Jafet Brito Perez

Ingénieur logiciel senior · IA, full-stack et cloud

Sudbury (Ontario), Canada · À distance (heure de l'Est) · (249) 532-1984

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

PROFIL

Ingénieur logiciel senior avec plus de 10 ans d'expérience dans la conception et le déploiement de solutions de recherche et d'assistants basés sur l'IA. Expert dans la construction de pipelines d'ingestion de données sécurisés et évolutifs et d'intégrations backend avec Python, TypeScript et AWS, ainsi que dans la mise en œuvre de flux de génération augmentée par récupération (RAG) et de prompt engineering. J'ai dirigé le développement d'un assistant de réunion IA et d'agents conversationnels multi-agents, améliorant nettement la précision et l'efficacité opérationnelle.

FORMATION

Licence en intelligence artificielle — Université autonome de l'État de Morelos (UAEM)

Août 2011 – Mai 2015

EXPÉRIENCE

Fondateur et ingénieur logiciel senior — Oliver Academy Juin 2026 – Aujourd'hui

- Développement d'un assistant de réunion IA qui enregistre les appels Zoom, Teams et Meet ainsi que les réunions en présentiel depuis un mobile ; stockage de l'audio avec AWS Lambda et S3, génération de notes avec Llama 3, rédaction d'e-mails de suivi et création de tâches dans le CRM. La relecture humaine a nettement réduit le temps de transcription manuelle et amélioré la précision des tâches créées.
- Mise en place d'un moteur de conformité « policy-as-code » pour effectuer des contrôles avant et après chaque réunion, appliquer les règles de conservation et les attestations et signaler les termes interdits ; les journaux de décisions auditables ont accéléré les revues réglementaires et réduit l'effort d'audit.
- Conception d'agents conversationnels de recommandation avec un modèle Llama 3 affiné (fine-tuning), de la génération augmentée par récupération sur des bases vectorielles Pinecone, un système de débat multi-agents avec LangGraph et des stratégies de prompt engineering, améliorant la pertinence des recommandations pour les tâches à forte intensité de connaissances.
- Direction de la conception d'un cadre décentralisé d'apprentissage par renforcement multi-agents (MARL) avec Ray/RLlib et PettingZoo, résolvant les problèmes d'attribution du mérite entre agents coopératifs grâce aux algorithmes QMIX et MADDPG et à des fonctions de récompense personnalisées.
- Déploiement de connecteurs MCP pour les CRM Salesforce, Wealthbox, Redtail et Microsoft Dynamics, avec une gestion robuste des jetons et de leur renouvellement, des écritures idempotentes et une résolution des conflits, garantissant une synchronisation fiable des données.

Fondateur et ingénieur logiciel senior — Rubic's Digital Solutions Sept 2023 – Aujourd'hui

- Conception et optimisation d'API REST haute performance avec FastAPI (Python) et Node.js (TypeScript), alignées sur les standards d'architecture d'entreprise, avec cache Redis et stratégies d'indexation pour porter

l'intégration de données à 60 000 requêtes par minute et réduire les temps de réponse.

- Conception et optimisation de backends GraphQL, en résolvant le problème des requêtes N+1 avec DataLoader pour regrouper et mettre en cache les requêtes, améliorant les performances et l'évolutivité.
- Amélioration des performances avec AWS Elastic Load Balancing, suppression des goulots d'étranglement de requêtes par seconde et atténuation des attaques DDoS grâce à la surveillance automatisée du trafic et à la limitation de débit.
- Résolution des problèmes de concurrence avec le verrouillage optimiste et des files de messages comme AWS SQS, améliorant l'intégrité des données et supportant jusqu'à 500 utilisateurs simultanés sans perte de performance.
- Création d'un site web responsive au pixel près avec React/Next.js et TypeScript, optimisé avec React Query pour la récupération et la mise en cache des données, entraînant une hausse de 25 % du taux de conversion.
- Développement d'applications avec rendu côté serveur (SSR) sous Next.js 13 et App Router, améliorant nettement la visibilité dans les moteurs de recherche et la navigation.
- Utilisation de Capacitor pour relier les applications web aux fonctions natives du mobile, publiant des versions hybrides iOS et Android en deux fois moins de temps que des développements natifs séparés.
- Amélioration des flux DevOps avec des pipelines CI/CD sous Jenkins, Docker et environnements Microsoft, en optimisant les couches Dockerfile, les builds multi-étapes et le cache Docker Hub / AWS ECR pour réduire les temps de construction.

Développeur senior IA et full-stack — Amazon `Févr 2021 – Juil 2023`

- Porteur de la vision technique à 2 ans d'Alexa Subscriptions, alignant 8 lignes de produits et plus de 50 ingénieurs ; la plateforme alimente tous les parcours d'abonnement vocaux (Alexa+, Prime, Kids+, Hey Disney!, Skills) et génère des revenus récurrents annuels de plusieurs millions.
- Création d'un pipeline automatisé d'évaluation de LLM comparant chaque semaine plus de 100 variantes de modèles et de prompts, réduisant de 95 % la validation manuelle et servant de contrôle qualité aux lancements d'Alexa+ sur 20 indicateurs clés.
- Standardisation du processus d'affinage (fine-tuning) des LLM internes, améliorant 8 indicateurs conversationnels.
- Pilotage de la conception d'un système de recommandation de contenu multi-appareils, déployé sur 6 expériences vocales et à écran.
- Mentorat de 9 ingénieurs, dont 6 promus SE2/SE3 ; plus de 70 entretiens techniques menés et 5 nouveaux recruteurs formés.
- Responsable d'astreinte incidents d'Alexa sur 10 événements de haute gravité, réduisant de plus de 50 % le temps moyen de rétablissement (MTTR) et instaurant des revues opérationnelles, des forums de revue de conception et des pratiques de partage de connaissances.
- Suppression de 190 ms de latence perçue grâce à la parallélisation et à l'exécution spéculative.
- Définition d'une stratégie automatisée de tests et de mises en production avec gestion de versions Git et pipelines CI/CD complets, remplaçant les procédures manuelles.
- Conception et livraison de fonctions de découverte multimodales (voix et visuel) pour l'Echo Show (2e génération), augmentant l'usage des fonctionnalités.

- Conception et lancement d'un service d'activation groupée de skills Alexa, permettant les offres de skills groupés.
- Intégration de passerelles de paiement, de services d'e-mail et d'API d'authentification sociale via des endpoints REST sécurisés.
- Conception et lancement d'abonnements à commande vocale pour l'application Hey Disney! avec AWS Lex et Node.js, contrats de données JSON et déploiement automatisé avec Jenkins ; l'interface vocale a permis des inscriptions mains libres.

Développeur full-stack — Tecsolt Févr 2018 – Déc 2020

- Direction de la conception et du développement d'applications web et mobiles à grande échelle pour des entreprises et des startups avec React, Next.js, Node.js, Go et Python (FastAPI), PostgreSQL et Redis.
- Traduction des besoins en schémas d'architecture évolutifs, API GraphQL et pipelines de déploiement AWS (ECS, Lambda, S3, CloudFront), accélérant les cycles de mise en production.
- Conception de microservices événementiels pour l'analytique, les notifications et le suivi des comportements avec Kafka, Redis Streams et des passerelles de fédération GraphQL, améliorant la fiabilité et le débit.
- Développement de couches de communication temps réel (WebSocket, Server-Sent Events et WebRTC) pour tableaux de bord en direct, chat et télémétrie, réduisant nettement la latence.
- Unification du code front-end avec une bibliothèque de composants partagée React / React Native et des outils de monorepo (Nx, Turborepo), réduisant de plus de 25 % le cycle de développement.
- Mise en place de pipelines CI/CD avec GitHub Actions et Docker, automatisant les tests d'intégration et les prévisualisations, avec des déploiements sans interruption.
- Intégration d'API tierces (Stripe, Firebase, Twilio, Google Maps) et création de flux d'authentification sécurisés avec JWT, OAuth 2.0 et AWS Cognito pour des applications multi-locataires.
- Mentorat sur le code propre, le TDD, la revue par les pairs et l'observabilité ; participation aux revues d'architecture, aux jurys de recrutement et à la gouvernance interne définissant les standards d'évolutivité, de tests et de sécurité.

Développeur full-stack junior — Laserfiche Janv 2016 – Déc 2017

- Livraison de fonctionnalités du tableau de bord client en React, dont les mises à jour Autoship, le suivi de commandes et la gestion des profils d'animaux, en étroite collaboration avec les équipes produit et UX.
- Migration de services existants vers AWS en conteneurisant les charges de travail et en les déployant sur Amazon EC2 avec Terraform, avec des pipelines CI/CD sous GitHub Actions.
- Réduction de la latence P99 des API de commandes et d'inventaire à fort volume grâce au cache distribué, à des requêtes SQL optimisées et à l'analyse des performances pendant le Black Friday et le Cyber Monday.
- Développement de tableaux de bord de supervision des performances, des erreurs d'API et des retards de livraison, accélérant la résolution des incidents.
- Intégration d'API de recommandation par apprentissage automatique dans les applications clients pour proposer des suggestions de produits personnalisées.

COMPÉTENCES

IA / ML : développement et affinage de LLM, RAG, LangChain, LangGraph, Langbase, LangSmith, Hugging Face, AutoGen, CrewAI, DSPy, API OpenAI, Llama 3, PyTorch, TensorFlow, Ray, MLflow, Pinecone, systèmes multi-agents, prompt engineering et évaluation, Amazon SageMaker, Google Vertex AI

Langages et environnements : Python, JavaScript, TypeScript, Go, Java, C#, C++, Rust, Julia, CUDA, vLLM, Hugging Face Inference

Backend et 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 et 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, conception SLO/SLA/SLI, VPC et réseaux, VPN, sécurité Zero Trust, mise à l'échelle automatique, architecture d'entreprise, environnements 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

Tests : TDD, BDD, pytest, unittest, Jest, Mocha, Vitest, xUnit, NUnit, Cypress, Selenium, Playwright, pipelines d'évaluation de modèles, Pandera, k6, Locust, JMeter, OWASP ZAP

Langues : Espagnol (natif) · Anglais (professionnel) · Français (notions)