Marc Castro

LEAD AI ENGINEER & TECH CONSULTANT · production LLM, RAG & agentic systems

Olongapo, Philippines · Remote (UTC+8) · Open to relocation: Singapore, Australia, EU MCastro.DSci@gmail.com
+63 919 902 5869 portfolio.marccastro.solutions linkedin.com/in/nipmcastro github.com/AstroMC98

SUMMARY

Forward-deployed AI engineer who ships production LLM systems inside client environments, not demos. 6+ years across insurance, genomics, contact centres, education and government. Leads end-to-end delivery of RAG platforms, agentic pipelines and document intelligence for a UK consultancy's enterprise clients, from scoping and architecture through audits, handover and training. An applied-physics background shows in the work: every pipeline ships with verification, tracing and a cost model.

CORE SKILLS

- **LLM systems:** RAG (semantic, vector, hybrid), HyDE-style query expansion, Cohere rerankers, vision models for image-heavy documents, structured output with Pydantic, knowledge graphs
- **Agents:** multi-agent pipelines (planner, extractor, builder, verifier), MCP servers, Claude Code skills and plugins, n8n workflows
- **Engineering:** Python, FastAPI, AsyncIO, WebSockets, Pub/Sub messaging, PostgreSQL, CosmosDB, ChromaDB, Docker
- **Cloud & LLMOps:** Azure (AI Search, AI Foundry, Container Apps, Entra ID), GCP (Vertex AI, Cloud Run, Pub/Sub), Databricks, LiteLLM cost controls, drift monitoring
- **Delivery:** discovery and scoping, production-readiness audits, per-run cost models, handovers, training non-technical teams

EXPERIENCE

Lead AI Engineer & Tech Consultant · Intelligent AI Solutions (UK, remote)

Sep 2024 – Present

UK consultancy building custom enterprise AI across insurance, media, real estate and professional services.

- Built a **4-phase agentic document pipeline** (layout analysis, adaptive extraction, HTML assembly, verification) on Vertex AI and Pub/Sub that turns complex insurance PDFs into semantic HTML at **10–30× throughput** through three levels of async parallelism, cutting turnaround from hours to minutes.
- Architected a **production RAG chat platform** on Azure Container Apps with switchable semantic, vector and hybrid retrieval, vision support for image-heavy documents, Entra ID role-based access and per-user monthly credit limits to cap spend.
- Built a **media monitoring pipeline** that analyses large article volumes for bias, with broadcast diarization, speaker attribution and a five-prompt LLM analysis pass, costed per run against Gemini Flash-Lite pricing.
- Owned delivery of an **insurance certificate processing platform** (Azure + Vertex AI): fixed a compliance-critical extraction defect, resolved Workload Identity Federation access across client tenants, and ran the handover.
- **Audited client AI systems** for production readiness, including a .NET/Semantic Kernel RAG stack and an n8n support chatbot. Surfaced gaps in identity propagation, circuit breaking and tracing, plus injection exposure and an exposed secret, each with a scoped remediation plan.
- Planned LLM cost management under Canadian data-residency rules for an insurance-sector client, and led internal research on retrieval, fine-tuning and LLM serving that feeds reusable components into client builds.

Co-founder & AI Lead · Stealth Startup

Concurrent

Data intelligence SaaS for SMEs, built on a semantic knowledge graph.

- Designed a three-layer AI operating system for non-technical business users (ingestion, graph construction, intelligence and retrieval), with the graph as a thin navigational index whose nodes hold locators, not payload.
- Built a browser voice agent with wake-word detection and a FastAPI WebSocket backend; wrote the technical whitepaper and a 68-ticket phased build backlog.

Global Data Science Manager · TDCX AI (Singapore)

Aug 2023 – Sep 2024

Newly launched AI consulting arm of Singapore-based TDCX.

- Built an LLM web app that generates realistic customer conversations for agent onboarding, cutting **time to proficiency by at least 20%**.
- Built a workforce-management forecasting model with a simulation layer, packaged in an interface operations teams used directly.

Data Scientist · AIA Digital+

Aug 2022 – Aug 2023

Technology arm of AIA, the largest listed life insurance group in Asia-Pacific.

- Built an end-to-end next-best-action pipeline for AIA Indonesia, cutting the **FTE needed for NBA model development by up to 95%**.
- Built a fraud detection solution to automate claim approvals, projected to cut processing time by at least 25%.

Senior Research Associate, Analytics & Visualization · Philippine Genome Center — Sep 2020 – Mar 2021

- Built data pipelines and analytical modules that shortened SARS-CoV-2 genomic surveillance reporting; co-authored papers on emerging variants.

CONCURRENT ROLES

AI Automation Instructor & Mentor · CIIT & Eskwelabs — Nov 2023 – Present

- Teach AI, no-code AI and data science to government agencies, corporate executives and mixed professional cohorts through workshops, bootcamps and hackathons.

Senior Data Science Consultant · PSHS Knowledge Innovation Division — Sep 2024 – Dec 2025

- Led data science on student performance management and data-driven curriculum enhancement for the Philippine Science High School system.

SELECTED PROJECTS

- **Customer service agent copilot:** FastAPI on Cloud Run with Pub/Sub, a Cohere reranker and a graph database; reads live conversations and returns structured JSON to flag journey stage, map resolution paths, score calls and forecast resolution. Cut average handle time and lifted customer satisfaction.
- **Automated media analysis:** scheduled Papermill notebooks on Databricks and Vertex AI with PostgreSQL storage and drift monitoring, replacing daily manual reporting across multiple analyst teams.
- **AI enablement curriculum:** a 5-module, 5–6 hour programme for commercial real estate firms covering AI fluency and confidentiality, Claude Cowork workflows (lease abstraction, rent rolls, offering memoranda, comps), Claude Code for non-developers, and governance.
- **Developer tooling:** Claude Code skills for multi-agent development planning and resumable execution, plus a 10-skill SaaS development plugin.

TRAININGS & TALKS

Jun 2025 – now	**Ayala Group companies**	Generative AI and low code for corporate teams
Aug, Nov 2025	**DICT ICT Literacy & Competency Development Bureau**	AI technology government trainings
Aug 2025	**Carousell Singapore (IT, marketing, HR)**	n8n automation for e-commerce
Mar 2025	**Asian Institute of Management**	Breaking into tech consulting within 5 years
Oct 2024	**DOH 20th National Health Research Forum for Action**	BANT.AI: fraud detection with unified health data
Jun 2023	**FEU Tech ACM**	Breaking bias in AI
Mar 2023	**UST Computer Science, HORAIZON**	The art of data storytelling and visualisation

RECOGNITION

- **Champion, DOH-DAP Hack4Health 2024** (100+ teams): all-in-one health fraud detection platform built in three days, later presented as BANT.AI.
- **Champion, ICSC Hack-A-Street 2025:** gravity-model tool that sites bike racks the way retailers site franchise stores, built in two days.
- **2nd place, UN Big Data Hackathon 2022, Youth Track** (123 teams worldwide); also Best Youth Team in Asia.
- **3rd place, USHER Hackathon 2022:** ML pipeline that reads early-warning signals of building damage.
- **Top 50 finalist, Google Cloud JAPAC AI Builder Cup 2026** (thousands of participants across 21 countries): Impromptu, live AI-hosted podcasts that listeners can call into mid-show.

PUBLICATIONS

- Detection and genome sequencing of SARS-CoV-2 B.1.1.7 in the Philippines. doi:10.1128/MRA.00219-21
- Emergent SARS-CoV-2 variant with multiple spike mutations, Central Visayas. doi:10.1101/2021.03.03.21252812
- A directed minimum connected dominating set for protein-protein interaction networks. ISBN 9780367814656

EDUCATION

- **MS Data Science**, Asian Institute of Management, 2022
- **BS Applied Physics (Instrumentation)**, University of the Philippines Diliman, 2020