

Sergio Brito

AI Agent Systems Engineer | Multi-Agent Orchestration · MCP · Multi-LLM · Eval & Safety | Claude Code | 11+ Years

Campo Grande, Brazil (Remote, US-friendly TZ) · sergio.ssbfb@gmail.com · +55 67 98155-8888 · sergio-brito.dev ·
linkedin.com/in/sergio-brito145 · github.com/SergioBrito

SUMMARY

AI Agent Systems Engineer with 11+ years shipping production software end-to-end, focused since 2022 on building agentic systems that actually run in production. I build multi-agent orchestrators, MCP tool servers, multi-LLM routers, and autonomous Claude Code workers — plus the unglamorous layer that keeps them alive: tier-enforced safety in code, eval gates, audit logging, circuit breakers, and observability. Deep pre-AI foundations (Rails, Python, distributed systems, PostgreSQL at scale) are what make these systems reliable instead of demos. Daily Claude Code driver since 2022; led company-wide AI-augmented dev adoption as Principal Engineer.

SELECTED PRODUCTION AI SYSTEMS (built solo, running in production)

Balider — Multi-Agent Orchestrator (flagship) — Agent harness with 3 coordinated agents (orchestrator + mission + marketing), 20+ MCP-style tools, and autonomous Claude Code headless workers as a sub-agent type. Tier-enforced safety in code (green/yellow/red via PreToolUse hooks; red ops require human confirmation), not in the prompt. Multi-LLM routing via OpenRouter with fallback, event-sourced audit log, scheduled smoke evals, circuit breakers. Bootstraps a new project (repo + DB + DNS + deploy) from a single tool call. WhatsApp Cloud API as operator UI.

Rails 8 · Claude Code SDK · MCP (Playwright) · OpenRouter · Solid Queue · DigitalOcean auto-deploy

Railter — Internal LLM Gateway — OpenAI-compatible API unifying Claude, GPT, and Gemini behind one interface; SSE streaming, per-key auth, usage metering and billing, retry/fallback. Every other system calls Railter instead of providers directly.

Rails 8 · OpenAI-compatible API · SSE · Multi-LLM routing

SelfScore — Multimodal Sports Analytics (selfscore.app) — 5-phase video pipeline (FFmpeg chunking → Whisper STT → Gemini Vision segmentation → cutting → AI analysis); live analysis running 8 simultaneous tool calls over WebSocket. Per-player coaching profiles, 5 narrator personalities.

Rails 8 · React Native · Gemini Vision · Whisper · WebSocket · Sidekiq

Giants — Multi-Agent Sales Platform — Multi-channel agents (WhatsApp, Instagram, Facebook) on Claude/GPT/Gemini with structured human takeover and an anti-promise detector that flags risky outputs before they reach a lead.

Rails · Multi-LLM · Hotwire · PostgreSQL

Also in production: Chailter (multimodal personal AI workspace w/ Kids-mode guardrails), Organaizer (multimodal assistant — Whisper + Vision + tool calling), BBE (RAG over proprietary catalogs w/ pgvector), plus dozens of production agents shipped for real businesses.

PROFESSIONAL EXPERIENCE

MONUMENT — August 2025 to May 2026

Senior Data Migration Engineer (AI-Augmented)

US self-storage SaaS startup. Led migration across dozens of enterprise clients, hundreds of facilities, thousands of units.

- Claude Code as primary development driver: agentic scaffolding, refactoring, and migration-diff review.
- Idempotent pipelines with multi-level validation reconciled to the cent before cutover.
- Internal tooling that significantly accelerated migration speed across clients.

Stack: Claude Code, TypeScript, Node.js, MySQL, AWS

VIK — May 2023 to July 2025

Principal Software Engineer / AI Platform Lead

Enterprise B2B healthtech. Designed and shipped the company's AI platform from zero; led platform re-architecture.

- Designed company AI platform from zero: agents, function calling, RAG, multi-LLM orchestration via OpenRouter and LiteLLM.
- Shipped 'Vik Nutri' AI to 1,000+ users — tool use, structured outputs, per-user memory across sessions.
- Drove company-wide AI-augmented dev adoption: prompting conventions, AI code-review patterns, CLAUDE.md standards.
- 99%+ performance gain on enterprise dashboards (query tuning, async jobs, Redis caching).
- Monolith to distributed: bounded contexts, microservices, event-driven messaging (Kafka, RabbitMQ).

Stack: Ruby on Rails, React, OpenAI, Claude, Gemini, OpenRouter, LiteLLM, Kafka, RabbitMQ, BigQuery

META | YOUSE (Caixa Digital Insurance) — November 2022 to May 2023

Senior Software Engineer (Consultant)

Brazil's largest digital insurer. Microservices within a 50+ service event-driven architecture; critical data migration for risk assessment.

- Microservices integrated into a 50+ service Kafka-based event mesh; React dashboards; AWS deployments with CI/CD.

Stack: Node.js, Ruby, React, AWS (S3, RDS, ECS), Kafka

ADMINER — September 2021 to November 2022

Software Engineer

E-commerce analytics platform serving B2C at scale (millions of requests/day).

- 'Adminer Analytics' Chrome extension with 50,000+ downloads; async processing at scale.
- Proxy orchestration and complex aggregations across PostgreSQL, MongoDB, and Elasticsearch.
- Frontend revamp (Angular + Tailwind); CI/CD reducing deployment errors by 90%.

Stack: Node.js, Ruby, Angular, PostgreSQL, MongoDB, Elasticsearch

MJV | BRADESCO INSURANCE — November 2020 to November 2021

Java Software Developer

Critical async batch processing for one of Brazil's largest banks.

- Async batch processing up to 1M records/day; ETL pipelines with Apache Airflow.
- PostgreSQL optimization (indexing, partitioning, query tuning) — cut average query time by 60%.

Stack: Java, Spring Boot, PostgreSQL, Apache Airflow

PSG | DETRAN (Government Traffic Department) — January 2017 to November 2020

Java Developer

Owned the full lifecycle of a high-availability government infractions system at multi-million-vehicle scale.

- High-availability infractions system integrated with multiple government APIs; 99.9% uptime over multiple years.
- Internal tooling, relational data modeling, and DevOps practices on Heroku.

Stack: Java, Spring Boot, PostgreSQL, Heroku

CORE SKILLS

Agent Systems: Multi-agent orchestration, MCP (tool servers + clients), function calling / tool-call loops, sub-agents and hand-offs, autonomous Claude Code workers, context engineering

Multi-LLM: Claude, GPT, Gemini, Llama; OpenRouter / LiteLLM routing, fallback, SSE streaming, structured outputs, cost-aware orchestration

Safety & Evals: Tier-based tool guards (in code), pre-tool hooks, eval frameworks (custom + RAGAS), audit / event sourcing, circuit breakers, human-in-the-loop confirmation

Multimodal: Video (FFmpeg), Vision (GPT/Gemini), Speech-to-Text (Whisper), real-time WebSocket / SSE

Backend: Ruby on Rails 8, Python (FastAPI), Node.js / TypeScript, REST / GraphQL / SSE, Sidekiq / Solid Queue

Data: PostgreSQL (tuning, partitioning), pgvector / RAG, BigQuery, Elasticsearch, ETL and migration

Cloud / DevOps: DigitalOcean, AWS (S3, RDS, ECS, Lambda), Docker, GitHub Actions / GitLab CI, auto-deploy

Foundations: Distributed systems (Kafka, RabbitMQ, DDD, event sourcing), design patterns / SOLID, TDD / BDD, technical leadership and mentorship

EDUCATION & LANGUAGES

B.Sc. Software Engineering — Federal University of Mato Grosso do Sul (UFMS)

Languages: Portuguese (Native) · English (C1 Advanced) · Spanish (Advanced)