Pedro Gomes

Engenheiro de Backend e IA

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

Resumo

Sou o Pedro, engenheiro de backend e IA em Salvador, Brasil, e trabalho remotamente. Trabalho com desenvolvimento e IA há 4 anos. Desenvolvo aplicações, integrações e automações que conectam os sistemas de uma empresa e reduzem o trabalho manual. Meu foco é o backend: APIs, bancos de dados, pagamentos, processamento em segundo plano e plataformas que atendem várias empresas, com permissões e dados separados.

Na área de IA, construo agentes que consultam informações e executam ações em sistemas reais: atualizar registros, agendar serviços, organizar solicitações e apoiar operações pelo WhatsApp, Telegram ou voz. Também desenvolvo soluções de busca e perguntas sobre documentos, com respostas acompanhadas das fontes, além de aplicações com transcrição, tradução e áudio em tempo real.

Meu cuidado está em como esses sistemas se comportam no trabalho do dia a dia. As regras de negócio ficam no código, ações importantes podem exigir revisão humana e as decisões ficam registradas para consulta. Trabalho com limites de gasto, tratamento de falhas e caminhos claros para acompanhamento por uma pessoa.

Entrego desde a arquitetura e o banco de dados até a interface e a aplicação em produção. Minha stack principal é Python com FastAPI e TypeScript com Node.js, React e Next.js, apoiada por PostgreSQL, Redis, Docker e integração contínua. Uso Claude Code e Codex no desenvolvimento, com revisão e testes para validar as entregas.

Hoje conduzo a Gomio, minha software house, onde desenvolvo sistemas sob medida, automações e produtos próprios. Também sou CTO da Peblink na América do Sul, responsável pela direção técnica de uma plataforma em planejamento. Atendo projetos em português e inglês.

Experiência

Engenheiro de Backend e IA (fundador), Gomio · ago/2026 – atual

gomio.com.br

- A Gomio é minha software house de uma pessoa só. Vende três serviços (Operações Gomio, Recuperação de IA em Produção e Sistemas sob Medida) e tem dois produtos: a Chancela, em Projetos, e a Gomio Voz, cuja recepcionista de demonstração, a Sofia, atende em voz.gomio.com.br (LiveKit Agents, Deepgram Nova-3, Cartesia), com conversas de avaliação roteirizadas em cada idioma que precisam passar antes de cada publicação.
- Construí as três demos de operação públicas (contabilidade, advocacia e imobiliária, dados fictícios): o código responde os fatos do registro, o Claude Haiku só classifica a intenção e o Claude Sonnet só redige o que uma pessoa aprova. Cada uma tem cópia isolada por visitante e teto de US$ 3 por mês de IA, e foi avaliada com o modelo real, em testes de navegador e num teste de carga antes de ir ao ar.

CTO, América do Sul (PJ), Peblink · ago/2025 – atual

- A Peblink opera incubadoras e programas de aceleração que conectam estudantes a projetos reais de empresas. Sou um dos parceiros e o CTO na América do Sul; a plataforma ainda está em planejamento.

Engenheiro de Backend e IA (PJ), startup de telefonia para consumidor (sob NDA) · abr/2025 – ago/2026, com trabalho de backend até out/2026

- Liderei o backend, sozinho a partir de março de 2026, de um app que liga para os menus telefônicos (URA) das empresas e chega ao departamento certo pelo usuário. Construí a maior parte do serviço principal em FastAPI (cerca de 16 mil linhas de Python, mais de 80% de minha autoria), os motores de mapeamento de menu e de navegação ao vivo e as travas de teste e deploy. Construí junto com outro dev a API do app e o painel administrativo em Laravel, e o contrato de API usado pelo app iOS nativo.
- Construí o núcleo de IA e telefonia: o áudio da ligação ao vivo alimenta a transcrição em tempo real, e um LLM com tool use interpreta cada menu e escolhe a próxima tecla. Cada árvore de menu fica no PostgreSQL; o caminho ao vivo conduz uma conferência até o destino e então passa a ligação ao usuário.
- Entreguei login por OTP no telefone (códigos CSPRNG, comparação em tempo constante, bloqueio e limite de tentativas), um cliente de push novo e exclusão, exportação e expurgo conforme a GDPR.
- Entreguei por pull requests, com suítes de teste no CI nos dois serviços e o deploy do serviço principal esperando a própria suíte completa, coordenando migrações entre serviços que dividem o mesmo banco.

Desenvolvedor Backend (PJ), Ehchad · out/2024 – jul/2025

- Backend da Ehchad, tradutor de voz para voz ao vivo em inglês, espanhol e francês: um servidor Flask-SocketIO com Whisper (Faster-Whisper ou Google Cloud), MarianMT em GPU e Google Cloud Text-to-Speech.
- Levei o app, que dependia de uma interface gráfica local, para um servidor sem interface na AWS EC2, e construí o cliente de navegador: um AudioWorklet captura o microfone, e o Socket.IO leva o áudio e traz a fala traduzida.
- Reescrevi o servidor em cima do RealtimeSTT, com Silero VAD e Faster-Whisper no lugar da detecção de voz escrita à mão, mantendo os eventos Socket.IO compatíveis com o cliente existente. Uma segunda versão roda o alinhador só para o par de idiomas ativo, nas seis direções entre inglês, espanhol e francês.

Projetos

Phonar · app de telefonia para consumidor (não lançado)

Um app que atravessa os menus telefônicos de empresas no lugar do usuário e passa a ligação a ele quando chega ao departamento certo. Fui o engenheiro de backend responsável (abr/2025 – ago/2026): construí a maior parte do backend e o núcleo de IA que decide o caminho da ligação, além dos testes e das verificações que toda versão precisava passar.

Clínica Aurora · demo de agente que grava numa agenda (cenário fictício)

aurora.pgdev.com.br · demo aberta, sem cadastro

Um agente marca, remarca e cancela consultas num ERP de clínica fictícia, separado, por HTTP, e recusa quando uma regra não permite. As cinco regras ficam num motor de política determinístico e versionado, e cada decisão é registrada com a sua regra. O agente só confirma depois que o ERP confirmou; um botão derruba o ERP e ele passa a conversa a uma pessoa. A agenda também é um servidor MCP com proposta e confirmação em toda escrita. O motor de regras tem 100% de cobertura de testes, exigida no CI.

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

Chancela · a IA recomenda esse negócio?

chancela.gomio.com.br · produto no ar

Mede se as respostas das IAs citam um negócio local quando alguém pergunta pelo "melhor dentista em Salvador" e quais fontes elas usam. A auditoria grátis roda ao vivo, pergunta por pergunta (em produção: 5 perguntas, 53 segundos, menos de R$ 1). A comparação de domínios respeita os rótulos, então uma citação de superodonto.com.br nunca conta para odonto.com.br, e uma chamada que falha deixa a auditoria inconclusiva em vez de apontar ausência.

Stack: Python 3.12, FastAPI, PostgreSQL (fila de jobs com SKIP LOCKED), OpenRouter (GPT, Gemini), Next.js 15, Docker, Traefik

Sofia · recepcionista de voz com IA (Gomio Voz)

voz.gomio.com.br · demo aberta, em português e inglês

Atende o telefone de um negócio: responde só o que está na base de conhecimento declarada, captura o contato e recusa o que é política (por exemplo, preço sem escopo). O motor é determinístico: ele escolhe a próxima pergunta, valida cada campo e bloqueia prompt injection antes de qualquer configuração; o modelo só redige a frase. Pipeline: LiveKit Agents (WebRTC e SIP, uma ligação por processo) → Deepgram Nova-3 (fala para texto) → Gemini 3.1 Flash-Lite pela LiveKit Inference, que chama uma ferramenta por campo ouvido → o motor valida, guarda e devolve a próxima pergunta → Cartesia Sonic-3 (voz). Traz teto diário de ligações, chave de desligamento e conversas de avaliação roteirizadas por idioma, que precisam passar antes de cada publicação.

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

GovLead · inteligência comercial para quem vende ao governo (em desenvolvimento)

Quem vende digita o CNPJ e vê as licitações abertas e os contratos vencendo que casam com o que já vende. A plataforma espelha dados públicos (PNCP, Compras.gov.br, Receita) com cota por fonte e guarda a resposta bruta, e classifica cada item por catálogo com votação de classificadores. A relevância foi medida numa amostra, com relatório reproduzível. Ainda sem clientes nem API: o primeiro passo é vender como relatório por CNPJ.

Stack: Python, FastAPI, PostgreSQL, arq, Alembic

agent-ops · extraída do BrainHub, minha plataforma de perguntas sobre documentos

github.com/DevPedroGomes/agent-ops · código aberto, v0.3.0

Um teto de gasto que recusa quando não lê o próprio contador, uma trilha de decisão que responde "por que ele fez isso" com um código de regra e uma fila que sobrevive a um redeploy. A maior parte dos testes roda sem Redis, banco ou rede.

Stack: Python 3.12, PostgreSQL, Redis, arq

BrainHub · perguntas e respostas multimodais com citação

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

Perguntas sobre PDFs, imagens, áudio, vídeo e páginas da web, respondidas com as fontes citadas. Avisa quando duas fontes se contradizem, responde perguntas "na data X" e mantém trilha de decisão. A busca combina pgvector, a busca textual do PostgreSQL, reranking e contexto adicionado a cada trecho na ingestão.

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

Outros projetos

- Agente de scraping com IA (no ar): uma URL e uma frase viram JSON validado, com defesa contra SSRF e DNS rebinding, conteúdo da página isolado do prompt e chave própria para 6 provedores. FastAPI, Playwright.
- Diária: SaaS multi-tenant de reservas para pousadas, com um papel por propriedade, Pix pelo Asaas, sincronização por iCal e CPF criptografado com AES-256-GCM. Next.js, Express, PostgreSQL.

Competências

Linguagens: Python, TypeScript, SQL, Node

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

IA: Anthropic Claude (tool use), OpenAI, Gemini, OpenRouter, LangGraph, MCP, RAG (pgvector, busca híbrida, reranking)

Mensageria e pagamentos: WhatsApp Cloud API, Telegram Bot API, Stripe, Mercado Pago, Asaas

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

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

Desenvolvimento assistido por IA: Claude Code, Codex

Formação

Bacharelado em Engenharia de Software, Universidade Católica de Salvador (UCSAL) · ago/2024 – atual

Certificado em Engenharia de Software Full-Stack, Springboard (online) · fev/2023 – fev/2024 · mais de 800 horas

Inglês para negócios, Stafford House International, San Diego, EUA · set – dez/2019

Idiomas

Português (nativo), inglês (fluente)