

Abdiel Rodríguez Lara

Agentic AI Engineer

abdielr10330@gmail.com | +1 249-680-0549 | Montréal, QC (relocating Aug 2026) | Open work permit

linkedin.com/in/abdiel-rodríguez-lara | github.com/Ericks0330 | ericks0330.github.io

Summary

Agentic AI Engineer with 2+ years shipping production applications, focused on building LLM-powered agents that plan, call tools, and execute multi-step workflows autonomously. Designs RAG pipelines, multi-agent orchestration, and API-driven backend integrations in Python and TypeScript, with guardrails, evals, and human-in-the-loop controls for reliable production behavior. Full-stack foundation across FastAPI, React (Next.js), and PostgreSQL, plus experience in cybersecurity and workflow-heavy enterprise systems. Relocating to Montréal in August 2026 on an open work permit.

Skills

AI & Agentic Engineering: Agent orchestration, autonomous planning, tool/function calling, multi-agent systems, RAG (embeddings, chunking, retrieval), structured outputs, guardrails and human-in-the-loop, prompt engineering, prompt-injection defense, LLM evals and observability (cost/latency/task-success tracing), token and cost optimization, voice and multimodal agents.

AI Stack: LangChain, LangGraph, MCP, AWS Bedrock (Knowledge Bases, Guardrails), Anthropic Claude API, OpenAI API, LangSmith, n8n, Vapi, Qdrant, Supabase/pgvector, PyTorch, TensorFlow, federated learning (FLEX/FLEXClash), adversarial ML, computer vision (YOLO, OpenCV).

Backend & Data: Python, TypeScript, Java, C, C++, FastAPI, Spring Boot, Nest.js, REST APIs, microservices, authentication workflows, PostgreSQL, PostGIS, Redis, Meilisearch, SQLite.

Frontend: React (Next.js), Angular, Vue.js, HTML5/CSS3, responsive design, WCAG 2.2 accessibility.

Delivery & Infrastructure: Docker, CI/CD (GitHub Actions), AWS (Bedrock, Lambda, S3, IAM), pytest, unit and integration testing, Git and code reviews, Agile/Scrum, Google Cloud (ADK, Colab), DigitalOcean, Render, Liquibase, functional and regression testing.

Experience

AIKoders — AI Engineer & Full Stack Developer (full-time) | Feb 2026 – Present | La Habana, Cuba

AI product studio delivering agentic systems for hospitality, logistics, and marketing clients.

- Built a three-agent bug-triage-to-pull-request pipeline (support → planner → coder) with a human approval gate before merge; 9 of the first 10 agent-generated PRs were merged after engineer review (90%).
- Shipped Guest Concierge (guest-concierge.ai), a production agent resolving guest inquiries over a pgvector/Supabase knowledge base and escalating critical cases to owners — owner-handled messages down from ~340 to ~155 per month across 12 properties (~54%).
- Cut token spend 26% and p95 latency from 11s to 7s on the production agent, by instrumenting LangSmith tracing for cost, latency, and task success per run, then iterating on retrieval, prompts, and model routing against the observed traces.
- Deployed voice and WhatsApp agents for EasyCargo (multi-country LatAm shipping) with Vapi, answering customs questions from a vector knowledge base and pulling live shipment status via API tool calls — 72% of ~1,200 monthly inquiries resolved without human handoff.
- Built EasyPosting, a multimodal publishing agent that enhances customer photos via Cloudinary behind a paid-plan entitlement check, writes brand-aligned captions from retrieved client context, and auto-publishes to Instagram and Facebook.

Technologies: Python, TypeScript, FastAPI, React (Next.js), Anthropic Claude API, LangSmith, n8n, Vapi, RAG, pgvector, Supabase, structured outputs, Cloudinary, PostgreSQL, PostGIS, Meilisearch, Docker, DigitalOcean.

Guajiritos — AI Engineer, RAG & Automation (contract, 2 months) | Jan 2026 – Feb 2026 | La Habana, Cuba

- Raised LLM answer accuracy on proprietary business data from 52% to 87% (+35 pp) on a 500-question internal benchmark, by redesigning chunking and embedding strategy over vector databases and enforcing structured outputs for reliable downstream integration.
- Built an eval and tracing harness (LangSmith) scoring cost, latency, and task success per run, cutting cost per query from \$0.045 to \$0.036 (~20%) through retrieval and model-routing changes.
- Automated 5 internal business processes with n8n workflows and MCP-based tool orchestration, connecting LLMs to APIs, vector databases, and backend services through function calling.

Technologies: Python, n8n, LangChain, vector databases, REST APIs, RAG architectures, multi-agent systems, MCP, OpenAI API.

INTRUST Security — Full Stack Developer (contract) | Nov 2025 – Jan 2026 | La Habana, Cuba

- Shipped a vulnerability-management platform (Spring Boot + Angular), consolidating 2 security scanning tools into a unified real-time monitoring dashboard for centralized risk analysis.
- Built automated domain-scanning tooling that increased security-assessment throughput, and managed versioned PostgreSQL migrations with Liquibase for repeatable releases.

Technologies: Java, Spring Boot, Angular, TypeScript, PostgreSQL, Liquibase, REST APIs.

Centro de Ingeniería Genética y Biotecnología (CIGB) — Bioinformatics Software Engineer (freelance, ~10 h/week) | Sep 2025 – Present | La Habana, Cuba

- Designed a quantitative scoring algorithm producing normalized confidence scores for Dengue-LRP1 protein-protein interactions, weighting evidence type, interaction frequency, and dataset characteristics into a reproducible framework.
- Built and deployed a production web analytics platform with advanced search by protein name, gene symbol, UniProt ID, and cell type, integrating the scoring engine into the application's search core.

Technologies: Angular 21, TypeScript, Python, bioinformatics data processing, algorithm design, REST APIs. Live: data-base-lrp1-web-app.onrender.com | github.com/Erick0330/Score-algorithm | github.com/Erick0330/data_base_lrp1_web_app

A11ySolutions — QA Engineer → Web Accessibility Auditor (contract) | Jan 2025 – Nov 2025 | La Habana, Cuba

- Delivered 10 months of continuous QA on production web platforms including BillShark.com (Mark Cuban-backed), running functional, regression, and cross-browser testing across critical user flows.
- Led WCAG 2.2 accessibility audits for Crownpeak enterprise sites, documenting compliance issues and partnering with development teams on remediation.

Technologies: Functional and regression testing, WCAG 2.2, screen readers (NVDA/JAWS/VoiceOver), accessibility auditing, cross-browser testing.

ALASolutions — Full-Stack Web Developer (part-time) | Oct 2024 – May 2025 | La Habana, Cuba

- Improved backend performance and usability of a real-time GPS fleet-monitoring system (Traccar-based), fixing defects and redesigning the UI for non-technical Gaviota users.
- Built API integrations for a smart EV-charging platform, including ETECSA Transfermóvil payments across distributed charging stations in Havana.

Technologies: TypeScript, Nest.js, Vue.js, Angular, PostgreSQL, REST APIs.

ETECSA — Backend Developer (contract) | Aug 2024 – Nov 2024 | La Habana, Cuba

- Designed the database and backend for a digital-transformation assessment platform that scores companies across dimensions, perspectives, and areas from structured questionnaires, delivered with a team of three developers.

Technologies: Java, Spring Boot, PostgreSQL, SQLite, REST APIs. Repository: github.com/Erick0330/TETRDIG

Education

B.Sc. Software Engineering — Ciudad Universitaria José Antonio Echeverría (CUJAE) | Feb 2023 – Jun 2026 | La Habana, Cuba

Certifications & Languages

AI Engineering & LLM Orchestration — DataCamp, 2026. [Multi-agent architectures \(LangGraph\)](#), [LLM pipeline orchestration \(LangChain\)](#), [Google Cloud ADK](#).

Machine Learning & Deep Learning — DataCamp, 2026. [Supervised](#) and [unsupervised predictive models](#), and [deep learning with PyTorch](#).

Languages — Spanish: first language. English: Advanced (C2), [EF SET certified](#).

Selected Projects

MARFL: A Robust Multi-Aggregation Operator for Federated Learning | Undergraduate Thesis, 2026

- Designed a robust aggregation operator defending federated training against model-poisoning attacks, composing multiple robust aggregators along a temporal axis (phase transition across rounds) and a spatial axis (multi-stage pipeline within each round).
- Benchmarked against 5 baseline aggregators under 6 poisoning attacks (MNIST, CNN, common protocol), sustaining 87% accuracy where the strongest single-operator baseline dropped to 61% under the same attack.

Technologies: Python, PyTorch, FLEX/FLEXClash, adversarial ML, design of experiments (DOE), NumPy, Matplotlib.

Real-Time Drowsiness Detection System | 2026 — Trained a YOLO (Ultralytics) detector for eye state and yawning on a custom Roboflow dataset and shipped an OpenCV + CustomTkinter desktop app that cuts false positives with debounce and temporal validation.