

Abelito Faleyrio Visese

Malang, East Java, Indonesia | +62 812-1601-7057 | afvisese@gmail.com | linkedin.com/in/abelito

PROFESSIONAL SUMMARY

AI Engineer specializing in LLM-powered systems, retrieval-augmented generation, and production-grade agent pipelines. Experienced in multi-step agentic workflows with tool/function calling, hybrid retrieval (dense + sparse + RRF), and evaluation pipelines measuring faithfulness and regression. Comfortable across Python RAG backends, TypeScript MCP servers, and AWS serverless infrastructure.

TECHNICAL SKILLS

- **LLM & Agent Engineering:** Claude API (tool-use, agent loops), AWS Bedrock (Strands Agents), RAG pipeline design, MCP server authoring, LangChain.
- **Retrieval & Evaluation:** FAISS, ChromaDB, BM25, Reciprocal Rank Fusion (RRF), Sentence Transformers, RAGAS, LLM-as-judge scoring, DeepEval.
- **AI & Machine Learning:** PyTorch, Computer Vision (YOLOv11, SAM 3, OpenCV), NLP (spaCy), Speech-to-Text (Wav2Vec2, Librosa), SVM.
- **Backend & Infrastructure:** Python (FastAPI, Django REST), TypeScript, Node.js, REST/GraphQL, Docker, AWS (EC2, Lambda, AppSync, Bedrock, CDK), PostgreSQL, DynamoDB, CI/CD.
- **Data:** Pandas, NumPy, Time-Series Analysis, Relational Schema Design.

WORK EXPERIENCE

KinetixPro

AI/ML Engineer Intern

Feb 2026 – Apr 2026

Singapore, Remote

- Architected an end-to-end Computer Vision Active Learning monorepo leveraging Docker Compose and the NVIDIA Container Toolkit to orchestrate 8 integrated microservices, encompassing YOLOv7 automated labeling, continuous training, and dataset QA pipelines.
- Engineered a high-throughput video sampling and recording engine using OpenCV and FFmpeg, directly extracting critical frames from MediaMTX RTSP live streams to build robust continuous evaluation loops.
- Deployed scalable GPU-accelerated API endpoints for real-time model inference and automated dataset synchronization to cloud storage, establishing resilient and strictly decoupled MLOps workflows.

Axrail

AWS Cloud Engineer Trainee

Jan 2026 – Mar 2026

Malaysia, Remote

- Built an **AI-powered conversational commerce agent** using AWS Bedrock (Nova Lite) and the Strands Agents framework, featuring **RAG retrieval** via Bedrock Knowledge Bases, **cross-session memory management**, real-time order trend analysis, and **MCP tool-use orchestration** with a 5-iteration agentic loop — enabling natural-language order placement over WebSocket.
- **Automated an enterprise employee timesheet management system** using AWS AppSync (GraphQL), 12 DynamoDB tables, and 5 EventBridge-scheduled Lambdas handling auto-provisioning, deadline reminders via SES, automatic submission enforcement, YTD chargeability tracking via DynamoDB Streams, and biweekly archival.
- Designed and deployed a multi-tenant serverless data platform using AWS CDK (Python) with 11 fully decoupled microstacks dynamically stitched via SSM Parameter Store — supporting OpenSearch full-text indexing, SQS FIFO queues, and Step Functions for third-party payment orchestration.
- Implemented production-grade security with Cognito RBAC, API Gateway authorizers, Bedrock Guardrails for content filtering, input sanitization, and per-user rate limiting — achieving zero critical findings in a third-party penetration test.
- **Adopted spec-driven development** using Kiro IDE to systematically define requirements, design documents, and implementation tasks before coding — reducing iteration cycles and ensuring traceability from specification to deployed feature.

Apple Developer Academy

Developer & Domain Expert Intern

Feb 2025 – Dec 2025

Surabaya, East Java

- **Talkative Pronunciation Coach (Voice AI & NLP Engine):** Architected a real-time ML inference backend using Django REST Framework, integrating PyTorch-based Wav2Vec2 models and Librosa DSP pipelines to perform phoneme-level pronunciation scoring on streaming audio. Achieved sub-2s latency per utterance by optimizing byte-stream decoding and dynamic time warping alignment across 44+ phoneme classes.
- **Cire (Edge-to-Cloud Computer Vision):** Architected a containerized edge-to-cloud CV pipeline decoupling YOLOv11 inference from the analytics warehouse, leveraging asynchronous HTTP workers with intelligent batching to reduce API calls by ~80% and sustain real-time object-detection throughput across 3 CCTV streams at 30 FPS with sub-200ms end-to-end latency.
- **GerakinAja (Real-Time Pose Analysis Engine):** Built an on-device ML inference pipeline processing 30 FPS video through Apple Vision and a custom CoreML action classifier trained on 60-frame pose sequences across 18 skeletal keypoints, with a 5-state finite automaton and 4 biomechanical constraint validators achieving sub-200ms per-frame latency for real-time repetition counting and form correction feedback.

SELECTED PROJECTS

Riset: AI Research Agent for Content Pipeline | *Python, Claude API, FAISS, BM25, MCP* *In Progress*

- Building a multi-step agentic system that researches AI/ML papers and news, evaluates what's worth covering, and drafts content angles for a short-form content channel — designed to demonstrate production-grade LLM engineering on a real, self-directed problem.
- RAG pipeline: arXiv API and AI/ML news RSS ingestion, chunked and embedded with open-source models (BAAI/bge-small or all-MiniLM-L6-v2), retrieved via hybrid dense (FAISS/Chroma) and BM25 sparse search fused with Reciprocal Rank Fusion (RRF).
- Agent harness: 5 custom tools (`search_arxiv`, `fetch_article`, `summarize_for_audience`, `score_relevance`, `draft_hook`) orchestrated through a hand-rolled multi-step loop via the Claude tool-use API, with per-step token/latency logging for cost analysis.
- Evaluation pipeline: a 10–15 example benchmark scored with RAGAS (faithfulness, context relevance) and a custom LLM-judge rubric, with before/after metrics across multiple prompt variants to catch quality regressions.
- MCP server (TypeScript): all 5 agent tools wrapped as a local MCP server and connected to Claude Code for live use — directly demonstrating the ability to build and extend custom agent tooling, not just consume it.

Manna Cooking Studio WhatsApp Chatbot | *TypeScript, Gemini LLM, RAG (ChromaDB), Google APIs, Baileys*

- Built an end-to-end AI customer service chatbot on WhatsApp handling 100% of inbound inquiries 24/7, automating class booking, payment validation, and scheduling workflows for a cooking studio — reducing manual admin workload by an estimated 80%.
- Implemented a RAG pipeline using Gemini embeddings and ChromaDB vector search to ground LLM responses in a curated knowledge base of 10+ documents, achieving contextual accuracy for class info, pricing, and policy questions.
- Engineered a stateful conversation engine with intent extraction, multi-step booking flow (class selection → scheduling → payment proof → admin verification), and automatic Google Sheets/Calendar sync for real-time inventory and appointment management.

Padel Court Analytics System | *Computer Vision, YOLO26, SAM 3, OpenCV, PyTorch*

- Engineered a real-time tracking vision pipeline integrating YOLO object detection with homography algorithms to algorithmically map spatial player movements from broadcast footage onto a scaled 2D minimap.
- Implemented robust multitracker solutions merging a high-speed YOLO ByteTrack module (using Hungarian spatial/HSV optimization) with a precise SAM 3 segmentation-based tracker designed to persist visual targets through heavy occlusions.
- Fine-tuned a custom pose estimation model using Roboflow datasets, and optimized global frame sampling utilizing structural similarity index measure (SSIM) logic.

Real-Time Traffic Congestion Detection System | *Computer Vision, YOLO11, OpenCV, PyTorch, SVM*

- Architected a scalable computer vision pipeline processing real-time CCTV streams, deploying parallel YOLO11 models for high-speed vehicle detection (90.8% mAP) and road segmentation.
- Engineered a robust feature extraction engine using Lucas-Kanade optical flow and spatial mapping to calculate road occupancy, vehicle density, and velocity, overcoming the limitations of traditional counting models.
- Developed and fine-tuned a Support Vector Machine (SVM) classifier achieving 97.4% accuracy and 14.4ms inference latency, delivering automated congestion monitoring to replace costly hardware sensors and mitigate urban traffic inefficiencies.

HONORS & AWARDS

Winner (1st Place), Hackfest 2025 | *Ciputra University* *Apr 2025*

- Engineered a custom Computer Vision pipeline enabling image-based semantic visual search, improving product discovery within a B2B marketplace.

4th Place, Jogjakarta Finance Case Competition | *Universitas Gadjah Mada*

May 2024

- Formulated complex financial restructuring valuations and predictive workforce projections modeling the strategic merger of two national telecommunication entities.

EDUCATION & CERTIFICATIONS

Amazon Web Services (AWS)

AWS Certified Developer – Associate

2026

DeepLearning.ai

Mathematics for ML, Machine Learning, TensorFlow Developer, TensorFlow Advanced Techniques, TensorFlow Data & Deployment *2023*

Petra Christian University

Bachelor of Computer Science, Data Science and Analytics

Surabaya, East Java

2021 – 2025

GPA: 3.61/4.00 | Cum Laude Graduate | Most Outstanding Freshmen in Data Science (2022)