Pedro Fioravante

AI Engineer · LLM Agents, Document AI, Computer Vision & Speech

Brazil (UTC-3), remote · pedro.fioravante.dev@outlook.com · linkedin.com/in/fioravante-dev · github.com/fioravante-dev

SUMMARY

AI Engineer with about 2 years of professional experience, almost all of it shipping GenAI systems to production end to end, from prototype to deployed service. Built a document-extraction pipeline that processes 30,000+ documents a month at 97% accuracy. Works across LLM integration, agents, RAG, OCR, computer vision and speech, plus the Python/TypeScript backends and infrastructure around them. Fluent English (taught English for 8 years).

EXPERIENCE

Statum — Ribeirão Preto, Brazil (remote) · May 2025 – Present

AI Engineer (R&D) · Full-time — **Aug 2025 - Present**

- Designed and built a document-extraction pipeline combining OCR and LLMs with structured outputs that processes **30,000+ tax documents a month** in production with **97% accuracy**; the rest is routed to manual review, mostly low-quality scans.
- Built LLM agents with tool calling (Claude API, MCP, LangGraph) that triage requests, read codebases and propose fixes as pull requests for human review.
- Build agents that act in real time during meetings: Whisper transcription, speaker diarization and low-latency responses.
- Train and benchmark computer-vision models (CNNs, YOLO family) and serve them behind APIs.
- Run hosted and local models, including LLMs and SLMs on llama.cpp for sensitive data and PII anonymization.
- Evaluate LLM pipelines with LangSmith tracing and in-house test sets.
- Ship Python (FastAPI), TypeScript (React, Next.js, Node.js) and Rust services with Docker and CI/CD on Azure DevOps.

AI & Software Developer Intern — **May 2025 - Sep 2025**

- Built a natural-language-to-SQL agent with LangChain over MySQL, served by a containerized FastAPI backend (Docker Compose).
- Built a RAG pipeline with LlamaIndex over multiple documents, including scanned PDFs through Tesseract OCR.
- Compared LLMs for reliability and latency, including Llama 3 via Groq.

Minas Gerais State University (UEMG) — Sep 2023 - Oct 2024

Undergraduate Researcher · Part-time

- Built an Android farm-management app (Kotlin, Jetpack Compose) with a Node.js/Express backend and JWT authentication.

English Teacher · language schools and private students — 2015 – 2023

PROJECTS

talk · github.com/fioravante-dev/talk
100% local voice dictation for Linux and Windows: press a hotkey, speak, and the text lands at the cursor. Silero VAD + Whisper (whisper.cpp on CUDA), written in Rust. ~100 MB of RAM and no VRAM while idle.

op2 · github.com/fioravante-dev/op2
Rust tray app driving a desktop client's local API over WebSocket, with data pulled through an MCP server. 1.5 MB binary, ~5 MB RAM.

Client SaaS · freelance, 2026
Multi-tenant scheduling SaaS for dental clinics (Postgres row-level security, audit trail) and an offline-first attendance PWA for a factory. Next.js, Supabase, test-first.

SKILLS

LLMs & agents: Claude API, Azure OpenAI, Groq, Google AI APIs, LangChain, LangGraph, LlamaIndex, LangSmith, MCP, prompt engineering, tool calling, structured outputs

Document AI & RAG: OCR (Tesseract, VLMs), structured extraction, embeddings (SentenceTransformers), vector databases (Pinecone), semantic search

ML & data: PyTorch, HuggingFace Transformers, scikit-learn, pandas, NumPy, YOLO, CNNs, Whisper, llama.cpp, dataset building and annotation

Backend & infra: Python, FastAPI, TypeScript, Rust, Next.js, Node.js, PostgreSQL/Supabase, MySQL, Docker, Azure, Azure DevOps CI/CD, Linux

EDUCATION & LANGUAGES

B.S. in Information Systems · Minas Gerais State University (UEMG) · 2025
Portuguese (native) · English (fluent)