Nicolas Fonseca Docolas

nicolas@docolas.com.br | https://linkedin.com/in/nicolasdocolas/ | https://github.com/ndocolas

Professional Summary

AI Engineer with 5+ years building and shipping production LLM systems across enterprise and founder roles, spanning **agentic platforms**, **RAG pipelines**, and **multi-agent orchestration**. Backend-first engineer working with **Python, FastAPI, LangChain/LangGraph, vector databases, Docker, Kubernetes, and Terraform** on **GCP, Azure, and AWS**, focused on robust LLM architecture and the engineering practices that drive accurate, grounded model responses from real document sources. Daily practitioner of spec-driven development through **Claude Code, Cursor, MCP, and custom Agent Skills**.

Highlights

- Founded **24p7**, a multi-tenant AI SaaS for real estate, and scaled it to **3 paying clients** and **3,000+ messages/month** within 2 months of launch as solo founder and lead engineer.
- Co-built **HP Agent AI Studio**, the agentic AI platform now serving **5,000+ HP employees** in production with database-layer RBAC for auditable access control.
- Drove **MySales360**'s LLM router from **65% to 98%** accuracy (BERTScore vs ground truth) through prompt and agent-description engineering on a multi-agent enterprise system.
- Led on-prem POC migrating an Azure OpenAI chatbot to a local Ollama stack, cutting **inference cost by 50%** while retaining **95% of cloud-model accuracy**.

Experience

- **24p7 — Helper [🌐]** *Feb 2026 – Present*
  Founder & Lead Engineer (Part-time) Porto Alegre, Brazil
  - *Python, FastAPI, LangChain, Google Gemini, PostgreSQL, pgvector, Redis, Docker, Terraform, GCP, React, TypeScript*
  - Architected and deployed a **multi-tenant AI SaaS** for real estate on GCP — onboarded **3 paying clients** and **3,000+ messages/month** within 2 months; single codebase serves all tenants via PostgreSQL **RLS**, provisioned with Terraform, containerized with Docker, system designed for **3,000 concurrent users**.
  - Built an **agentic AI system** (LangChain + Gemini) with dynamic tool calling: **RAG knowledge retrieval**, WhatsApp image delivery, user-interest capture for targeted broadcast campaigns, and native **voice transcription** via Gemini multimodal API.
  - Hardened for production: Redis message buffer (groups rapid user inputs into one LLM request), per-user **rate limiter**, and **prompt injection defenses** (input sanitization, UNTRUSTED-marker wrapping).
  - Implemented **evaluation metrics** to validate the Agentic RAG pipeline using **RAGAS** (faithfulness, answer relevancy, context precision/recall) and **LLM-as-judge**, enabling regression testing and quality tracking across prompt and model changes.
  - Delivered **React + TypeScript** admin panel with live **RAG document management**: operators upload or replace knowledge base on-the-fly, AI uses updated context in the very next conversation — no redeployment needed.
- **TELUS Digital (Poatek) [🌐]** *Mar 2026 – Present*
  AI Engineer Porto Alegre, Brazil
  - Develop **AI microservices** (Python, FastAPI, LangChain) and **RAG pipelines** for enterprise NLP workloads on cloud infrastructure.
  - Apply **prompt engineering** and **LLM evaluation harnesses** (BERTScore, LLM-as-judge) to improve response accuracy and reliability of production endpoints.
  - Contribute to **model selection** and **cost/latency optimization** for client-facing GenAI features.
- **Hewlett-Packard Inc. [🌐]** *Jan 2024 – Feb 2026*
  Machine Learning Engineer Porto Alegre, Brazil
  - *Promoted from Junior Machine Learning Engineer in Oct 2024.*
  - **HP Agent AI Studio:** Built backend services (Python, FastAPI) for agentic AI platform serving **5,000+ HP employees** on Microsoft Foundry; designed **RBAC** system at the database layer for structured, auditable access control.
  - **MySales360:** Designed multi-agent orchestration (LangChain, LangGraph) with **DAG-based query decomposition**, parallel node execution, and Redis-backed state persistence; improved **LLM router accuracy from 65% → 98%** (BERTScore vs ground truth) via prompt and agent-description engineering.
  - **HP Assistant:** Built core orchestration layer (Python, FastAPI, LangChain) routing tasks across specialized agents (translation, summarization, DALL·E image gen, email, PDF); added evaluation harnesses to measure and improve reliability at scale.

- **On-Prem Migration (POC):** Migrated Azure OpenAI chatbot to **Ollama local stack** — benchmarked Qwen, Llama, DeepSeek with quantization, orchestrated with Kubernetes; achieved **50% cost reduction** retaining **95% of cloud model accuracy**.
- **Delfos:** Built GitHub-integrated AI assistant (Python, FastAPI, LangChain, Gemini, **ReAct**) that autonomously parses repositories, detects misconfigurations, and generates reproducible onboarding instructions.

- **LIS — Software Innovation Laboratory (HP/PUCRS) [🌐]** *Jan 2022 – Dec 2023*
  Software Engineer Porto Alegre, Brazil
  - Benchmarked **8+ multimodal LLMs** (LLaVA, BLIP, MiniGPT, Qwen-VL families) on resource-constrained HP hardware; evaluated **quantization** (INT8/INT4) and model-selection strategies to maximize throughput within tight memory and compute budgets.
  - Designed scalable **Python data pipelines** for ingesting, transforming, and evaluating large multimodal datasets feeding LLM behavior studies.
  - Authored internal research (HP/PUCRS) on **efficient on-device LLM deployment** patterns later reused across enterprise projects.

Skills

- **AI-Assisted Development:** Claude Code, Cursor, Custom Agent Skills, MCP Tools, AI-Augmented Workflows, Prompt-Driven Development
- **Core AI/LLM:** LangChain, LangGraph, RAG, Agentic Systems, Tool Calling, Multi-agent Orchestration, Prompt Engineering, Evaluation Pipelines
- **LLM Providers & Models:** Google Gemini, Azure OpenAI, Qwen, Llama, DeepSeek, Transformers, Embeddings (sentence-transformers, Gemini), Quantization (INT8/INT4)
- **ML & Model Serving:** Classical ML (regression, classification, clustering), scikit-learn, Vector Search (pgvector, FAISS), Model Serving & Inference (Ollama, vLLM), Throughput & Latency Optimization, GPU/CPU Benchmarking
- **Backend:** Python, FastAPI, Go, Java, SQL, Pydantic, PyTest, Locust
- **Data & Infra:** PostgreSQL (pgvector, RLS), Redis, MongoDB, FAISS, Docker, Kubernetes, Terraform, GCP, Azure, AWS, CI/CD
- **Frontend:** React, TypeScript

Projects

- **RAG PDF Assistant — Gemini + LangChain** *Sep 2025*
  Python, FastAPI, LangChain, Gemini, FAISS, Streamlit, PyPDF2, Pydantic [🐙]
  - Built complete **RAG pipeline**: FAISS vector store, Gemini embeddings, cosine similarity retrieval, dynamic prompt injection; enforces context-grounded answers, source citation, and user language mirroring.
  - Modular FastAPI backend (RetrievalService, ChatService) + Streamlit UI; session-aware orchestration for multi-user workloads.

Achievements

- **HP International Case Competition Winner** *Dec 2024*
  1st place — designed and implemented new AI feature for HP AI Companion Certificate: 🌐
- **AGES Featured Project** *Jun 2024*
  University Award — Experimental Software Engineering Agency Certificate: 🌐

Education

- **Pontifical Catholic University of Rio Grande do Sul (PUCRS)** *Jan 2023 – Dec 2026*
  B.E. Software Engineering Porto Alegre, Brazil
  - **GPA:** 9.3/10 (∼3.7/4.0)
  - **Coursework:** Machine Learning, AI, NLP, Distributed Computing, Algorithms & Data Structures, Databases

Additional Information

- Portuguese (Native)
- **English (Fluent)** — certified C2 in two 2-month immersion programs abroad: **Fort Lauderdale, FL (2024)**, **Vancouver, Canada (2025)**
- Spanish (Basic A2)