Pedro Gomes

Backend & AI Engineer

Salvador, Brazil (remote) · +55 71 98792-0958 · devpedrogomes@gmail.com · linkedin.com/in/devpgomes · github.com/DevPedroGomes · portfolio.pgdev.com.br

Summary

I'm Pedro, a backend and AI engineer based in Salvador, Brazil, working remotely. I have been working in development and AI for 4 years. I build applications, integrations and automations that connect a business's systems and reduce manual work. My focus is backend development: APIs, databases, payments, background processing and platforms serving multiple businesses with separate data and permissions.

In AI, I build agents that retrieve information and take action in real systems: updating records, scheduling services, organizing requests and supporting operations through WhatsApp, Telegram or voice. I also develop document search and question-answering systems with sources attached to their answers, along with applications for transcription, translation and real-time audio.

I focus on how these systems behave during routine work. Business rules live in code, important actions can require human review, and decisions are recorded for inspection. My work covers spending limits, failure handling and clear paths for human follow-up.

I deliver everything from architecture and databases to the interface and the application running in production. My main stack is Python with FastAPI and TypeScript with Node.js, React and Next.js, supported by PostgreSQL, Redis, Docker and continuous integration. I use Claude Code and Codex during development, with review and testing to validate what I deliver.

Today I run Gomio, my software business, where I develop custom systems, automations and my own products. I'm also CTO of Peblink in South America, responsible for the technical direction of a platform currently in planning. I work on projects in Portuguese and English.

Experience

Backend & AI Engineer (founder), Gomio · Aug 2026 – present

gomio.com.br

- Gomio is my one-person software house. It sells three services (Operações Gomio, Recuperação de IA em Produção and Sistemas sob Medida) and has two products: Chancela, under Projects, and Gomio Voz, whose demo receptionist, Sofia, answers at voz.gomio.com.br (LiveKit Agents, Deepgram Nova-3, Cartesia), with scripted evaluation conversations in each language that must pass before every release.
- Built the three public operation demos (accounting, law firm and real estate, fictional data): code answers the facts from the records, Claude Haiku only classifies the intent and Claude Sonnet only drafts replies that a person approves. Each has an isolated copy per visitor and a US$ 3 monthly AI cap, and was evaluated with the real model, browser tests and a load test before release.

CTO, South America (contract), Peblink · Aug 2025 – present

- Peblink runs workforce incubators and acceleration programs that connect students with real projects from companies. I'm a partner and lead it in South America; as CTO I'm responsible for the platform, still in planning.

Backend & AI Engineer (contract), consumer telephony startup (under NDA) · Apr 2025 – Aug 2026, backend work into Oct 2026

- Lead backend engineer, sole from March 2026, on an app that calls company phone menus (IVR) and reaches the right department for the user. Built most of the core FastAPI service (about 16k lines of Python, more than 80% authored), its menu-mapping and live-navigation engines and its test and deploy gates. Co-built the Laravel app API and admin panel with another developer, and the API contract the native iOS app uses.
- Built the AI and telephony core: live call audio feeds streaming speech-to-text, and an LLM with tool use reads each menu and decides the next key. Each phone tree is stored in PostgreSQL; the live path drives a conference bridge until it reaches the destination, then hands the call to the user.
- Shipped phone-OTP login (CSPRNG codes, constant-time comparison, lockout and rate limits), a push client that replaced a broken library, and GDPR delete, export and purge.
- Delivered through pull requests, with test suites in CI for both services and deploys of the core service gated on its full suite, coordinating schema migrations across services that share one database.

Backend Developer (contract), Ehchad · Oct 2024 – Jul 2025

- Backend of Ehchad, a live speech-to-speech translator for English, Spanish and French: a Flask-SocketIO server with Whisper transcription (Faster-Whisper or Google Cloud), MarianMT translation on GPU and Google Cloud TTS.
- Moved the app off a local desktop GUI to a headless server on AWS EC2, and built the browser client: an AudioWorklet captures the microphone, and Socket.IO carries the audio and the translated speech.
- Rewrote the server on RealtimeSTT, with Silero VAD and Faster-Whisper replacing the hand-written voice detection, and kept the Socket.IO events compatible with the existing client. A second version runs the aligner only for the active language pair, across all six directions between English, Spanish and French.

Projects

Phonar · consumer telephony app (unreleased)

An app that works through companies' phone menus on the user's behalf and hands the call to them once it reaches the right department. I was the lead backend engineer (Apr 2025 – Aug 2026): I built most of the backend and the AI core that decides the path of the call, plus the tests and checks every release had to pass.

Clínica Aurora · demo agent that writes to a schedule (fictional setting)

aurora.pgdev.com.br · open demo, no signup

An agent books, reschedules and cancels appointments in a separate, fictional clinic ERP over HTTP, and refuses when a business rule says no. Five rules live in a versioned deterministic policy engine, and every decision is logged with its rule. The agent only confirms after the ERP has confirmed; a button takes the ERP down to show it handing the conversation to a person. The schedule is also an MCP server where every write takes a proposal and a commit. The policy engine has 100% test coverage, enforced in CI.

Stack: Python 3.12, FastAPI, Claude Sonnet 5, MCP, server-sent events, Docker, Traefik, GitHub Actions

Chancela · does the AI recommend this business?

chancela.gomio.com.br · live product

Measures whether AI answers mention a local business when someone asks for "the best dentist in Salvador", and which sources they cite. The free audit runs live, question by question (in production: 5 questions, 53 seconds, US$ 0.13). Domain matching respects label boundaries, so a citation of superodonto.com.br never counts for odonto.com.br, and a failed engine call makes the audit inconclusive instead of reporting absence.

Stack: Python 3.12, FastAPI, PostgreSQL (SKIP LOCKED job queue), OpenRouter (GPT, Gemini), Next.js 15, Docker, Traefik

Sofia · AI voice receptionist (Gomio Voz)

voz.gomio.com.br · open demo, in Portuguese and English

Answers a business's phone: it replies only from the declared knowledge base, captures the contact and refuses what is policy (for example, a price without scope). The engine is deterministic: it picks the next question, validates every field and blocks prompt injection before any configuration; the model only writes the sentence. Pipeline: LiveKit Agents (WebRTC and SIP, one call per process) → Deepgram Nova-3 (speech to text) → Gemini 3.1 Flash-Lite through LiveKit Inference, which calls one tool per field it hears → the engine validates, stores and returns the next question → Cartesia Sonic-3 (voice). It ships with a daily call cap, a kill switch and scripted evaluation conversations in each language that must pass before every release.

Stack: Python 3.12, LiveKit Agents, Deepgram, Gemini, Cartesia, Docker, Traefik, GitHub Actions

GovLead · commercial intelligence for government suppliers (in development)

A seller types a CNPJ and sees the open tenders and expiring contracts that match what they already sell. The platform mirrors public data (PNCP, Compras.gov.br, Receita) with a per-source quota and keeps the raw response, and classifies each item by catalog with a vote across classifiers. Relevance was measured on a sample, with a reproducible report. No customers or API yet: the first step is selling it as a per-CNPJ report.

Stack: Python, FastAPI, PostgreSQL, arq, Alembic

agent-ops · extracted from BrainHub, my document Q&A platform

github.com/DevPedroGomes/agent-ops · open source, v0.3.0

A spend cap that refuses when it cannot read its own counter, a decision trail that answers "why did it do that" with a rule code, and a queue that survives a redeploy. Most of its tests run with no Redis, database or network.

Stack: Python 3.12, PostgreSQL, Redis, arq

BrainHub · multimodal Q&A with citations

group-documents.pgdev.com.br · github.com/DevPedroGomes/group-documents

Questions over PDFs, images, audio, video and web pages, answered with cited sources. It flags sources that contradict each other, answers "as of date X" questions and keeps a decision trail. Retrieval combines pgvector, PostgreSQL full-text search, reranking and per-chunk context added at ingest.

Stack: Python, FastAPI, LangGraph, PostgreSQL + pgvector, Redis, Voyage AI, Cohere Rerank, Gemini, Claude, Next.js, Docker, Traefik

More projects

- AI Web Scraping Agent (live): a URL and a sentence become validated JSON, with SSRF and DNS-rebinding defense, page content isolated from the prompt and bring-your-own-key for 6 providers. FastAPI, Playwright.
- Diária: multi-tenant reservation SaaS for Brazilian inns, with a different role per property, Pix payments through Asaas, iCal sync and CPF encrypted with AES-256-GCM. Next.js, Express, Drizzle, PostgreSQL.

Skills

Languages: Python, TypeScript, SQL, Node

Backend and data: FastAPI, Flask, Node.js, Express, Laravel, WebSockets, SSE, PostgreSQL + pgvector, Redis, SQLite

AI: Anthropic Claude (tool use), OpenAI, Gemini, OpenRouter, LangGraph, MCP, RAG (pgvector, hybrid search, reranking)

Messaging and payments: WhatsApp Cloud API, Telegram Bot API, Stripe, Mercado Pago, Asaas

Voice: LiveKit Agents, Deepgram Nova-3, Cartesia, Faster-Whisper, Silero VAD, Google Cloud STT and TTS, OpenAI

Frontend and infrastructure: Next.js, React, Tailwind CSS, Docker, Traefik, GitHub Actions, GHCR, AWS EC2, Linux VPS

AI-assisted development: Claude Code, Codex

Education

Bachelor's in Software Engineering, Universidade Católica de Salvador (UCSAL) · Aug 2024 – present

Full-Stack Software Engineering certificate, Springboard (online) · Feb 2023 – Feb 2024 · 800+ hours

Business English, Stafford House International, San Diego, USA · Sep – Dec 2019

Languages

Portuguese (native), English (fluent)