

Jesús Martínez

AI Engineer | LLM Systems, Autonomous Agents, Production ML

jesus@datasmarts.net • Bogotá, Colombia (UTC-5) • [linkedin.com/in/jesusamartinezv](https://www.linkedin.com/in/jesusamartinezv) • github.com/jesus-a-martinez-v

SUMMARY

AI and back-end engineer with 14+ years of experience and 10+ years shipping production machine learning systems. I build LLM orchestration layers, autonomous multi-agent platforms, RAG architectures, and document intelligence pipelines, and I own them from architecture through production deployment. My AI work sits on a deep systems foundation: distributed architectures, monolith-to-microservices migrations, and data pipelines processing millions of records.

TECHNICAL SKILLS

AI and LLM	RAG architectures, autonomous multi-agent systems, prompt engineering, LLM orchestration and dynamic model routing (OpenRouter), open-source fine-tuning (Llama 2/3, Mistral, Falcon, PEFT), LangChain, LlamaIndex, DSPy, Langfuse, GPT-4, Claude, Gemini
ML and NLP	Production ML (10+ years), NER, text classification, summarization, document extraction, entity resolution, TensorFlow, Keras, BERT, Hugging Face, Scikit-learn, SpaCy, NLTK, SARIMAX forecasting, OpenCV
Languages	Python (12+ years), TypeScript / JavaScript, SQL, Go, Java (Spring Boot), Scala, Clojure
Back end and data	FastAPI, REST API design, PostgreSQL, Supabase, Pinecone, Redis, MongoDB, Elasticsearch, Neo4j, Kafka, PySpark, GraphQL, RabbitMQ, React
Infra and practice	AWS (SageMaker, Lambda, EC2), GCP, Azure, Docker, Render, CI/CD, Jenkins, n8n, pytest, TDD, strict mypy

EXPERIENCE

Lead AI Engineer / SpotFeed (via DataSmarts AI) Oct 2025 – Present

- Sole engineer on a political ad spending intelligence platform, owned end to end: architecture, LLM pipeline, backend, database, admin frontend, deployment, and the client relationship.
- Built LLM-first multimodal extraction over FCC contract PDFs whose layouts vary by station and change without notice, with model-per-stage routing through OpenRouter (Claude for classification and entity resolution, Gemini for line-item extraction).
- Processed 1,664 documents at a 98.8% extraction success rate into 52,768 queryable spend records covering roughly \$77M in tracked political TV ad spend; passed the client's 75% accuracy criterion at 80%+ over a two-week validation.
- Cut entity resolution from ~3 seconds to under 100ms with an alias cache that turns every human review decision into a permanent cache hit; low-confidence results route to review queues instead of being silently trusted or dropped.
- Shipped 2,500+ tests at 92% backend coverage and took the integration tier from 55 minutes to 22 seconds without deleting a test.

AI Solutions Architect / Crescendo Creative LLC Apr 2025 – Mar 2026

- Architected "Forge", a multi-agent platform of eight specialized AI assistants covering business planning, social data mining, and creative asset generation.
- Built a master-agent hierarchy over 10+ nested n8n workflows: the master agent parses high-level Slack requests and autonomously delegates sub-tasks across the agent network.
- Routed model selection dynamically through OpenRouter to balance cost, latency, and reasoning quality, and integrated Midjourney and Kling for automated image and video generation.

AI Solutions Architect / JIA NOMADS LIMITED Oct 2024 – Feb 2026

- Built a production AI email intelligence pipeline for 50+ users: multi-stage NLP on GPT-4 and Gemini 1.5 Pro classifying and summarizing 300+ emails daily and extracting tasks into Supabase at 80% accuracy.
- Designed multi-tenant conversational agents managing Google Calendar in natural language across Signal, Slack, WhatsApp, and Telegram for 20+ users, on a platform-agnostic abstraction layer normalizing incompatible channel APIs under strict per-channel data isolation.
- Led migration from Make to self-hosted n8n and pure Python: response times from 3 minutes to under 20 seconds, operating costs down 90% (\$500/mo to \$50/mo).

LLM Engineer / Cortado, Inc. Aug 2024 – Nov 2025

- Implemented Pinecone vector retrieval consolidating disparate property data, improving agent response accuracy and customer satisfaction.
- Migrated data ingestion from a monolith to distributed, event-based microservices, cutting agent response latency from minutes to under a minute.

AI Engineer / United Claims Specialists, LLC

Jul 2025 – Oct 2025

- Architected a company-wide RAG chatbot for internal policy questions, significantly reducing internal support tickets.
- Automated lead follow-up tracking for 12+ executives with n8n and AI validation: hundreds of leads per week processed in under 2 minutes instead of 5 hours.
- Built daily-refreshed telephony dashboards on Google Apps Script, n8n, and RingCentral data, replacing manual monthly reporting.

AI Engineer / Padrino's Restaurants Inc

Dec 2024 – Jul 2025

- Bilingual English and Spanish operational assistant across 5 locations resolving 20+ daily staff queries, cutting manager interruptions an estimated 75%.

Data Scientist / ML Engineer / ContractPod Technologies Limited

Dec 2023 – Jul 2024

- Fine-tuned Llama 3 and Mistral to match GPT-4 on legal document risk analysis, running in secure isolated environments to meet data privacy requirements.
- PEFT-based prompt and inference optimization delivering substantial cost savings with faster classification and summarization.

Senior Data Engineer / Komorebi Investment Pty Ltd

Mar 2024 – Jun 2024

- NL-to-SQL query agent on LangChain cutting safari planning time 20%, with schema enforcement and query validation guardrails over tens of tables.

AI / Machine Learning Engineer / Mahisoft

Dec 2020 – May 2023

- Rebuilt a legacy NER architecture on SpaCy and NLTK: 4x faster with improved extraction accuracy on legal documents.
- Built a Java and Spring Boot RESTful framework for training and deploying text clustering and classification models.
- Earlier stint as Senior Software Engineer (Aug 2017 to Apr 2020): multi-channel notification and authentication microservices in Go and Java adopted by four major production clients; re-engineered SQL for a multi-billion dollar client, cutting a critical process from 2 hours to 9 minutes.

Machine Learning Engineer / Blue Orange Digital

Nov 2019 – Dec 2022

- NLP topic tagging of pharmaceutical documents surpassing human reviewer accuracy by 10%.
- Quarterly pharma sales forecasting with SARIMAX and classical ML; computer vision preprocessing to rescue OCR quality on low-resolution scans.

Earlier / Wivo, Mesfix, S4N, freelance computer vision

2014 – 2019

- Computer vision (2019): zenithal RGB people counter deployed across retail venues; RGB-D height estimation for access control in 10+ locations.
- Wivo (2016 to 2017): led monolith-to-microservices transformation in Scala, Clojure, and GraphQL, halving maintenance time; Docker sandboxes cutting client environment spin-up from days to minutes.
- Mesfix (2015 to 2016) and S4N (2014 to 2015): led migration of a core financial platform to microservices, built the company's first automated CI/CD pipelines, and delivered stateful online onboarding for SURA that reduced a one-week manual process to about three hours.

EDUCATION**BSc in Computer Engineering**, Universidad Simón Bolívar, Caracas, Venezuela

2008 – 2013

Languages: Spanish (native), English (native level)