

Yoandry Collazo Barroso

SENIOR SOFTWARE DEVELOPER · LLM PLATFORM ENGINEERING

Calgary, AB, Canada / ycbarrosocu@gmail.com / linkedin.com/in/ycbarrosocu

PROFILE

I build the engineering *around* large language models — the part that decides whether an LLM feature is safe to put in front of customers. At NDAX, a Canadian crypto exchange, that has meant building an AI platform end to end: agent orchestration, a layered guardrail stack, a replayable eval harness, an authenticated gateway between models and customer accounts, and an MCP tool layer over internal services. Behind the AI work sits more than a decade of production systems engineering — data-centre infrastructure, fintech, aviation and e-commerce — which is why the AI work looks like systems engineering: contracts, tests, observability and an honest account of the failure modes. In a regulated environment, “the model said so” is not an acceptable answer to a compliance reviewer, and designing for that constraint from the start is most of the work.

LLM platform engineering

Agent orchestration

RAG & grounding

LLM evaluation

Guardrails & AI safety

Model Context Protocol

Regulated finance

EXPERIENCE

Jan 2026 – Present

NDAX — *Senior Software Developer*

Calgary, AB

AI platform engineering for NDAX’s customer-facing LLM products. I own the architecture across the three services that make up the platform and review every change into them.

- **Agent orchestrator** — Built the service that turns a customer message into a grounded answer: intent routing, prompt assembly with few-shot selection, tool calling against internal APIs, streamed responses over WebSockets, session continuity and erasable user memory. TypeScript, Azure AI Foundry, Redis, Socket.IO.
- **Guardrails** — Designed the response-safety layer — a layered set of independent checks covering groundedness, structured-output validation with fallback, formatting integrity, provenance and compliance-sensitive content. Replaced embedding-similarity grounding with an entailment-based LLM judge after demonstrating that similarity both false-blocked safe replies and passed answers the corpus did not support.
- **Eval harness** — Built the in-repo evaluation framework — runner, assertion library, LLM judge, versioned baselines — so a prompt or model change is a measurable regression against a replayable suite rather than a hand assessment.
- **AI gateway** — NestJS service that is the single authenticated ingress for AI traffic and the only path to exchange internals, exposing the exchange’s internal APIs to agents as typed tools with generated parameter schemas, per-tool access control and full audit logging.
- **MCP tooling** — Model Context Protocol servers exposing OpenAPI specs and internal service operations as AI-queryable tools over STDIO and HTTP, so engineers and support staff can query production systems from any MCP client, under the same access controls as the underlying APIs.
- **Chat surface** — Preact/Vite chat widget compiled to a standalone CDN bundle, embeddable on any page, plus the assistant and marketing surfaces that consume it.
- **Technical direction** — Author the technical backlog, own the cross-service contracts — WebSocket events, HTTP headers, payload shapes, schema — and run code and security review for the engineers building on the platform.

Jun 2025 – Jan 2026

Independent Contractor — *Confidential client*

Contract software engineering engagement. Client and scope are covered by a non-disclosure agreement.

Jan 2023 – Jun 2025	Credit App — <i>Senior Software Developer</i> Digital finance platform engineering.
Aug 2022 – Jan 2023	Quantifi — <i>Senior Software Developer</i> Backend platform engineering.
Jan 2018 – Aug 2022	Front10 — <i>Lead Software Engineer</i> Full-stack Developer, Jan 2018 – 2019 · promoted to Lead 2019 Miami, FL · remote Led the software development team, building aviation and e-commerce products for North American clients across a four-and-a-half-year tenure.
Jan 2017 – Jan 2018	LUPO Barcelona — <i>Full-stack Developer</i> Full-stack product development. Barcelona, Spain
Sep 2012 – Dec 2016	Dirección de Tecnologías y Sistemas (DTS) — <i>Systems Engineer, Data Centres</i> Data-centre systems engineering: infrastructure build-out and operations across a four-year tenure. Havana, Cuba
TECHNICAL	
PROGRAMMING	TypeScript · JavaScript · Python · Go · SQL
AI & LLM	Azure AI Foundry · Model Context Protocol · agent orchestration · RAG and grounding · LLM-as-judge evaluation · guardrail design · prompt engineering
BACKEND	NestJS · Node.js · GraphQL · WebSockets / Socket.IO · REST and OpenAPI · Redis
FRONTEND	React · Preact · Vite
PLATFORM	Docker · Kubernetes · Terraform · Azure · AWS
EDUCATION & CERTIFICATION	
2012	Ingeniero en Ciencias Informáticas Engineering in Informatics Sciences — Universidad de las Ciencias Informáticas (UCI). Havana, Cuba
	Assessed by World Education Services (2023) as equivalent to a Canadian four-year Bachelor's degree .
2015	HCNA · HCNP — <i>Huawei Certified Network Associate & Professional</i> Data-centre technologies track. Huawei University, Shenzhen. Since renamed HCIA · HCIP . Shenzhen, China
LANGUAGES	
SPOKEN	Spanish — native · English — professional