+6287822803050
South Jakarta, ID
yudhymccodey@gmail.com
linkedin.com/yudhy-prayitno

Yudhy McCodey

AI/ML Engineer | Applied NLP

yudhymccodey.xyz
github.com/yudhyy
yintelligence.space
x.com/mccodey

Recent Informatics Engineering graduate with hands-on experience developing and evaluating machine learning systems across applied NLP, Vision-Language Models, and LLM applications. Experienced with Python, PyTorch, Docker, CI/CD, Vertex AI, and ML experimentation tools.

TECHNICAL SKILLS

Proficient
Programming: Python and PyTorch
Inference & GenAI: vLLM, LiteLLM, LMCache, LangGraph, LangChain, Langfuse, Milvus
Backend & Data: FastAPI, PostgreSQL, Redis, Milvus, S3 (MinIO, Wasabi, RustFS)
Infra & MLOps: Docker, CircleCI, Google Cloud Vertex AI, MLflow, Weights & Biases, Git, Oxen

EXPERIENCE

Data Scientist (Freelance) — **Oct 2025 – Jan 2026**
Ismaya Group — *Remote*

- Engineered an agentic AI helpdesk with Redis caching and sliding-window rate limiting, reducing API calls by 90% while supporting 1,000+ F&B employees with sub-second response times.
- Integrated 3 MCP servers and 30+ retrieval tools into an AI chatbot, synchronizing menu data from Google Sheets with Redis across 30+ F&B brands.
- Developed an ACOS extraction pipeline covering Google Maps review collection, batched LLM-assisted annotation yielding 24k+ quadruples from 4,415 reviews, and multi-label BERT-CRF modeling; achieved 70.0 macro-F1, 93.9% aspect-opinion pairing accuracy, and 90.6% sentiment-category accuracy.

PROJECTS

Klaudia [link] — **Feb 2026 – Present**
Agentic Accountant Platform — *Open Source*

- Developed an open-source agentic accounting platform for spreadsheet-based financial workflows using FastAPI, LangGraph, PostgreSQL, Redis, S3-compatible storage, and MCP. Integrated Gemini via Google Cloud Vertex AI alongside OpenAI-compatible model endpoints.
- Built and deployed a fine-tuned Qwen 3.5-4B VLM for receipt/invoice key-information extraction using vLLM; benchmarked variants across six metrics, achieving 87.02 KIEVal F1, 93.99 ANLS*, and 94.58% digit accuracy on 1,000 human-verified images.
- Implemented a deterministic evaluation sandbox for tool use, guardrails, accounting tasks, and multi-step workflows.

Tunnel Engine [link] — **Jul 2026**
LLM Inference Gateway — *Open Source*

- Orchestrated a unified LLM gateway with LiteLLM and vLLM, exposing multiple models through an OpenAI-compatible endpoint with per-key rate limits, virtual keys, spend tracking, and priority scheduling.
- Designed multi-model inference routing with LMCache, S3-backed model streaming, and load-aware fallback strategies to improve availability during traffic spikes and model failures.
- Optimized long-context serving through KV-cache offloading across GPU, CPU memory, and storage, reporting an 8.4× improvement in time-to-first-token (TTFT). Added XGuard guardrails and Prometheus/Grafana observability.

EDUCATION

Bachelor of Informatics Engineering — **Sep 2022 – Aug 2026**
GPA: 3.86/4.00 — *Halu Oleo University*

Activities & Leadership: Bangkit Academy ML Cohort (Google, Gojek, Tokopedia, & Traveloka); First-authored Sinta-indexed paper on embedding model analysis for Medical RAG (4 citations); Educational volunteer across 3T regions, AIESEC local initiatives, and AI Content Researcher at College Copilot.