Marco Jácome Jara

Quito, Ecuador
Móvil: +593 993 999 079 | Email: marco@marcojacome365.com | marcojacome10@gmail.com
linkedin.com/in/marco-jácome-jara-1abb79b3 | github.com/Klvo666
marcojacome365.com

Resumen Profesional

ML/AI Systems Architect con **10+ años** diseñando y desplegando pipelines end-to-end (datos → ML → producción) en entornos de alta criticidad (servicios financieros, cobranzas). Especialización en **sistemas RAG escalables**, **arquitecturas multi-agente**, y **mitigación de alucinaciones en LLMs** mediante validación multi-capa y grounding estricto. Trabajo publicado en **procesamiento de lenguaje natural para segmentación comportamental** (593 Digital Publisher CEIT, peer-reviewed). Expertise en traducir objetivos de negocio en soluciones de IA productivas con impacto cuantificable: +12–18% engagement, −15–22% costos operativos, time-to-insight de días a minutos. Enfoque en **sistemas verificables** (RLS, embeddings semánticos, búsqueda híbrida), **ingeniería de concurrencia** (STT multi-motor con fallback), y **ROI medible** en deployments a escala.

Educación

Pontificia Universidad Católica del Ecuador (PUCE) — Quito, Ecuador
Magíster en Marketing — 2022
Nota: 10/10

Escuela Politécnica Nacional (EPN) — Quito, Ecuador
Ingeniero Matemático — 2003
Mención Cum Laude

Habilidades Técnicas

Lenguajes: Python, TypeScript, SQL

LLM Systems & Architectures: RAG (hybrid search, re-ranking), DSPy, Multi-agent orchestration (CrewAI), Hallucination mitigation, LLM evaluation frameworks, Prompt optimization, Function calling & tool use

ML/AI Production: Fine-tuning (LoRA, QLoRA), Model compression & quantization, Propensity modeling, Scoring systems, PLN para segmentación comportamental (peer-reviewed publication)

Frameworks & Infrastructure: LlamaIndex, LangChain, Hugging Face ecosystem, FastAPI, React/Next.js, Supabase (RLS, pgvector), PostgreSQL, Redis, Docker, AWS

Embeddings & Vector Search: pgvector, ChromaDB, Cohere-embed-v3, Semantic chunking, Hybrid retrieval (BM25 + cosine)

Speech & Multimodal: STT pipeline engineering (Whisper, ElevenLabs, Gemini), Concurrent transcription with fallback, VLM integration, Audio preprocessing (FFmpeg)

MLOps & Production: Model versioning, A/B testing, Observability, Data quality monitoring, Concurrency management, Rate limiting strategies

Experiencia Profesional

CobroAndina Quito, Ecuador (Híbrido)
Cofundador & ML/AI Systems Architect enero 2023 – presente

- Diseñé e implementé plataforma end-to-end: **RAG system con mitigación de alucinaciones** (hybrid search + re-ranking), **scoring predictivo multi-dimensional** (gradient boosting + neural nets), y **STT pipeline concurrente** con adaptive fallback (Whisper/ElevenLabs/Gemini).
- Arquitectura en **AWS + Supabase**: PostgreSQL con RLS para multi-tenancy, pgvector para semantic search, Redis para caching/rate limiting, FastAPI backend, React/TS frontend, Docker containerization.
- **Impacto cuantificable:** +12–18% engagement | −15–22% operational costs | QA time 72h → <6h | reporting 1–2 días → <2 min | infraestructure costs −20–35%.
- Desarrollé **multi-agent systems** (CrewAI + Ollama): CSV Analyzer, Code Reviewer, Research Synthesizer, automatizando complex operational workflows con DSPy para prompt optimization.

Lasintec Ecuador
Gerente de Operaciones agosto 2022 – presente

- Dirección operativa integral, coordinación de equipos multidisciplinarios y mejora continua de niveles de servicio con enfoque en eficiencia operacional y control riguroso de KPIs.
- Implementación de cultura data-driven y optimización de procesos mediante análisis predictivo y dashboards en tiempo real.

Recover (Servicios Financieros) Quito, Ecuador
Head of Operations & ML/BI Systems noviembre 2018 – octubre 2022

- **Transformación data-driven:** implementé sistema integral de BI con modelos predictivos para propensión de pago, riesgo de incobrabilidad, y churn prediction, maximizando portfolio ROI.
- Liderazgo de estructura operativa escalable (6 jefaturas, 12 supervisores, 250 agentes), con aplicación de **ML para scoring y segmentación dinámica** de carteras.
- **KPI impact:** portfolio payback 17 → 9 meses mediante optimization de contact strategies con gradient boosting models y A/B testing sistemático de canales digitales vs tradicionales.

Gestiona GTX Ecuador
Subgerente de Operaciones mayo 2013 – septiembre 2018

- Operación integral de **4 campañas de servicio al cliente y 2 de cobranzas** (4 supervisores, 82 agentes), con recaudación anual superior a **USD $6 millones**.
- Administración de plataforma CTI (Computer Telephony Integration), seguimiento de KPIs financieros/servicio, atención de líneas 1700/1800 y liderazgo de proyectos multicanal.
- Optimización de procesos operativos mediante análisis de datos, reduciendo AHT (Average Handle Time) y aumentando FCR (First Call Resolution).

Proyectos Destacados

RAG System with Multi-Layer Hallucination Mitigation

Stack: Python, LlamaIndex, pgvector, Cohere-embed-v3, FastAPI, PostgreSQL (RLS), React/TS

- Arquitectura de **Retrieval-Augmented Generation** con sistema híbrido: búsqueda vectorial (cosine similarity) + BM25 lexical search + cross-encoder re-ranking para maximizar precision@k.
- **Mitigación de alucinaciones** mediante validación multi-capa: strict grounding (respuestas ancladas a chunks), citation cross-referencing automático, y confidence scoring con thresholds adaptativos.
- Sistema de **embeddings semánticos** con chunking estratégico (sentence-window, semantic boundaries) y metadata filtering para queries contextuales complejos.
- **Row-Level Security (RLS)** en PostgreSQL para multi-tenancy garantizando aislamiento de datos a nivel de base de datos, no aplicación.
- Reducción de latencia p99 de 8.3s a 1.2s mediante caching inteligente (Redis) y query optimization en vector search (HNSW indexes).

Concurrent STT Pipeline with Adaptive Fallback

Stack: Python (AsyncIO), FFmpeg, Whisper API, ElevenLabs Scribe, Gemini, Redis, PostgreSQL

- **Sistema de transcripción concurrente** con manejo inteligente de rate limits heterogéneos (Whisper: 50 req/min, ElevenLabs: 120 req/min, Gemini: sin límite) mediante token bucket algorithm y priority queuing.
- **Fallback automático multi-motor:** circuit breaker pattern para detectar degradación de servicio, failover transparente entre motores, y retry logic con exponential backoff, logrando 99.9% availability.
- Audio preprocessing pipeline con FFmpeg: normalización dBFS, noise reduction (spectral subtraction), VAD (Voice Activity Detection), y format optimization para maximizar WER (Word Error Rate).
- **Structured logging** (JSON) con full traceability: file_id, engine_id, timestamps, metadata, latency metrics, permitiendo análisis post-mortem y optimización continua del pipeline.
- Aplicación de **NLP downstream**: clasificación de intenciones (promesas de pago, objeciones, escalamientos), sentiment analysis, y auto-scoring de calidad mediante fine-tuned BERT models.

Hybrid Vector Search with Cross-Encoder Re-ranking

Stack: Python, LlamaIndex, Cohere-embed-v3, PostgreSQL (pgvector), FastAPI

- **Ingesta multi-formato** (PDF, DOCX, MP3/MP4 via STT, images via OCR) con semantic chunking strategies: sentence-window, sliding context, y boundary detection basado en topic coherence.
- **Hybrid retrieval**: vector search (cosine similarity sobre Cohere embeddings 1024-dim) + BM25 lexical search con score fusion (RRF - Reciprocal Rank Fusion), seguido de cross-encoder re-ranking para precision optimization.
- Sistema de **citation generation** con source tracking: cada respuesta LLM incluye hyperlinks precisos a chunks originales con metadata (doc_id, page, paragraph), enabling fact verification.
- **Query optimization:** HNSW indexes en pgvector para ANN search, metadata filtering pre-retrieval, y caching semántico (Redis) de embeddings frecuentes, logrando p95 latency < 800ms.

NLP for Behavioral Segmentation (Peer-Reviewed Publication)

Stack: Python, scikit-learn, BERT fine-tuning, PostgreSQL, Pandas

- **Publicación académica** en 593 Digital Publisher CEIT: metodología completa para segmentación de clientes en cobranzas usando procesamiento de lenguaje natural sobre interacciones textuales.
- Pipeline de **feature engineering**: TF-IDF sobre transcripciones, sentiment analysis, entity extraction (promesas, montos, fechas), y behavioral patterns (objeciones recurrentes, tone shifts).
- **Clustering multi-dimensional**: K-means sobre embeddings + hierarchical clustering para refinamiento, con silhouette score optimization y interpretación de segments via keyword extraction.

- Sistema desplegado en producción para **dynamic prioritization** de carteras, mejorando contact effectiveness y optimizando allocation de recursos humanos basado en segment complexity.

Multi-Agent Systems with CrewAI – Task Orchestration & Delegation

Stack: Python, CrewAI, Ollama (Llama 3.2, Mistral), LangChain, ChromaDB, DSPy

- Arquitectura de **agentes especializados** con roles definidos: Analyzer (datos tabulares, series temporales), Reviewer (code quality, security vulnerabilities), Synthesizer (research papers, technical docs), Analyst (invoices, contracts, legal docs).
- **Orquestación multi-agente** con CrewAI: task delegation basada en capabilities, inter-agent communication via shared memory, validation cruzada de outputs, y hierarchical planning para workflows complejos.
- Uso de **modelos locales (Ollama)** para data privacy en entornos empresariales, con quantization (4-bit) para optimización de recursos y latencia sub-200ms en inferencia.
- **DSPy integration** para optimización automática de prompts: compilación de signatures, few-shot example optimization, y chain-of-thought prompting adaptativo basado en task complexity.
- Trabajo con arquitecturas sobre datos estructurados (CSV, JSON), documentos (RAG), entradas multimodales (imágenes, videos), análisis de código y síntesis de investigación. Código y repositorios: perfil de GitHub.

Publicaciones Académicas

Jácome-Jara, M. & Cevallos-Campoverde, M. (2022). *Aplicación del procesamiento de lenguaje natural para segmentar clientes en una empresa de cobranza.* 593 Digital Publisher CEIT, 7(5-2), 99–113.
DOI: 10.33386/593dp.2022.5-2.1431
Revista científica ecuatoriana indexada en Latindex (ISSN 2588-0705), acceso abierto.

Certificaciones Seleccionadas

Docker	Codings Academy	diciembre 2024
AWS Web Services for Data Science	SDC LATAM	noviembre 2022
Certified Aheeva Integrator	Aheeva Technology	enero 2014
Machine Learning Specialization	Coursera (Stanford University)	2023
Deep Learning Specialization	Coursera (deeplearning.ai)	2023

Idiomas

Español: Nativo
Inglés: Profesional (lectura técnica avanzada, comunicación escrita/oral)

Enlaces y Recursos

LinkedIn: linkedin.com/in/marco-jácome-jara-1abb79b3
GitHub: github.com/Klvo666
Portafolio Web: marcojacome365.com
Código y repositorios: Perfil público de GitHub
Publicación Académica: doi.org/10.33386/593dp.2022.5-2.1431