guardrails*
High-performance agentic inference service, hybrid RAG database, and secure gateway 2026

- Engineered a scalable FastAPI AI service orchestrated on Kubernetes, ensuring high availability and low latency.
- Designed hybrid RAG retrieval combining pgvector and Elasticsearch for dense/sparse semantic queries.
- Deployed an AI security proxy with LLM guardrails (NeMo Guardrails) to mitigate injection vulnerabilities.

Automated Document Compilation Pipeline *GitHub Actions, XeLaTeX, MCP arXiv API*
CI/CD automation for document compilation and citation management 2025

- Designed and deployed an automated CI/CD pipeline using GitHub Actions to compile complex typesetting files and manage citation synchronization dynamically.
- Integrated Model Context Protocol (MCP) APIs to pull real-time metadata and BibTeX citations from scientific repositories (e.g., arXiv), automating citation assembly.
- Architected a modular XeLaTeX compilation system widely adopted by developers and researchers (published open-source on GitHub).
- Open-source: github.com/aditwicaksonodinus/udinus-thesis-template-latex

EDUCATION

Universitas Dian Nuswantoro *Semarang, Indonesia*
Master of Computer Science (M.Kom.), Informatics Engineering 2024 – 2026

Universitas Negeri Semarang *Semarang, Indonesia*
Bachelor of Education (S.Pd.), Physics 2018 – 2023