

Kaio Alexandre Marques

João Pessoa, PB, Brazil / Remote | +55 (21) 98177-5327 | alexandremarqueskaio@gmail.com
linkedin.com/in/marqueskaio | github.com/kaiomarques93

Professional Summary

Senior Software Engineer with over 5 years of experience specializing in AI-powered backend services, distributed systems, and full-cycle product delivery. Built production OCR and transcription services for the public sector and designed verifiable legal-AI workflows with durable orchestration, privacy-conscious observability, evaluation gates, and strict cost controls. Former Forward Deployed Engineer embedded with U.S. healthcare clients, delivering HIPAA-aligned Revenue Cycle Management automation. Founder of scalable SaaS products built with Python, TypeScript, React, NestJS, event-driven architecture, and AWS.

Professional Experience

Software Engineer, AI Services

Mar 2026 – Present

Public Defender's Office of the State of São Paulo, São Paulo, Brazil (Remote)

- Architected and delivered two production AI services for legal-document OCR and audio/video transcription using Python, FastAPI, Kafka, AWS ECS Fargate, DynamoDB, S3, Terraform, and GitHub Actions OIDC.
- Architected a durable legal case-summarization workflow with LangGraph, modeling five checkpointed stages backed by PostgreSQL so failed jobs resume without repeating completed OCR or paid LLM calls.
- Integrated LangSmith for privacy-conscious tracing and deterministic evaluation of citation coverage, checklist completeness, and explicit gaps; hid legal inputs and outputs by default and persisted human-review outcomes.
- Built a provider-neutral LangChain LCEL adapter with strict JSON Schema outputs, stage-specific model routing, token accounting, and a hard per-job cost ceiling enforced before each model call.
- Established a pre-committed evaluation gate across 4 public court decisions, measuring citation fidelity and narrative completeness while keeping inference costs near R\$0.02 per page.

Forward Deployed Engineer

Jul 2025 – Mar 2026

Thoughtful AI, United States (Remote)

- Embedded with U.S. healthcare clients to translate complex Revenue Cycle Management workflows, including eligibility and prior authorization, into HIPAA-aligned production automations with Python, Playwright, and Selenium.
- Improved eligibility-check pipeline performance by 400% by auditing the data flow and consolidating sequential payer requests into a unified parallel-processing boundary.
- Architected a reusable multi-payer prior-authorization platform that separated payer-agnostic validation from thin portal adapters, supporting multiple insurers without forks in core logic.
- Replaced scattered feature flags with auditable environment-policy contracts across development, sandbox, and production, improving behavioral governance and traceability.

Full Stack Developer

Nov 2023 – Jul 2025

Cubo Tecnologia, Brazil (Remote)

- Built a reusable component platform distributed through AWS CodeArtifact, eliminating duplicated frontend code across 6+ repositories in a multi-tenant B2G SaaS product.
- Implemented technical SEO with Next.js server-side rendering, structured data, and dynamic sitemaps, achieving first-page Google rankings for targeted B2G search terms and increasing organic lead generation.
- Introduced production observability with Sentry and Core Web Vitals, using collected data to diagnose rendering bottlenecks and improve application responsiveness.

Full Stack Developer

May 2023 – Nov 2023

Lab101, Brazil (Remote)

- Diagnosed a critical database bottleneck through query-log instrumentation and designed a change data capture pipeline with Debezium, Kafka, and Elasticsearch, moving 100% of search traffic off the primary MongoDB database.
- Built a daily ingestion pipeline for thousands of videos from the YouTube and Twitch APIs, handling rate limits through scheduled batching, retries, and record deduplication.

Selected Founder Projects

Langu – AI Language-Learning Platform

2026 – Present

Founder & Principal Engineer, Available on Google Play

- Led end-to-end development of a mobile and web AI chat platform: a NestJS backend spanning 14 hexagonal modules, a 6-stage RabbitMQ pipeline with dead-letter queues and a transactional outbox, PostgreSQL, and an offline-first React Native app.
- Designed an idempotent multi-provider billing model for Stripe, Apple, and Google Play with reconciliation workflows and compliant account-deletion and data-purge flows.

Velaris – AI Content Operations Platform

2026 – Present

Founder, Production

- Shipped an AI-assisted political communications platform in 3 weeks, reaching R\$2,000 in monthly revenue against R\$630 in operating costs from its first month.
- Reduced Claude API spend through prompt and token-budget optimization, while implementing a human-in-the-loop approval workflow and audit trail for Brazilian electoral compliance.

Education

Visiting Scholar

Completed

Massachusetts Institute of Technology (MIT), Cambridge, MA, United States

B.Sc. in Materials Engineering

Completed

Federal University of Paraíba (UFPB), João Pessoa, Brazil

Technical Skills

Languages: Python, TypeScript, JavaScript, SQL, HTML5, CSS3

AI Engineering: LangGraph, LangChain, LangSmith, OpenAI and Claude APIs, LLM evaluation, structured outputs, prompt engineering, human-in-the-loop workflows, inference cost optimization

Backend & Messaging: FastAPI, NestJS, Node.js, Kafka, RabbitMQ, microservices, hexagonal architecture, Domain-Driven Design, transactional outbox, change data capture

Frontend & Mobile: React, Next.js, React Native, Tailwind CSS, Redux/Zustand, server-side rendering, technical SEO

Data & Cloud: PostgreSQL, DynamoDB, MongoDB, Redis, Elasticsearch, AWS (ECS Fargate, S3, IAM, CodeArtifact), Terraform, Docker, GitHub Actions CI/CD

Languages Spoken: Portuguese (native), English (fluent / C2), Spanish (intermediate)