

Michael Geraldin Wijaya

Jakarta, Indonesia | +62-895-6043-39001 | michaelwjy2@gmail.com

[LinkedIn](#) | [GitHub](#) | [Portfolio](#)

EXPERIENCE

AI Engineer

12/2025 – Present

AI Engineer - Intern

02/2025 – 11/2025

Maknadata, Jakarta, Indonesia

- Developing a real-time conversational AI voice agent for sales roleplay training, building a cascaded voice pipeline (ASR → LLM → TTS) from scratch.
- Implemented LLM tool calling to control each conversation turn, grounding responses in scenario facts through Qdrant vector retrieval (Vertex AI embeddings).
- Built a multimodal LLM-as-a-judge pipeline scoring recorded sales conversations against a versioned rubric.
- Researching and benchmarking proprietary and open-source speech/language models to evaluate cost-efficiency and self-hosting feasibility, reducing dependency on third-party APIs.
- Architected backends as layered DDD services with enforced module boundaries, and dispatched long-running evaluation jobs to Cloud Tasks workers on Cloud Run.
- Implemented logging and monitoring (observability) for a compliance audit backend service.
- Developed an automated system that merges draft reports and interview transcripts using LLMs, streamlining document integration and enhancing workflow efficiency.
- Built AI Agent (Agentic AI) system (RAG, text-to-SQL, etc.) with FastAPI and PostgreSQL backends, containerized using Docker.
- Analyzed business use cases to identify opportunities for AI-driven solutions.

Tutor Bina Nusantara University

Bina Nusantara University

09/2023 – 01/2024

- Conducted weekly study classes for Statistics and Linear Algebra, supporting students throughout one semester.
- Facilitated learning by providing explanations, answering questions, and offering guidance to ensure student success in these subjects.

Project-Based Virtual Intern: Data Scientist – ID/X Partners x Rakamin Academy

ID/X Partners, Jakarta, Indonesia

08/2023 – 09/2023

- Finishing various task related with the activity of Data Scientist from ID/X Partners, such as Big Data Fundamental, Statistics & Data Analytics, and Machine Learning Implementation.

EDUCATION

Bina Nusantara University

Master of Computer Science | 2025–Present

Streaming: Data Science

Latest GPA: 4.00

- National Finalist ASEAN Data Science Explorers 2024 Competition

- Top 20 Finalist School of Computer Science Hackaton 2024 BINUS x Microsoft (AI4A)

Bina Nusantara University

Bachelor of Computer Science | 2022–2025

Streaming: Intelligent System

GPA: 3.96 (Dean's List)

Achievements:

- Awardee Widia Partial Scholarship BINUS University
- Awardee of Beasiswa Unggulan Kemendikbudristek 2023
- Published paper: *Comparative Analysis of Machine Learning Algorithms and Data Balancing Techniques for Credit Card Fraud Detection*, presented at the International Conference on Computer Science and Computational Intelligence (ICCSOI), and published in *Procedia Computer Science* (Elsevier)

Senior High School 4 Palembang

Natural Science | 2019–2022, Grade: 97.29

Achievements:

- 1st Place Winner of the National Independent Science Competition - Mathematics
- A+ (Gold Medalist) in the Indonesian Science Competition 2022 - Mathematics
- A+ (Gold Medalist) in the Indonesian Science Competition 2022 - Technology and Information
- Gold Medalist in the 2022 National Numeracy Olympiad for High School Students
- Gold Medalist in the National Student Science Competition - Mathematics

PROJECTS

AI Voice Agent - Sales Roleplay

- Real-time voice agent that helps sales representatives practice with an AI customer persona, using a cascaded voice pipeline (ASR → LLM → TTS) built from scratch. Researching open-source vs. proprietary speech/language models to evaluate cost-efficiency and self-hosting feasibility.

Vision QC (Fiber Installation Audit)

- Vision LLM system auditing FTTH field installation photos against audit category standards for an Indonesian ISP. Achieved 83.6% overall accuracy and 85.5% defect recall on a benchmark of 286 photos. Benchmarked open and closed VLMs behind a modular provider layer (Qwen3-VL on self-hosted vLLM, Claude/Kimi K2.5/Nova via AWS Bedrock, Gemini via Vertex AI) and selected self-hosted Qwen3-VL-32B to meet data residency constraints. Added an OpenCV quality gate to filter unusable photos before any LLM call.

Report and Interview Integration – Automated System

- AI platform that turns interview recordings and a draft into a structured assessment report. It transcribes the audio, merges the transcripts into the draft section by section, and writes an analysis of the interview.

Alma - Agentic Conversational System

12/2024

- Conversational agentic system that enables users to interact with multiple data sources through natural language. The system integrates RAG (Retrieval-Augmented Generation) for document-based queries, text-to-SQL for database operations (e.g., checking product stock), and web search capabilities for real-time information retrieval.

Research Assistant Bot_GenAI_LLM

12/2024

- Research Assistant Bot (RAB) is a user-friendly research tool for retrieving information. Users can input article URLs or upload a PDF file and ask questions related to the content. It is built using Groq, LangChain, FAISS, Llama 3, and Streamlit.

Plant Disease Detection

10/2024 – 12/2024

- This plant disease detection system utilizes traditional ML methods like SVM and Random Forest with feature extraction techniques (Hu Moments, Haralick textures, and Color histograms) and pretrained CNNs such as ResNet and MobileNet. It includes a FastAPI backend with TensorFlow Serving. The model is deployed on Google Cloud Platform (GCP) using Cloud Functions and Cloud Storage, with an additional Gradio interface for easy access on Hugging Face Spaces.

Reachout Mail Generator_GenAI_LLM

11/2024

- Outreach Mail Generator is a tool that helps companies to generate outreach emails using Groq, LangChain, ChromaDB, Llama 3.3, and Streamlit/Gradio. It allows users to input the URL of a company's careers page, extract job postings from it, and create personalized emails. These emails include relevant portfolio links retrieved from a vector database, tailored to match specific job descriptions.

Personal Medical Chatbot with LLM RAG

06/2024 – 07/2024

- Personal medical chatbot project using Large Language Models (LLM) and Retrieval Augmented Generation (RAG). The chatbot provides detailed answers by leveraging medical knowledge from PDF documents, employing LlamaCPP for LLM integration and llama_index for efficient document handling. This approach enhances natural language responses and enables advanced retrieval of health-related information.

CERTIFICATES AND CERTIFICATIONS

- Building Transformer-Based Natural Language Processing Applications** | NVIDIA | 2024
- ASEAN Data Science Explorers 2024 National Final Indonesia** | ASEAN Foundation | 2024
- Machine Learning Specialization** | DeepLearning.AI | 2024
- Unsupervised Learning, Recommenders, Reinforcement Learning** | DeepLearning.AI | 2024
- Advanced Learning Algorithms** | DeepLearning.AI | 2024
- Supervised Machine Learning: Regression and Classification** | DeepLearning.AI | 2024
- Oral Presentation** | International Conference on Computer Science and Computational Intelligence (ISSCSI) | 2024

- Data Science and Machine Learning Fundamentals** | Corporate Finance Institute (CFI) | 2022
- Machine Learning with Python** | IBM | 2022
- Statistics Fundamentals** | Corporate Finance Institute (CFI) | 2022
- Data Analysis using Python** | IBM | 2022
- Python for Data Science** | IBM | 2022

SKILLS

- Programming:** Python, SQL, C, Java
- GenAI, LLM, RAG & Agents:** LangChain, LangGraph, CrewAI, LlamaIndex, Google ADK
- Vector Databases:** Qdrant, Milvus, ChromaDB, FAISS
- Models:** Gemini (Vertex AI), Claude (AWS Bedrock), GPT (OpenAI), DeepSeek, Llama, Qwen3-VL, Kimi K2.5, Nova

- **LLM Serving & Infrastructure:** vLLM, SGLang, Ollama, LiveKit, Pipecat, self-hosted GPU deployment, real-time voice pipelines (ASR → LLM → TTS)
- **Backend:** FastAPI, Flask, Celery
- **Databases:** MySQL, PostgreSQL (Neon, Supabase)
- **Machine Learning & Deep Learning:** ML algorithms and techniques, PyTorch, TensorFlow, Keras, NLP
- **Computer Vision:** OpenCV, scikit-image
- **Evaluation & Observability:** LLM-as-a-judge, RAG evaluation metrics, confusion matrix, RAGAS, MLflow, LLM tracing, LangSmith, Langfuse
- **Cloud Platforms:** GCP (Vertex AI, Cloud Run, Cloud Tasks, Cloud Storage, BigQuery, Cloud Functions), AWS (Bedrock, EC2)
- Docker, Git/GitHub/DagsHub, PySpark, Data Visualization

ORGANIZATIONS AND ACTIVITIES

- Data Science Club (DSC) and Ureeka BINUS University (2024 – Present)
- Badminton (2023 - Present) and Binusian Gaming (2022 – Present) BINUS University
- Tutor BINUS University (2023 – 2024)

VOLUNTEERING

- Anti Bullying Educator - Christian Eunike Orphanage Foundation, Semarang, Central Java (11/2022)
- Patriotism Educator - Bethlehem Salvation Army Orphanage, Semarang, Central Java (06/2023)