

Nicolas Japas

Senior AI Engineer · Production GenAI Systems on AWS · Retrieval & Agents

Lisbon, Portugal • +351 918 254 312 • German & Argentinian Citizen (no visa required)

nicolasjapas@gmail.com • linkedin.com/in/nicolas-japas • github.com/nicojapas

SUMMARY

Engineer building production LLM systems end-to-end, from the retrieval layer (hybrid search, chunking, re-ranking, metadata filtering) and agent orchestration through the APIs and interfaces that expose them to real users. Currently delivering the retrieval, ingestion, agent and application layers of a live conversational AI product for an enterprise client. Deep AWS background in event-driven serverless architecture, distributed systems design and infrastructure-as-code. Also fine-tunes small open-weight models for on-device use.

CORE SKILLS

Languages: Python, TypeScript, JavaScript, SQL, C++, Bash

AI / LLM: Prompt Engineering, RAG, LangGraph, LangChain, Tool Calling, Amazon Bedrock, OpenSearch Serverless, FAISS, Embeddings (Titan v2), Guardrails, Fine-tuning (LoRA), llama.cpp, LLM Evals, vLLM

Backend: FastAPI, Node.js, REST APIs, Event-Driven, OpenAPI, GraphQL, Pydantic, asyncio, Pytest, Sentry

Frontend: React, Next.js, TypeScript, Vite, MUI, Tailwind, Webpack, Playwright, Jest, Cypress

Cloud: AWS (Lambda, Bedrock, OpenSearch, EventBridge, API Gateway, S3, DynamoDB, Glue, CloudWatch, Cognito, IAM), AWS CDK; Cloudflare (Workers, Pages, D1, KV, R2)

Infrastructure: Terraform, CloudFormation, Docker, Kubernetes, Microservices, GitHub Actions, CI/CD, IaC

Databases: PostgreSQL, DynamoDB, MongoDB, MySQL, Neo4j, Cypher, SQLAlchemy, DuckDB

CERTIFICATIONS

AWS Solutions Architect Associate (SAA-C03), 2026 | **HashiCorp Terraform Associate 004**, 2026 | **Neo4j Certified Professional**, 2026 | **Claude Certified Architect – Professional**, 2026 | **AWS Generative AI Developer – Professional** (AIP-C01), in progress

EXPERIENCE

Senior AI Engineer (B2B Contract) - PwC AppDev | Remote Jul 2026 – Present

- Build and tune the retrieval layer of a production conversational AI assistant: hybrid search over Amazon OpenSearch Serverless (vector kNN with metadata filtering, SigV4 auth), Bedrock Titan v2 embeddings, and chunk re-ranking to control answer precision.
- Fixed the crawler that diffs source-of-truth data into the knowledge base, keeping retrieval current.
- Deliver across the full service boundary in Python and TypeScript: React 18 SPA, FastAPI/Pydantic API on AWS Lambda, and a LangGraph agent using function calling with engineered prompts and structured response types (text, action buttons, data cards).
- Test-driven delivery; structured logging and monitoring for service behaviour, I/O, and performance; ship through QA, security review, and deployment alongside data scientists and DevOps.

Independent Senior AI Engineer / Founder - Japas Audio | Remote Nov 2025 – Present

- Built and released **Sequoia**, a VST3 AI sequencer (C++/JUCE) that writes melodies from text with a LoRA fine-tuned Qwen3-0.6B running offline via embedded llama.cpp, plus its Python data pipeline (~2,000 distilled and code-generated examples) and a 158-request eval with code checks and an LLM judge on Amazon Bedrock.
- Released **Digital Decay**, a commercial VST plugin (C++, JUCE, DSP), at japasaudio.com ↗.
- Built a self-hosted LLM inference stack (Modal, vLLM, Qwen3) in Python behind an OpenAI-compatible API.

Full-Stack Engineer, Backend & Cloud - Brother ISM GmbH | Germany (remote) Jul 2023 – Oct 2025

- Built core backend services for Myze, a B2B SaaS platform for industrial print-on-demand serving ~30 companies across Europe; designed event-driven serverless infrastructure (Lambda, EventBridge, API Gateway, DynamoDB) for multi-step order workflows and cross-company fulfillment.
- Designed multi-tenant REST APIs (OpenAPI) with JWT/OAuth2 and Cognito auth; built time-limited demo accounts via EventBridge, with provisioning on signup and tier downgrade on trial expiry; enabled non-EU shipping with automated customs docs (ShipStation, SendCloud) and Shopify order ingestion (GraphQL).
- Owned full-stack delivery on customer-facing features, from API design through React/TypeScript UI; maintained and migrated unit and E2E test suites (Pytest, Jest, Cypress, Playwright).

Software Engineer - DataZoo GmbH | Germany (remote) Jun 2021 – Jun 2023

- Productionized ML forecasting models via SageMaker-backed APIs, bridging data science output and production systems for industrial time-series applications.
- Built interactive tooling for large-scale industrial time-series data; implemented processing pipelines (DuckDB, Dataslayer) and custom JupyterLab visualization components (Plotly) for high-density datasets.

Earlier (2005–2021): Mechanical Engineer (XPAK USA, TECNA) and University Professor (UNA, Argentina).

TECHNICAL PROJECTS [full list ↗]

AI Document Q&A - Hybrid BM25 + vector RAG with RRF fusion and Voyage reranking [Python, FastAPI, Pydantic] ↗

Menger World - Real-time voice AI agent driving a live WebGL fractal [Python, WebGL, FastAPI, LangGraph, ElevenLabs] ↗

EDUCATION, LANGUAGES & HONORS

Data Analytics (Ironhack) | **M.A. Musical Arts** (UNA) | **Mathematics** (UBA) | **Electronics Tech.** (ITSB)

Spanish: Native | **English:** C2 | **German:** B2 (TELC Certificate) | **Portuguese:** A2

Ironhack Berlin Hackshow Winner (2021) • **Mathematical Olympiad** National Finalist