GILBERTO CARRILLO

AI Engineer & Team Lead | Agentic AI, LLM Platforms & Python Backend

Venezuela · Open to Remote | Spanish (Native) · English (Professional Working Proficiency)
gilbertojosecarrilloparra@gmail.com | linkedin.com/in/gilberto-carrillo | github.com/gilbertocarrillo

PROFILE

AI Engineer with 3+ years building production LLM systems at MAIDEPOT, a multi-tenant platform for AI workflows and agents used by enterprise clients such as LG, KOTRA and The Korea Times. Was the primary engineer behind its core backend and AI orchestration layer, and a hands-on Team Lead for 2.5 years. Looking for AI engineering roles building agentic products.

EXPERIENCE

MAIDEPOT — Multi-tenant platform for AI workflows & agents *Remote · South Korea-based company*

Backend Engineer & Team Lead — AI Platform *Jan 2024 – Aug 2026*

- Promoted to Team Lead within six months; led a cross-functional engineering team (frontend, full-stack and QA) while remaining the main backend contributor and owning architecture decisions.
- Built a model-agnostic LLM layer across 7 providers, with retries, rate-limit failover and load balancing, so features could switch models through configuration.
- Built the platform's multi-model chat on that layer, with SSE streaming, file attachments (documents, images, audio and video) and per-message credit accounting; it later became the foundation for the agent experience.
- Turned the workflow engine into a platform that enterprise clients such as KOTRA and The Korea Times built on themselves, by re-architecting it from a linear prompt queue into a JSON-defined DAG with parallel execution, LLM-evaluated conditional branches and map-reduce loops over documents, images, audio and video.
- Delivered two of the platform's most-used features: translation of documents up to 1,000 pages, built as a workflow on the engine, and transcription of long audio and video with Gemini or AssemblyAI. Both keep terminology consistent across chunks and resume automatically after failures.
- Extended the platform to PydanticAI agents: workflows became tools they can call, alongside ∼40 native tools, RAG knowledge bases (pgvector and Cohere rerank) and scheduled routines. Delivered client agents such as a sales-training assistant for LG.
- Built the public API (API tokens stored as hashes, rate limits per operation) and usage-based billing priced per model and token type, with revenue share for app creators.
- Owned the core FastAPI backend (multi-tenancy, Redis locks and rate limiting) and its releases on Azure App Service across dev, staging and production, using slot swaps for zero-downtime deployments.

Backend Engineer *Jul 2023 – Dec 2023*

- Migrated the platform's core backend, including its original text and image generation apps, from Google Cloud Functions to a dedicated FastAPI service, eliminating cold starts and consolidating the logic into a single, maintainable API.

SKILLS

AI & Agentic Systems: PydanticAI, MCP, Workflow Orchestration, Multi-provider LLM Integration, Tool Use / Function Calling, RAG, Embeddings & Reranking, pgvector, Prompt Engineering, Context Management

LLM Providers: OpenAI / Azure OpenAI, Anthropic Claude, Google Gemini (Vertex AI), DeepSeek, GLM, LG EXAONE

Document & Speech AI: Azure Document Intelligence (OCR), AssemblyAI

Backend & Cloud: Python, asyncio, FastAPI, PostgreSQL, Redis, Elasticsearch, Firebase, Azure (App Service, Functions, OpenAI, Blob & Queue Storage), GCP, Docker

RESEARCH & PUBLICATIONS

Reinforcement Learning for Estimating Student Proficiency in Math Word Problems — *2022 XVII Latin American Conference on Learning Technologies (LACLO), IEEE.* DOI: 10.1109/LACLO56648.2022.10013399

EDUCATION

Universidad de los Andes (Venezuela) — B.Sc. in Systems Engineering *2022*

Undergraduate thesis: *"Design of Human-Robot Interaction for Learning Multiplication Tables and Implementation of Speech Recognition Models"* — 2nd place, VII National Contest of Special Degree Projects, Venezuelan Computing Society (certificate).