Jafet Brito Perez

Ingeniero de Software Senior · IA, Full-Stack y Cloud

Sudbury, ON, Canadá · Remoto (hora del Este) · (249) 532-1984

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

PERFIL

Ingeniero de software senior con más de 10 años de experiencia diseñando y desplegando soluciones de búsqueda y asistentes basados en IA. Experto en construir pipelines de ingesta de datos seguros y escalables e integraciones de backend con Python, TypeScript y AWS, y en implementar flujos de generación aumentada por recuperación (RAG) e ingeniería de prompts. Lideré el desarrollo de un asistente de reuniones con IA y de agentes conversacionales multiagente, mejorando de forma notable la precisión y la eficiencia operativa.

FORMACIÓN

Licenciatura en Inteligencia Artificial — Universidad Autónoma del Estado de Morelos (UAEM)

Ago 2011 – May 2015

EXPERIENCIA

Fundador e Ingeniero de Software Senior — Oliver Academy Jun 2026 – Actual

- Desarrollé un asistente de reuniones con IA que graba llamadas de Zoom, Teams y Meet, además de sesiones presenciales desde el celular; usé AWS Lambda y S3 para almacenar el audio, Llama 3 para generar notas, redactar correos de seguimiento y crear tareas en el CRM. La revisión con intervención humana redujo de forma notable el tiempo de transcripción manual y aumentó la precisión de las tareas creadas.
- Apliqué un motor de cumplimiento basado en políticas como código (policy-as-code) para hacer verificaciones antes y después de cada reunión, aplicar reglas de retención y certificaciones, y señalar términos restringidos; los registros de decisiones auditables agilizaron las revisiones regulatorias y redujeron el esfuerzo de auditoría.
- Construí agentes conversacionales de recomendación con un modelo Llama 3 ajustado (fine-tuning), generación aumentada por recuperación sobre bases vectoriales en Pinecone, un sistema de debate multiagente con LangGraph y estrategias de ingeniería de prompts, mejorando la relevancia de las recomendaciones en tareas intensivas en conocimiento.
- Dirigí el diseño de un marco descentralizado de aprendizaje por refuerzo multiagente (MARL) con Ray/RLlib y PettingZoo, resolviendo problemas de asignación de crédito entre agentes cooperativos con los algoritmos QMIX y MADDPG y funciones de recompensa personalizadas.
- Desplegué conectores MCP para los CRM Salesforce, Wealthbox, Redtail y Microsoft Dynamics, con manejo robusto de tokens y renovaciones, escrituras idempotentes y resolución de conflictos, logrando una sincronización de datos confiable y menos conflictos de edición.

Fundador e Ingeniero de Software Senior — Rubic's Digital Solutions Sep 2023 – Actual

- Diseñé y optimicé APIs REST de alto rendimiento con FastAPI (Python) y Node.js (TypeScript), alineadas con estándares de arquitectura empresarial, aplicando caché en Redis y estrategias de indexación para llevar la

integración de datos a 60,000 solicitudes por minuto y reducir los tiempos de respuesta.

- Diseñé y optimicé backends GraphQL, resolviendo el problema de consultas N+1 con DataLoader para agrupar y cachear consultas a la base de datos, mejorando el rendimiento y la escalabilidad en consultas complejas.
- Mejoré el rendimiento de las aplicaciones con AWS Elastic Load Balancing, eliminando cuellos de botella de solicitudes por segundo y mitigando ataques DDoS con monitoreo automatizado de tráfico y limitación de solicitudes (rate limiting).
- Resolví problemas de concurrencia con bloqueo optimista y colas de mensajes como AWS SQS, mejorando la integridad de los datos y soportando hasta 500 usuarios simultáneos sin degradar el rendimiento.
- Creé un sitio web responsivo y de diseño preciso con React/Next.js y TypeScript, optimizado con React Query para la obtención y caché de datos, lo que aumentó un 25% la tasa de conversión.
- Construí aplicaciones con renderizado del lado del servidor (SSR) en Next.js 13 con App Router, mejorando de forma notable la visibilidad en buscadores y la navegación.
- Usé Capacitor para conectar aplicaciones web con funciones nativas del celular, publicando versiones híbridas para iOS y Android en la mitad del tiempo que requieren desarrollos nativos separados.
- Mejoré los flujos de DevOps con pipelines de CI/CD en Jenkins, Docker y entornos Microsoft, optimizando capas de Dockerfile, compilaciones multietapa y caché en Docker Hub y AWS ECR para reducir los tiempos de construcción de imágenes.

Desarrollador Senior de IA y Full Stack — Amazon `Feb 2021 – Jul 2023`

- Impulsé la visión técnica a 2 años de Alexa Subscriptions, alineando 8 líneas de producto y a más de 50 ingenieros; la plataforma impulsa todos los embudos de suscripción por voz (Alexa+, Prime, Kids+, Hey Disney!, Skills) y genera ingresos recurrentes anuales multimillonarios.
- Construí un pipeline automatizado de evaluación de LLMs que compara cada semana más de 100 variantes de modelos y prompts, reduciendo un 95% la validación manual y sirviendo como control de calidad de los lanzamientos de Alexa+ en 20 indicadores clave.
- Estandaricé el proceso de ajuste fino (fine-tuning) de los LLMs propios, mejorando 8 indicadores conversacionales.
- Lideré el diseño de un recomendador de contenido entre dispositivos, desplegado en 6 experiencias de voz y pantalla.
- Fui mentor de 9 ingenieros, de los cuales 6 fueron promovidos a SE2/SE3; realicé más de 70 entrevistas técnicas y formé a 5 nuevos entrevistadores.
- Fui responsable de guardia de incidentes de Alexa en 10 eventos de alta severidad, reduciendo más de 50% el tiempo medio de recuperación (MTTR) y estableciendo revisiones operativas, foros de revisión de diseño y prácticas de intercambio de conocimiento.
- Eliminé 190 ms de latencia percibida por el usuario mediante paralelización y ejecución especulativa.
- Definí una estrategia automatizada de pruebas y lanzamientos con control de versiones en Git y pipelines completos de CI/CD, reemplazando los procesos manuales.
- Diseñé y entregué funciones de descubrimiento multimodal (voz e imagen) para Echo Show (2.ª generación), mejorando el descubrimiento y aumentando el uso de funciones.
- Diseñé y lancé un servicio de activación masiva de skills de Alexa, que hizo posibles los paquetes de skills.

- Integré pasarelas de pago, servicios de correo y APIs de autenticación con redes sociales mediante endpoints REST seguros.
- Diseñé y lancé suscripciones activadas por voz para la app Hey Disney! con AWS Lex y Node.js, definiendo contratos de datos en JSON y automatizando el despliegue con Jenkins; la interfaz de voz permitió registrarse sin usar las manos.

Desarrollador Full Stack — Tecsolt `Feb 2018 – Dic 2020`

- Lideré el diseño y desarrollo de aplicaciones web y móviles a gran escala para empresas y startups con React, Next.js, Node.js, Go y Python (FastAPI), PostgreSQL y Redis.
- Convertí requisitos en diagramas de arquitectura escalables, APIs GraphQL y pipelines de despliegue en AWS (ECS, Lambda, S3, CloudFront), acelerando los ciclos de lanzamiento.
- Construí microservicios orientados a eventos para analítica, notificaciones y seguimiento de comportamiento con Kafka, Redis Streams y gateways de federación GraphQL, aumentando la confiabilidad y la capacidad de procesamiento.
- Desarrollé capas de comunicación en tiempo real (WebSocket, Server-Sent Events y WebRTC) para tableros en vivo, chat y telemetría, reduciendo de forma notable la latencia.
- Unifiqué el código del frontend con una biblioteca de componentes compartida en React / React Native y herramientas de monorepo (Nx, Turborepo), reduciendo más de 25% el ciclo de desarrollo.
- Implementé pipelines de CI/CD con GitHub Actions y Docker, automatizando pruebas de integración y vistas previas, con despliegues sin tiempo de inactividad.
- Integré APIs de terceros (Stripe, Firebase, Twilio, Google Maps) y construí flujos de autenticación seguros con JWT, OAuth 2.0 y AWS Cognito para aplicaciones multiinquilino.
- Fui mentor de ingenieros en código limpio, TDD, revisión entre pares y observabilidad, y participé en revisiones de arquitectura, paneles de contratación y gobierno interno que definieron los estándares de escalabilidad, pruebas y seguridad.

Desarrollador Full Stack Junior — Laserfiche `Ene 2016 – Dic 2017`

- Entregué funciones del panel de clientes en React, como actualizaciones de Autoship, seguimiento de pedidos y gestión de perfiles de mascotas, en estrecha colaboración con los equipos de producto y UX.
- Migré servicios heredados a AWS contenerizando cargas de trabajo y desplegándolas en Amazon EC2 con Terraform, con pipelines de CI/CD en GitHub Actions.
- Reduje la latencia P99 de APIs de pedidos e inventario de alto volumen con caché distribuida, consultas SQL optimizadas y análisis de rendimiento durante Black Friday y Cyber Monday.
- Desarrollé tableros de monitoreo operativo del rendimiento, errores de API y retrasos de entrega, agilizando la resolución de problemas.
- Integré APIs de recomendación con aprendizaje automático en aplicaciones para clientes, ofreciendo sugerencias de productos personalizadas.

HABILIDADES

IA / ML: desarrollo y ajuste fino de LLMs, RAG, LangChain, LangGraph, Langbase, LangSmith, Hugging Face, AutoGen, CrewAI, DSPy, API de OpenAI, Llama 3, PyTorch, TensorFlow, Ray, MLflow, Pinecone, sistemas

multiagente, ingeniería y evaluación de prompts, Amazon SageMaker, Google Vertex AI

Lenguajes y entornos: Python, JavaScript, TypeScript, Go, Java, C#, C++, Rust, Julia, CUDA, vLLM, Hugging Face Inference

Backend y APIs: 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 y nube: 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, diseño de SLO/SLA/SLI, VPC y redes, VPN, seguridad Zero Trust, escalado automático, arquitectura empresarial, entornos Microsoft

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

Pruebas: TDD, BDD, pytest, unittest, Jest, Mocha, Vitest, xUnit, NUnit, Cypress, Selenium, Playwright, pipelines de evaluación de modelos, Pandera, k6, Locust, JMeter, OWASP ZAP

Idiomas: Español (nativo) · Inglés (profesional) · Francés (básico)