Kenny Valent Winalda Sembiring

AI / Machine Learning Engineer · LLM Infrastructure · Local-first AI

kennyvws1@gmail.com | +62 896-3707-2497 | github.com/KennyUMN | kennyvws-dev.vercel.app | Tangerang, Indonesia

EDUCATION

Universitas Multimedia Nusantara (UMN) 2024 – Est. Aug 2028

S1 Informatika GPA 3.53 / 4.00

- Relevant coursework: Machine Learning, Computer Networking, Object-Oriented Programming (Kotlin)

EXPERIENCE

Asset Management Division — Kelompok Studi Pasar Modal 2025 – Present

UMN Student Organization · Tangerang

- Analyzed Indonesian-listed equities (TUGU) for Stock Quest 2026 competition; built full three-statement financial model for TUGU with 2026–2028 projections
- Developed valuation frameworks incorporating PSAK 117 / IFRS 17 adoption impacts on insurance sector reporting
- Produced investment memos and pitch materials for peer review and competition submission

PROJECTS

Forge — Open-Source Agentic Coding CLI 2026 – Present

github.com/KennyUMN/forge · TypeScript / Node.js · Open Source

- Co-building an open-source agentic coding CLI kernel in TypeScript (Node 20+); derived the architecture from a comparative study of 10 existing agentic coding tools (opencode, Cline, Aider, OpenHands, SWE-agent, Goose) before writing an original codebase
- Shipped the kernel phase end to end: session manager, model-provider abstraction, agent loop with permission gate, plugin/tool registry with MCP client, and a CLI with session resume — 154/154 tests passing
- Added multi-provider model routing through OpenRouter (300+ models) behind a single provider interface
- Roadmap targets native multi-agent orchestration, a plugin ecosystem, and deep local dev-tool integration (git, LSP, sandboxing)

FinRAG-ID — Indonesian Financial Document Q&A 2025

github.com/KennyUMN/finrag-id

- Built a production-style local-first RAG pipeline on LlamaIndex + Qdrant + Ollama; targets Indonesian financial documents (laporan tahunan, prospektus)
- Implemented hybrid retrieval (BM25 with Sastrawi Indonesian stemmer + BGE-M3 dense embeddings) fused via Reciprocal Rank Fusion for superior recall on numeric financial data
- Runs fully offline via Ollama (Llama 3.1 8B Q5_K_M GGUF quantized), with OpenRouter as an automatic cloud fallback for higher-capability models
- Evaluated with RAGAS metrics (faithfulness, answer relevancy, context precision) + per-query latency tracking; containerized via Docker Compose for one-command deployment
- FastAPI backend with /query, /upload, /health endpoints; Streamlit UI with source citation display

Lembar Janji — Overseas Job Offer Compliance Engine Sep 2026

IFEST 2026 Hackathon (4th Place) · Next.js (App Router) + TypeScript · Fullstack

- Built a 3-tier deterministic legal compliance validation engine evaluating overseas job offer flyers against 10 statutory requirements under Law No. 18/2017 (UU PPMI)
- Engineered modular architecture decoupling vision-model extraction from core validation rules: offline P3MI agency matching, cost compliance rules, and zero-PII card rendering (399 Vitest tests passing)

PPE Detection System — YOLOv9c Semi-Supervised Learning 2025 – 2026

IF441 Group Project · Kaggle: kennyvws

- Trained YOLOv9c on a merged Kaggle PPE dataset; implemented a two-phase SSL pipeline: supervised baseline → pseudo-label generation → self-training
- Designed a Compliance State Machine using IoU-based Person-to-PPE mapping for structured safety event detection

– Debugged training pipeline issues (CLI vs Python API), managed multi-phase model checkpointing on Kaggle

Self-Hosted AI Homeserver 2026 – Present

Personal Infrastructure Project · Ubuntu 24.04 / CasaOS

– Running Ubuntu 24.04 + CasaOS with Dockerized services (Ollama, OpenClaw AI gateway, Tailscale VPN overlay)
– Configured OpenClaw to route Telegram/WhatsApp requests through OpenRouter (Llama, Gemini, Claude) with JSON-level config management via SSH
– Practical exposure to local LLM serving, container networking, bind-mount management, and API key rotation

Health Insurance Claims Forecasting — DSC MCF ITB Feb 2025

– Reduced MAPE from 14.13% to 6.46% using multiplicative seasonal decomposition to preserve monthly seasonality patterns
– Outperformed flat ensemble baseline by explicitly modeling claim volume cycles from 2024 historical data

TECHNICAL SKILLS

AI / LLM	Ollama, vLLM (familiar), LlamaIndex, LangChain (familiar), RAGAS, GGUF quantization, OpenRouter API, HuggingFace
ML / Data	PyTorch, YOLOv9c, Scikit-learn, Pandas, NumPy, Tableau, semi-supervised learning, seasonal decomposition
Infrastructure	Docker, Docker Compose, Ubuntu Server, CasaOS, Tailscale, SSH, Qdrant, ChromaDB
Languages	Python (primary), TypeScript, Kotlin, SQL, Bash
Other	FastAPI, Streamlit, Git/GitHub, REST API integration

COMPETITIONS & ACTIVITIES

– IFEST 2026 — 4th Place (Juara Harapan 1), National Hackathon by Universitas Padjadjaran (Lembar Janji)
– Stock Quest 2026 — Equity Research Competition (Indonesian capital markets, multi-sector analysis)
– DSC MCF ITB 2025 — Data Science Competition (health insurance claims forecasting, top MAPE improvement)
– Dinamic 2025 — Equity Research Competition (Indonesian capital markets, multi-sector analysis)