Jafet Brito Perez

Senior Software Engineer · AI, Full-Stack & Cloud

Sudbury, ON, Canada · Remote (Eastern Time) · (249) 532-1984

jafetbrito028@gmail.com LinkedIn GitHub jafetbrito.com jafetbrito.art jafetbrito.blog

SUMMARY

Senior Software Engineer with over 10 years of experience architecting and deploying AI-driven search and assistant solutions. Skilled in building secure, scalable data ingestion pipelines and backend integrations using Python, TypeScript and AWS, and in implementing retrieval-augmented workflows and prompt engineering. Led the development of an AI meeting assistant and multi-agent conversational agents, significantly improving accuracy and operational efficiency.

EDUCATION

Bachelor's Degree in Artificial Intelligence — Autonomous University of the State of Morelos (UAEM)

Aug 2011 – May 2015

EXPERIENCE

Founder & Senior Software Engineer — Oliver Academy Jun 2026 – Present

- Developed an AI meeting assistant that recorded Zoom, Teams and Meet calls as well as in-person sessions via mobile; used AWS Lambda and S3 to store audio, applied Llama 3 for note generation, drafted follow-up emails and created CRM tasks. The human-in-the-loop review significantly reduced manual transcription time and increased task-creation accuracy.
- Applied a policy-as-code compliance engine to run pre- and post-meeting checks, enforce retention and attestations, and flag restricted terms; the auditable decision logs enabled faster regulatory reviews and lowered compliance audit effort.
- Engineered conversational recommendation agents using a fine-tuned Llama 3 model with retrieval-augmented generation over Pinecone vector stores, a multi-agent debate system built with LangGraph, and prompt engineering strategies, improving recommendation relevance for knowledge-intensive tasks.
- Orchestrated the design of a decentralized MARL framework with Ray/RLlib and PettingZoo, resolving credit-assignment challenges in cooperative agents by implementing QMIX and MADDPG alongside custom reward shaping.
- Deployed CRM MCP connectors for Salesforce, Wealthbox, Redtail and Microsoft Dynamics with robust token and refresh handling, idempotent writes and conflict-resolution logic, ensuring reliable data synchronization and fewer user-edit conflicts.

Founder & Senior Software Engineer — Rubic's Digital Solutions Sep 2023 – Present

- Architected and optimized high-performance RESTful APIs with FastAPI (Python) and Node.js (TypeScript), aligned with enterprise architecture standards, applying Redis caching and indexing strategies to improve data-integration throughput to 60,000 rpm and reduce API response times.

- Designed and optimized GraphQL backends, solving the N+1 query problem with DataLoader for efficient batching and caching of database queries, improving performance and scalability for complex data fetching.
- Boosted application performance with AWS Elastic Load Balancing, addressing RPS bottlenecks and mitigating DDoS attacks through automated traffic monitoring and rate limiting.
- Resolved concurrency challenges with optimistic locking and message queues such as AWS SQS, improving data integrity and supporting up to 500 concurrent users without performance degradation.
- Crafted a responsive, pixel-perfect website with React/Next.js and TypeScript, optimizing performance with React Query for efficient data fetching and caching, leading to a 25% increase in conversion rates.
- Engineered high-performance server-side rendered (SSR) applications with Next.js 13 and the App Router, significantly improving SEO visibility and navigation.
- Leveraged Capacitor to bridge web apps with native mobile features, shipping hybrid iOS and Android releases in half the time of separate native builds.
- Enhanced DevOps workflows by designing CI/CD pipelines with Jenkins, Docker and Microsoft environments, optimizing Dockerfile layers, multi-stage builds and Docker Hub / AWS ECR caching to reduce image build times.

Senior AI & Full Stack Developer — Amazon `Feb 2021 – Jul 2023`

- Championed the 2-year technical vision for Alexa Subscriptions, aligning 8 product lines and over 50 engineers; the platform powers all voice-driven subscription funnels (Alexa+, Prime, Kids+, Hey Disney!, Skills) and generates multi-million annual recurring revenue.
- Engineered an automated LLM evaluation pipeline benchmarking over 100 model/prompt variants weekly, reducing manual validation by 95% and gating Alexa+ launches across 20 KPIs.
- Standardized the fine-tuning pathway for in-house LLMs, improving 8 conversational KPIs.
- Spearheaded the design of a cross-device content recommender deployed across 6 voice and screen experiences.
- Mentored 9 engineers, leading to 6 promotions to SE2/SE3; conducted over 70 technical interview loops and trained 5 new interviewers.
- Served as Alexa on-call incident lead for 10 high-severity events, cutting MTTR by over 50% and establishing organization-wide ops reviews, design-review forums and knowledge-sharing practices.
- Eliminated 190 ms of user-perceived latency through parallelization and speculative execution.
- Formulated an automated testing and release strategy with Git-based version control and full CI/CD pipelines, replacing manual testing and release procedures.
- Designed and delivered multimodal (voice and visual) feature discovery for Echo Show (2nd Gen), improving discovery and increasing feature usage.
- Architected and launched a bulk Alexa skill enablement service, enabling skill bundling products.
- Integrated third-party payment gateways, email services and social authentication APIs through secure RESTful endpoints.
- Designed and launched voice-activated subscriptions for the Hey Disney! app with AWS Lex and Node.js, defining JSON data contracts and automating deployment with Jenkins; the voice interface enabled hands-free sign-ups.

Full Stack Developer — Tecsolt Feb 2018 – Dec 2020

- Led the design and development of large-scale web and mobile applications for enterprise and startup clients with React, Next.js, Node.js, Go and Python (FastAPI), PostgreSQL and Redis.
- Translated requirements into scalable architecture diagrams, GraphQL APIs and AWS deployment pipelines (ECS, Lambda, S3, CloudFront), enabling faster release cycles.
- Engineered event-driven microservices for analytics, notifications and behavior tracking with Kafka, Redis Streams and GraphQL federation gateways, increasing reliability and throughput.
- Developed real-time communication layers (WebSocket, Server-Sent Events and WebRTC) for live dashboards, chat and telemetry, significantly reducing latency.
- Unified front-end codebases with a shared React / React Native component library and monorepo tools (Nx, Turborepo), cutting development cycle time by more than 25%.
- Implemented CI/CD pipelines with GitHub Actions and Docker, automating integration tests and staging previews and enabling zero-downtime deployments.
- Integrated third-party APIs (Stripe, Firebase, Twilio, Google Maps) and built secure authentication flows with JWT, OAuth 2.0 and AWS Cognito for multi-tenant applications.
- Mentored engineers on clean code, TDD, peer review and observability, and took part in architecture reviews, hiring panels and internal governance shaping engineering standards for scalability, testing and security.

Junior Full Stack Developer — Laserfiche Jan 2016 – Dec 2017

- Delivered React-based customer dashboard features, including Autoship updates, order tracking and pet profile management, in close collaboration with product and UX teams.
- Migrated legacy services to AWS by containerizing workloads and deploying them on Amazon EC2 with Terraform, with CI/CD pipelines in GitHub Actions.
- Reduced P99 latency on high-volume order and inventory APIs through distributed caching, optimized SQL queries and performance analysis during Black Friday and Cyber Monday.
- Developed operational monitoring dashboards for application performance, API errors and fulfillment delays, supporting faster troubleshooting.
- Integrated machine-learning recommendation APIs into customer-facing applications to deliver personalized product suggestions.

SKILLS

AI / ML: LLM development & fine-tuning, RAG, LangChain, LangGraph, Langbase, LangSmith, Hugging Face, AutoGen, CrewAI, DSPy, OpenAI API, Llama 3, PyTorch, TensorFlow, Ray, MLflow, Pinecone, multi-agent systems, prompt engineering & evaluation, Amazon SageMaker, Google Vertex AI

Languages & runtime: Python, JavaScript, TypeScript, Go, Java, C#, C++, Rust, Julia, CUDA, vLLM, Hugging Face Inference

Backend & API: REST, GraphQL, gRPC, Node.js, Flask, FastAPI, Django, Gin, Fiber, ASP.NET Core, OAuth 2.0, JWT, SSO, RBAC, ABAC, Prisma, TypeORM, SQLAlchemy, Redis, Kafka, RabbitMQ, AWS SQS, WebSockets, WebRTC, SSE, Pub/Sub, Amazon API Gateway, NGINX, Kong, Cloudflare

DevOps & cloud: AWS (ECS, Fargate, Lambda, S3, RDS, CloudWatch, EventBridge), GCP, Azure, Terraform, AWS CDK, CloudFormation, Ansible, Docker, Kubernetes, CI/CD, GitHub Actions, Jenkins, GitLab CI, Prometheus, Grafana, ELK Stack, OpenTelemetry, SLO/SLA/SLI design, VPC & networking, VPN, Zero Trust security, auto scaling, enterprise architecture, Microsoft environments

Frontend: React, Next.js, Redux, TanStack Query, SWR, SSR, SSG, SEO, Core Web Vitals (LCP, CLS), TTI, Tailwind CSS, Styled Components, MUI, Figma, Motion, D3.js, Chart.js, Webpack, Vite, Angular, Vue.js

Testing: TDD, BDD, pytest, unittest, Jest, Mocha, Vitest, xUnit, NUnit, Cypress, Selenium, Playwright, model evaluation pipelines, Pandera, k6, Locust, JMeter, OWASP ZAP

Languages: Spanish (native) · English (professional) · French (basic)