

Thiago Peraro — Back-End / AI Engineer

Email perarothiago@gmail.com · WhatsApp +55 62 98216-9350 · Portfolio thiago-peraro.netlify.app LinkedIn linkedin.com/in/thiago-peraro · GitHub github.com/tperaro · Goiânia, Brazil (remote, UTC-3)

Professional Summary

Senior back-end / AI engineer with **6+ years** shipping production systems in regulated, high-stakes environments. Currently the **sole tech lead** at a US crypto-tax consultancy, where I designed and operate a production multi-agent platform on FastAPI + LangGraph powering sales-call ingestion, structured client profiling, automated email and contract workflows, and a customer-facing onboarding portal — running today against a paying client base.

I have built **RAG pipelines on pgvector in production** (oral-cancer surveillance integrated with Brazil's public health network — Best Paper SBCAS 2025), **multi-agent orchestration with tool calling** (LangGraph Supervisor with structured Pydantic outputs), and used **Anthropic Claude as a production tool**, embedded as a knowledge manager inside an internal company-as-code platform.

I treat compliance and audit logging as **architectural constraints, not afterthoughts** — every system ships with architecture decision records in git, signed-webhook verification, fail-closed startup guards, Row-Level Security by role, and validation rails on every public schema.

Experience

Sole Tech Lead — Count on Sheep (Dec 2025 – Present)

US crypto-tax consultancy · DAR (Digital Asset Reconciliation) methodology · Remote

Built and operate the **entire technology stack alone** — backend, AI, infrastructure, dashboards, integrations.

Multi-agent AI on FastAPI + LangGraph

- Multi-agent extraction pipeline converting raw sales-call transcripts into structured client profiles (crypto holdings, DeFi exposure, tax posture) — used by sales and tax teams as a scorecard. LangGraph with parallel-phase orchestration, Pydantic structured output per agent, full LLM tracing via Langfuse + LangSmith.
- LiteLLM Router** providing automatic provider failover (OpenAI → Gemini) and per-key load balancing — eliminated single-vendor outages as a class of incident.
- Tool-calling architecture**: agents invoke typed tools for scheduling, information retrieval and structured extraction, with sanitized inputs and idempotent execution.
- Dual-source real-time call transcription (AWS Transcribe Streaming + Deepgram) fused into a single canonical record; domain-specific LLM corrector for crypto terminology; keyword-spotting sidecar to flag high-signal terms in real time.

RAG, vector stores & audit logging

- Production **RAG on pgvector** (Supabase) — embeddings ingestion, vector search, and a custom adaptive chunker (n-gram overlap rather than fixed-window) for cleaner retrieval; hybrid retrieval with FAISS as a hot-path secondary store.
- Active learning** of crypto terminology so the transcription pipeline keeps improving on real call data — detected mistranscriptions are fed back as fine-tune candidates.
- Audit logging at every layer** — append-only activity log, signed webhook receipts, architecture decision records in git, and validation rails (Pydantic + Postgres CHECK constraints) on every public schema.

Deep PostgreSQL / Supabase

- Distributed leader election in PostgreSQL** for exactly-once execution of background jobs across multiple workers — no ghost duplication, no missed runs.
- Row-Level Security by role** across four production applications, ensuring every read and write is scoped at the database layer, not just in code.
- Query optimization under load** for analytics endpoints powering the executive dashboard — channel attribution and funnel analytics.

Anthropic Claude API in production

- Internal company-as-code platform**: a deployed knowledge hub using **Claude Code as a central knowledge manager** with custom slash-commands, custom skills, and tool-use hooks that auto-validate bidirectional cross-links — letting non-engineers contribute to a fully audited process map.
- Anthropic + OpenAI APIs** powering a daily reporting daemon that summarizes engineering activity for leadership.
- Real-time **Chrome Extension** capturing Google Meet audio client-side, streamed to a FastAPI WebSocket and corrected by an LLM in flight to handle crypto-specific terminology mistranscribed by generic ASR models.

Operations automation & customer SPA

- Deal-stage state machines wiring Stripe payments → customer-portal events → HubSpot CRM transitions, with idempotency keys, leader election, and automatic kanban cards in Slack for the operations team — replacing a manual checklist that ops had been running by hand.
- Customer-facing onboarding portal (React 19 + Vite, served behind Traefik) — designed and built end-to-end so each onboarding client gets a private, status-aware portal: adaptive polling, timezone-aware expiry countdown, build-time CSP parameterized per environment, Sentry session replay with source maps stripped from the served bundle.

Production hardening (regulated-industry posture)

- Multi-stage CI/CD with linting, type-checking, automated tests with coverage, **Bandit SAST**, **Trivy container scanning** failing the build on high-severity findings, and a status gate.
- Weekly dependency-vulnerability scanning and license-compatibility checks; **fail-closed startup guards**, signed-webhook verification on every external integration, payload size caps and rate limiting across the API surface.
- **Sentry** for errors, traces, profiling and cron monitoring of the background job runner — configured with PII scrubbing for compliance.

Stack: TypeScript · React 19 · Next.js 15 · Vite · Tailwind 4 · Python 3.12 · FastAPI · LangGraph · LangChain · LiteLLM · Anthropic Claude · OpenAI · Gemini · **Supabase (Postgres + RLS + pgvector)** · Redis · Docker Swarm · Traefik · GitHub Actions · GitLab CI · Trivy · Bandit · Sentry

NLP Researcher — CEIA & AKCIT / UFG (Dec 2024 – Present)

South America's largest AI research hub · EMBRAPA National Competency Center · CNPq fellow

- **Project Koru:** multi-agent system for assessing leadership soft skills — pydantic-ai agents with structured LLM output, observability dashboard (Streamlit + Plotly + TimescaleDB).
 - **Automatic Red Team adversarial testing:** prompt injection, bias and toxicity evaluation using DeepTeam — systematic test-case generation, not ad-hoc manual probing.
 - Maintains research-team server infrastructure, including self-hosted LLM serving for the group.
-

CNPq Research Fellow — Sobrevidas Project / OCDP / UFG (Nov 2024 – Dec 2025)

Oral-cancer surveillance integrated with Brazil's public health network (SUS, 200M+ citizens) · Best Paper SBCAS 2025

- Built the **RAG chatbot on pgvector** that supports clinicians querying screening protocols and historical cases — custom adaptive chunker (auto-tuned by content length, n-gram overlap) for cleaner retrieval on medical PDFs, hybrid retrieval with FAISS for the hot path, per-session conversation memory.
 - Co-built a **decision-support tool** turning plain-language questions from health managers into SQL against a guarded view, with automatic visualization selection (timeseries / categorical / KPI / table) — eliminating the need to learn the schema or BI tooling.
 - Java back-end: Spring Boot 3 + Clean Architecture, Keycloak (OAuth2/OIDC), RabbitMQ, **PostgreSQL with an FHIR (HL7) schema for health-system interoperability**, observability via Prometheus + Grafana + Loki.
 - **Best Paper Award — SBCAS 2025** (Brazilian Symposium on Computing Applied to Healthcare).
-

Founder & AI Lead — Clivia (Mar 2024 – Dec 2025) · Closed

Multi-agent LLM platform for medical clinics (WhatsApp) · multi-tenant production

- **LangGraph Supervisor architecture** routing each incoming patient message to the right specialist agent before final synthesis — with a separate classifier LLM for sales-funnel stage gating.
 - **Redis-backed graph state** persisted across multiple workers — eliminated message-ordering and race-condition bugs that plagued the first version.
 - **Tool-calling agents** for client lookup, sales-funnel state updates, procedure information and live scheduling — typed inputs, structured outputs.
 - **5-stage automated follow-up** with AI-generated personalized messages timed across the 72 hours after first contact — measurably increasing booking-conversion rates per partner clinic; LLM observability via Langfuse + OpenLit.
-

Backend Developer — GoiásPrev (State Pension System) (Jan 2022 – Sep 2024 · ~3 years)

Government · ISO 12207 process compliance · 120,000 active beneficiaries in production

- Built and maintained Java + Spring Boot **APIs in a regulated environment** where errors affect a retiree's income for years.

- **Deep PostgreSQL work:** query optimization, complex joins on a multi-decade schema; JasperReports modules generating official audit-trail reports consumed by state controllers.
- Dual-track: kept the legacy JSP/Java 8 system running while shipping new Java 11 + Spring Boot 3 services — **no production breakage during the modernization sprint**.
- ISO 12207 software lifecycle, SWEBOK engineering standards, formal sign-off on every change. Promoted intern → CLT employee in 10 months; returned later as PJ contractor for a modernization sprint.

Stack: Java 8/11 · Spring Boot · Spring MVC · Spring Security · JPA · JDBC · JasperReports · PostgreSQL · Jenkins · SonarQube

Skills

- **AI / LLM:** LangGraph · LangChain · CrewAI · pydantic-ai · LiteLLM Router · multi-agent orchestration · tool calling · structured Pydantic outputs · Langfuse · LangSmith · DeepTeam (Red Team)
- **Speech / real-time:** AWS Transcribe Streaming · Deepgram · KWS sidecar · LLM domain correctors · dual-source fusion
- **Back-end:** Python 3.12 · FastAPI · Flask · Java 11/17 · Spring Boot 3 · Spring Security · Keycloak (OAuth2/OIDC) · RabbitMQ · Redis · APScheduler
- **Front-end:** TypeScript · React 19 · Next.js 15 · Vite · Tailwind 4 · shadcn/ui · Streamlit
- **Data & storage:** PostgreSQL (deep) · Supabase (RLS + pgvector) · FAISS · TimescaleDB · FHIR/HL7 schemas
- **Integrations:** HubSpot CRM · DocuSign · Stripe · Gmail API · Google Workspace APIs · n8n · webhooks · Evolution API (WhatsApp)
- **Infra & DevOps:** Docker · Docker Swarm · Traefik · GitHub Actions · GitLab CI · Jenkins · SonarQube · Trivy · Bandit · Linux server administration
- **Observability & compliance:** Prometheus · Grafana · Loki · Sentry · Langfuse · LangSmith · ADRs · MkDocs · company-as-code

Education

B.Sc. Computer Science — Federal University of Goiás (UFG) · 2019 – present (~83% complete)
GPA **7.40** vs class average 5.88 — perfect scores in Computer Architecture, Graph Theory and Electronics; near-perfect in Software Engineering, Parallel Computing, HCI, Probability & Statistics, Databases.
UFG is a top-5 federal university in Brazil; the program emphasizes **scientific methodology** — hypothesis, controlled experimentation, peer-reviewed publication — which I apply to system design.

Awards & Recognition

- **Best Paper — SBCAS 2025** (Brazilian Symposium on Health-Applied Computing) — tools category, Sobrevidas platform.
- **CNPq Research Fellow** — Brazil's national research funding agency (NLP/AI applied to healthcare).
- **Speaker — Campus Party Goiás 5** (Nov 2025, ~30,000 attendees) — credential: "Founder CLIVIA & CEIA Researcher".
- **6th place / top 11%** — **Startup School I9+ / LINK** (2025) — among 54 startups; LINK is Brazil's most prestigious business school.
- **Standout Leader — EMBRAPA Industry 4.0 Training Program** — among 60 fellows.

Languages

- **Portuguese:** Native
- **English:** Full professional proficiency — **EF SET certified** (efset.org/cert/X7aVys)