

MARK JOSH N. ALVEAR

Pasig City, Philippines • mjalvear44@gmail.com • +63 956 537 1842 • github.com/Meiji

EXPERIENCE

AI Developer | **Avvoka**

Sep 2026 – Present

London, UK

- Design and deploy autonomous AI agents that automate legal drafting and contract-management workflows within Avvoka's document automation platform.
- Build Retrieval-Augmented Generation (RAG) systems that ground AI outputs in contract precedents and institutional knowledge, improving drafting accuracy and reducing hallucinations.
- Develop end-to-end automation pipelines that streamline legal drafting, negotiation, and internal operational workflows, reducing manual effort for legal teams.

AI Software Engineer | Mind **You Mental Health Systems**

Feb 2026 – Present

Manila, PH

- Developed and deployed autonomous AI agents using LangGraph and Claude Code, applying rigorous software engineering principles to production-ready agentic architectures aligned with enterprise AI delivery standards.
- Architected end-to-end data pipelines on AWS EC2 (Ubuntu), replacing legacy Excel workflows with automated scraping and cleaning scripts — demonstrating cloud-native development and MLOps fundamentals.
- Built a specialized LLM system using a proprietary 13-category distress-detection framework, integrating transformer-based NLP with multi-layered ethical and security safeguards for sensitive AI use cases.
- Engineered real-time client dashboards and automated psychometric scoring systems with end-to-end encryption, delivering measurable improvements in data reliability and compliance.

AI Engineer – Software Engineer for AI Training | **Outlier (Contractual)**

Aug 2024 – April 2026

California, US (Remote)

- Quality-checked the LLM's coding output in Python and other languages for syntactic correctness and security, establishing rigorous evaluation frameworks for auditable, reliable, and testable AI systems.
- Spearheaded the testing and validation of specialized models, including Data Science LLMs and Vision LLMs, identifying inefficiencies and edge-case failures to optimize model reasoning for multi-step workflows.
- Collaborated with prompt engineering teams to monitor and refine complex LLM behavior, successfully reducing bad-prompt and error rates by 30% while establishing robust best practices for reducing hallucinations.
- Evaluated AI outputs for sensitive content compliance and factual accuracy, directly contributing to the platform's safety guardrails and ensuring transparency in agent decision-making.

Systems Engineer | **JFC Corporations (Hybrid)**

Sep 2024 – April 2026

Manila, PH

- Led Agile development of a custom HR dashboard to manage employee records, payroll processing, and workflows — replacing legacy spreadsheets with a scalable digital system.
- Deployed an NFC-based timekeeping portal integrating RFID readers with automated timestamp uploads, reducing payroll processing time by 40% and demonstrating full-stack, process-automation capabilities.
- Designed and implemented a building-wide mesh Wi-Fi network, ensuring secure and reliable connectivity across office and restaurant environments.

PROJECTS

Retinal Vessel Segmentation | **Best Thesis Award** | Python, PyTorch, Deep Learning, Graph Neural Networks, OpenCV

- Proposed a hybrid CNN-GNN architecture integrating a (DA-U)²Net backbone with a Graph Attention Network (GAT) for global topological reasoning, evaluated on DRIVE, CHASE_DB1, and HRF benchmark datasets for retinal vessel segmentation.
- Introduced a Dynamic Adaptive Feature Fusion (AFF) mechanism with a learnable spatial weight parameter (μ) to intelligently balance local CNN and global GNN features, achieving statistically significant IoU gains over static fusion ($\Delta+0.0296$, $p=0.000$) on DRIVE.
- Achieved state-of-the-art Specificity (0.975) and +17% Precision on CHASE_DB1 over the baseline VGN, effectively resolving vessel thickening artifacts while improving Sensitivity to 0.8573 on fine-vessel detection.
- Optimized graph construction using KD-Tree spatial indexing and Windowed Fast Marching Method, reducing complexity from $O(N^2)$ to near-linear $O(N \log N)$ for high-resolution fundus images.

- Validated results through rigorous experimental design on NVIDIA RTX 4070/4060 Ti hardware, applying CLAHE + gamma correction preprocessing, bootstrapped 95% confidence intervals (B=10,000), and an ablation study confirming Dynamic AFF outperforms static fusion across all metrics on DRIVE and CHASE_DB1.

SavageStats | **Offline-First Hybrid AI Fitness Platform** | Kotlin, Jetpack Compose, MediaPipe, Room, TFLite, ML Kit, CameraX, Health Connect

- Built an offline-first AI fitness tracker for Android using Kotlin, Jetpack Compose, and MVVM architecture, running a 528MB Gemma 3 1B INT4-quantized LLM entirely on-device via Google MediaPipe with zero cloud dependency, accounts, or telemetry.
- Engineered a hybrid AI routing system (AIEngineRouter) that detects hardware capabilities and routes between local Gemma 3 inference and a Gemini Nano path architecture-ready for flagship devices (Pixel 8/9, Galaxy S24/S25) via ML Kit GenAI Prompt API.
- Implemented an offline RAG pipeline pairing a custom TFLite food classifier (2,000 classes, ML Kit, CameraX) with a 411MB pre-packaged USDA nutrition database using Room + FTS4 full-text search, enabling camera-to-macro food logging with zero internet dependency.
- Designed a dual-database architecture separating mutable user data (DailyLogs, Missions, CustomFoods via Room v7) from a read-only USDA nutrition asset DB, preventing migration conflicts while supporting independent schema evolution across subsystems.
- Integrated Android Health Connect for real-time biometric sync (steps, sleep, calories) across Samsung Health, Strava, and Fitbit, and implemented a Mifflin-St Jeor TDEE/macro engine with an XP-based Savage Rank progression system.
- Implemented a coroutine-safe LLM inference pipeline using SupervisorJob + Dispatchers.Default with a five-state load machine (Uninitialized → Loading → Ready | NeedsDownload | Error), streaming token generation via MediaPipe, and an OkHttp foreground download service with 64KB-buffered atomic writes (.part → final rename) to prevent corrupt model states.

DataMuse | **AI-Powered Data Visualization & Storytelling Tool**

- Built an AI-powered data visualization and storytelling platform with a conversational analyst (Muse) that runs real queries on uploaded CSVs, generates dynamic charts, and explains findings in plain language — built with React 19, TypeScript, FastAPI, and Docker.
- Engineered a tiered LLM model router across 8 models from 3 providers (Groq, Cerebras, Gemini) with complexity-based routing, automatic escalation, 429 failover with TTL cooldowns, and thread-safe execution — supporting ~47K+ requests/day at \$0 infrastructure cost.
- Implemented a RAG pipeline using Qdrant vector DB and fastembed (ONNX) for grounded, context-aware data answers, alongside an AI story builder that generates chapter-based narratives with embedded visualizations across refinement and generation endpoints (/api/story/refine, /api/story/generate).
- Designed a zero-latency regex-based complexity classifier that routes each user message to one of three model tiers without an extra LLM call, with tier pinning (conversations only upgrade, never downgrade), per-model pinning for coherence, and automatic TPM-triggered context trimming with same-model retry.
- Engineered thread-safe failover logic handling 429 rate limits (60-second TTL cooldown per model), tool_use_failed recovery with model substitution, and automatic stripping of chain-of-thought <think> tags from reasoning model outputs to ensure clean user-facing responses.

BPI Sparks Academy | **Datawave 2025 AI Employee Growth Platform** | React, TypeScript, Next.js, Supabase, PostgreSQL, Gemini Pro *Aug 2025*

- Architected and delivered a full-stack, AI-driven employee growth ecosystem targeting attrition reduction and continuous skill development, built and deployed end-to-end within the hackathon window using AI-assisted tooling.
- Implemented a production-ready back-end with Supabase and PostgreSQL, enforcing row-level security (RLS), role-based access control (RBAC), and encrypted storage of user profiles, assessment records, and anonymised team data.
- Built an AI-powered coaching bot using the Gemini Pro API with a Retrieval-Augmented Generation (RAG) architecture backed by a pgvector store, delivering context-aware operational guidance with low hallucination rates.
- Developed interactive adaptive-learning interfaces and analytics dashboards in React, TypeScript, and shadcn/ui, with server-side rendering (SSR) via Next.js for fast initial load and SEO compliance.
- Conducted user-centred design sprints in Figma, validated workflows through iterative prototyping, and generated synthetic datasets for end-to-end performance and security testing.

PwestoGo | **AI-Powered Location Consulting Mobile App** | Flutter, FastAPI, Gemini, Supabase, PostgreSQL, MongoDB, Qdrant, Riverpod, LiteLLM

- Architected an AI-powered location consulting mobile app in Flutter with a FastAPI modular monolith backend, featuring a dual-model LLM gateway (primary Gemini + OpenRouter 503 fallback), SSE real-time streaming, and Riverpod state management.
- Designed an agentic backend with an Orchestrator delegating to 6 specialized sub-agents (LocationAnalysis, Comparator, Commute, News, Context/RAG, GeneralQuery) via an A2A protocol, with tier-based tool access control across 4 subscription tiers (Free → Enterprise).

- Built a 10+ tool registry with abstract base classes, normalized ToolResult contracts, and tier enforcement — integrating TomTom Traffic, Open-Meteo Flood, OSM Overpass, PAGASA, and OSRM APIs for real-time multi-domain location scoring.
- Implemented a tri-database architecture (PostgreSQL for relational data, MongoDB for tool results and audit logs, Qdrant for skill/context embeddings) with a Qdrant-backed RAG SkillLoader for context-aware prompt assembly and dynamic persona inheritance.
- Engineered a database-driven dynamic onboarding question tree (configurable without code changes), a Pugo AI persona with GO/WATCH/AVOID verdict scoring across flood, safety, and traffic dimensions, and a chat-like iterative feedback loop with multi-turn context accumulation.

Movie Search with Vectors | [Semantic Search Engine & RAG Tutorial Project](#) | Django, Qdrant, Sentence Transformers, Groq, Pandas, Matplotlib

- Built a full-stack semantic movie search engine using Django, Qdrant (local mode), and Sentence Transformers (all-MiniLM-L6-v2), converting ~4,800 TMDb movie descriptions into 384-dimensional vectors for meaning-based cosine similarity search — surfacing thematically relevant results that keyword search misses entirely.
- Implemented a complete RAG pipeline (Retrieve → Augment → Generate) integrating Qdrant vector search with Groq's Llama 3.3 70B, grounding AI responses in real database context to eliminate hallucinations on natural language movie queries.
- Designed a shared-core architecture (qdrant_utils.py) consumed identically by five Django management commands (CLI) and Django views (web UI), ensuring consistent behavior across both interfaces with genre filtering, min-rating constraints, and adjustable result limits.
- Built a vector visualization suite using Matplotlib and PCA to compress 384-dimensional embeddings into interpretable 2D charts — including genre clustering maps, query-to-result proximity plots, cross-genre similarity heatmaps, and raw embedding bar charts.

LEADERSHIP & ACTIVITIES

Executive Associate to the Director for Applied Research | [UST Technovation Society](#) Sep 2025 – Present

- Director & Project Head, InnovaTalks (Current): lead the society's flagship InnovaTalks initiative, directing the entire project lifecycle from strategic planning and conception through to successful execution and delivery.
- Co-founded the organization, playing a pivotal role in establishing its core mission, foundational structure, and long-term strategic direction.

Technology Development and External Affairs Staff | [UST Computer Science Society](#) Sep 2024 – Present

- Cultivated relationships with corporate partners and external stakeholders, securing sponsorships and collaborations for society initiatives.
- Managed and updated the official Computer Science Society (CSS) website, enhancing the organization's online presence and communication with members.

Logistics Staff | [UST CISCO](#) Sep 2024 – Jun 2025

- Collaborated with the procurement team to strategically source and manage essential resources, ensuring the successful execution of all organizational events.

Member | [UST AWS Learning Club](#) Oct 2025 – Present

- Actively participated in workshops, seminars, and networking events focused on Amazon Web Services (AWS), deepening knowledge in cloud computing and related technologies.
- Engaged with the AWS Club Philippines General Assembly to expand professional networks and track current industry developments.

EDUCATION

University of Santo Tomas Manila, PH
 B.S. Computer Science, Specializing in Data Science • Dean's Lister and Best Thesis • August 2022 - June 2026

CERTIFICATIONS

- **Machine Learning Specialization** — Stanford Online (Coursera) | [View Credential](#)
- **AI Agent Fundamentals** — KodeKloud | [View Credential](#)
- **AI Fundamentals: Foundations for Understanding AI** — IBM (Credly) | [View Credential](#)
- **Gemini Certified Educator** — Google | [View Credential](#)
- **Databases and SQL for Data Science with Python (with Honors)** — IBM (Coursera) | [View Credential](#)
- **Data Science Foundations** — Great Learning
- **Building RAG Solutions with Azure AI Studio** — LinkedIn Learning | [View Credential](#)

- **Responsive Web Design** — freeCodeCamp | [View Credential](#)

TECHNICAL SKILLS

Languages & Frameworks: Python, Kotlin, Java, TypeScript/React.js, Next.js, PHP, Django, R, SQL

AI & Machine Learning: LLM Frameworks (LangChain, LangGraph), Agentic AI, RAG Pipelines, Prompt Engineering, On-Device AI (Gemma 3, Gemini Nano), Google MediaPipe, ML Kit, TFLite, Gemini Pro API, Graph Attention Networks (GAT), Computer Vision, PyTorch, TensorFlow, Transformers, pgvector

Cloud & Infrastructure: AWS (EC2, Ubuntu Servers), Docker, Supabase, PostgreSQL, Qdrant, FastAPI, RESTful APIs, Git, MCP, Claude Code, SQLite, Room DB, Neo4j, Android Health Connect, CameraX, OkHttp

DevOps & MLOps: Data Pipeline Automation, End-to-End System Deployment, Agile Development, Technical Workflow Automation, Row-Level Security (RLS), Role-Based Access Control (RBAC), FTS4 Full-Text Search, Jetpack Compose, MVVM Architecture

Additional: Vector Databases (RAG/SQLite), Multimodal AI, System Design, Data Scraping & Cleaning, Analytical Skills