JOAO VITOR COELHO

SOFTWARE ENGINEER

Email: joaovcoelho.contact@gmail.com

Phone: +55 21 96532-9850

Location: Rio de Janeiro, Brazil

LinkedIn: linkedin.com/in/sr-staff-john

Senior Python Developer & AI Engineer with 10+ years architecting LLM-powered microservices, RAG systems, and marketing automation platforms for fintech and enterprise AI deployments. Expert in Python, FastAPI, LangChain, AWS Bedrock, Vector Databases, Agentic AI, Prompt Engineering with proven track record: 30% conversational AI adoption increase, 42% fraud detection error reduction, sub-50ms API performance at scale. Specialized in building real-time data pipelines, autonomous agent workflows, and cloud-native architectures (AWS/GCP) for highstakes environments processing $50M+ quarterly transaction volumes. Passionate about leveraging AI-driven automation, cross-functional collaboration, and agile methodologies to deliver measurable ROI in AdTech, fintech, and enterprise SaaS ecosystems.

TECH STACKS

Languages:

- Python
- JavaScript

Technologies / Frameworks:

- FastAPI
- Django
- Flask
- Node.js
- Express
- LangChain
- LangGraph
- Pydantic
- RESTful APIs
- Microservices Architecture
- Agentic AI
- Retrieval-Augmented Generation (RAG)
- Prompt Engineering
- Natural Language Processing (NLP)
- BERT
- Design Patterns
- SOLID Principles
- Clean Code

Application / Database Servers:

- PostgreSQL
- MySQL
- MongoDB
- Redis
- DynamoDB
- Apache Kafka
- Apache Spark
- Vector Databases: Pinecone, Weaviate, LanceDB

Other Tools:

- AWS (Lambda, EC2, S3, IAM, Bedrock)
- AWS SageMaker
- GCP
- Docker
- Kubernetes
- Terraform
- Serverless Architecture
- Observability (Logging & Monitoring)
- Data Pipelines
- ETL

Build Tools / VCS:

- Git
- Jenkins
- GitHub Actions
- Azure DevOps

EDUCATION

Federal University of Rio de Janeiro (UFRJ)

Bachelor's Degree, Mechanical Engineering

EXPERIENCE

Avalara

Jul 2024 – Nov 2025

Software Developer

Project: GenAI Tax Advisor Platform – AI Platform Engineering, LLM Deployment, Conversational AI Optimization for Global Tax Compliance SaaS Clients

- LangChain multi-agent orchestration achieving 30% increase in conversational AI adoption through RAG pipeline architecture connecting proprietary tax compliance documents to Llama 3.1 70B on AWS Bedrock, resulting in 50% faster query resolution for Fortune 500 clients and 20% reduction in support escalations across 12 enterprise deployments
- FastAPI microservices achieving 99.8% uptime for LLM inference APIs through Kubernetes auto-scaling (handling 10,000+ requests/day), resulting in 15% cost optimization via spot instance orchestration and sub-300ms p95 latency for document summarization workflows in compliance-heavy environments
- Prompt engineering & LLM fine-tuning achieving 40% improvement in response accuracy through few-shot learning on tax compliance datasets (leveraging Weights & Biases for A/B testing), resulting in GenAI platform deployment to 12 enterprise clients within 6-month pilot and 92% user satisfaction scores
- Vector database integration (Pinecone) for semantic search across 500K+ regulatory documents, enabling real time RAG retrieval with 85% relevance scores and cutting manual research time by 60% for tax professionals
- Mentored 5 junior developers on LLM integration best practices, reducing onboarding time 40% through comprehensive API documentation and architectural decision records (ADRs) for distributed team collaboration

Tech Stacks: Python, FastAPI, LangChain, LangGraph, AWS Bedrock, Pinecone, Pydantic, Docker, Kubernetes, PostgreSQL, Redis, Intel Gaudi Accelerators, LangSmith Observability, Weights & Biases

WEPayments

Jun 2021 – Apr 2024

Software Developer

Project: Real-Time Fraud Detection & Payment Orchestration – Cloud Automation for High-Volume LATAM Fintech Transactions

- Real-time fraud detection pipeline achieving 42% error rate reduction through BERT-based NLP models (trained on 3M+ transaction patterns via TensorFlow/PyTorch), resulting in $8M+ annual fraud prevention and sub-50ms inference latency on Spring Boot 2 + Kafka microservices processing $12M+ daily transaction volume
- AWS Lambda automation achieving 10x audit scaling (from 500 to 5,000 transactions/hour) through serverless
- Python orchestration (integrating PostgreSQL + Redis), resulting in $120K annual cost savings and PCI-DSS compliance for 15+ financial institution clients across Brazil and LATAM
- API-first architecture achieving 30% throughput improvement through RESTful microservices migration (Node.js Express + Docker), resulting in seamless payment gateway integration with Stripe, PayPal, and custom banking APIs for multi-currency transaction processing ($50M+ quarterly volume)
- Anomaly detection system (unsupervised ML with Isolation Forest) flagging $2M+ suspicious transactions quarterly, reducing manual review overhead by 15% and enabling automated alerting for compliance teams
- Cross-functional collaboration with Product, Data Science, and Security teams to translate business requirements into scalable architectures, delivering 8 major releases in Scrum 2-week sprints with 98% on-time delivery rate

Tech Stacks: Python, Spring Boot, Java, Apache Kafka, BERT, TensorFlow, PyTorch, PostgreSQL, Redis, AWS Lambda, Node.js, Express, Docker, Git, PCI-DSS Compliance

JOAO VITOR COELHO

SOFTWARE ENGINEER

EXPERIENCE

BRQ Digital Solutions

Sep 2018 – Jun 2021

Software Developer

Project: Cloud-Native Microservices Migration & Healthcare Integrations – DevOps Automation for Banking and Hospital Management Platforms

- Kubernetes-managed microservices migration achieving 30% throughput enhancement through monolith decomposition (12-service architecture with Docker + Terraform), resulting in 99.9% SLA compliance for HL7/FHIR healthcare integrations and 20% infrastructure cost reduction for hospital management systems
- Anomaly detection system achieving 25% fraud identification improvement through unsupervised machine learning (Isolation Forest + PostgreSQL), resulting in automated alerting for $2M+ suspicious transaction flagging in banking sector and 15% reduction in manual review overhead
- CI/CD pipeline automation achieving 50% deployment time reduction through Jenkins + Git + Docker orchestration, resulting in twice-weekly production releases (up from monthly) and zero-downtime deployments for banking clients serving 500K+ daily active users
- RESTful API development (Django/Flask) for patient data management systems, integrating HL7/FHIR standards and serving 30K+ daily API requests with 99.95% uptime for healthcare providers
- Led code reviews for 6-engineer squad, establishing SOLID principles and clean code standards that reduced bug rates by 35% and improved test coverage to 85%+

Tech Stacks: Python, Django, Flask, Java, Spring Boot, Docker, Kubernetes, Terraform, PostgreSQL, MySQL, Jenkins, Git, HL7/FHIR, AWS EC2/S3

CI&T

Feb 2016 – Aug 2018

Software Developer Intern

Project: Full-Stack E-commerce Development – Agile Collaboration on Client-Facing Retail & Finance Platforms

- Developed Python-based automation scripts for data migration and ETL pipelines, reducing manual processing time by 70% for retail client onboarding
- Built RESTful APIs and React front-end components for e-commerce platforms, collaborating in Scrum teams to deliver 12+ client projects across retail and finance sectors
- Participated in agile ceremonies (sprint planning, retrospectives, daily standups), gaining foundational experience in cross-functional collaboration and iterative development methodologies
- Contributed to PostgreSQL database optimization, improving query performance by 40% for high-traffic applications serving 100K+ concurrent users

Tech Stacks: Python, Java, JavaScript, React, PostgreSQL, Git, Agile/Scrum