

Marcelo Prates, PhD

Senior AI/ML Engineer | Full-Stack GenAI Systems, Agentic RAG, and Production ML

Porto Alegre, Brazil (UTC-3) | Open to international remote roles

marceloorp@gmail.com | linkedin.com/in/marceloprates | github.com/marceloprates

Summary

Senior AI/ML engineer with 7+ years across data science, ML engineering, and applied AI, plus PhD research experience. Builds and ships production GenAI and ML systems across React/Next.js frontends, FastAPI and TypeScript backends, retrieval/reranking, agent runtimes, evaluation, and operational reliability.

Core Skills

GenAI and LLM Systems: OpenAI APIs, Azure OpenAI, LlamaIndex, Agentic RAG, ReAct Tool-Using Agents, DSPy, PydanticAI, MCP Integration, Tool Calling and Orchestration, Structured Outputs, Metadata Filtering, Hybrid Retrieval and Reranking, Memory and Context Management, Vercel AI SDK; AI Stack Ecosystem: LangChain, LangGraph, OpenAI Agents SDK, CrewAI, Semantic Kernel, Hugging Face Transformers, Pinecone, Qdrant, ChromaDB, FAISS, MongoDB, Convex; Full-Stack Product Engineering: Next.js 15, React, TypeScript, Tailwind, Clerk, FastAPI, SSE Streaming, REST APIs; Data and Retrieval Infrastructure: PostgreSQL, pgvector, Supabase, Redis, Elasticsearch, SQL, Vector Search; ML and Evaluation: Python, PyTorch, Scikit-learn, MLflow, Bayesian Optimization, Model Evaluation, Experimentation, Regression QA; Platform and Delivery: Docker, Kubernetes, CI/CD, OpenTelemetry, Prometheus, Grafana, Vercel, DigitalOcean App Platform; System Design.

Experience

Lead Engineer

HopHR, Remote, Brazil

Present

Oct 2025– Present

- Led Moon's production candidate-search platform, unifying retrieval, scoring, reranking, and explainable output so ambiguous hiring requests could become auditable shortlist decisions.
- Implemented backend APIs, async workflows, and runtime infrastructure on Supabase/PostgreSQL with Redis-backed caching and queues to support reliable, high-throughput hiring operations.
- Improved release safety through CI/CD hardening, test-driven refactors, and stronger observability across search, streaming, and agent-runtime components.
- Designed production AI search architecture combining Elasticsearch filters, semantic retrieval, multi-criteria reranking, and LLM-based assessments to improve shortlist quality while keeping results inspectable by recruiters.
- Built chat-driven hiring workflows with planning, tool orchestration, memory-aware context injection, and AI SDK-compatible SSE streaming across backend and product surfaces.
- Designed and operated a PydanticAI-centric multi-agent runtime: specialist profiles, delegated subtasks, memory-aware shortlist re-evaluation, and feedback-triggered workflow adaptation.
- Built persistent LLM memory and session-aware context-injection infrastructure for candidate-search learnings with confidence-scoped retrieval across sessions.
- Strengthened runtime observability with request correlation, trace propagation, safe logging/redaction, and metrics/tracing patterns suited for debugging long-running agent workflows.
- Stack: TypeScript, Next.js 15, React, Tailwind, Clerk, Supabase, PostgreSQL, Redis, Elasticsearch, OpenAI APIs, PydanticAI, Vercel AI SDK.
- Developed model-to-filter translation that turns learned SHAP/feature-importance signals into executable Elasticsearch filter clauses.
- Partnered with recruiters and hiring managers to convert ambiguous role requirements into testable search/ranking hypotheses.

Lead Data Scientist

Dataside, Remote, Brazil

Oct 2025

Oct 2023– Oct 2025

- Promoted to Lead Data Scientist; mentored team members and owned client-facing discovery, solution design, and technical delivery across multiple accounts.
- Standardized evaluation, documentation, deployment, and stakeholder-alignment routines so ambiguous business problems became measurable ML roadmaps, data requirements, and KPIs.
- Designed and shipped agentic RAG systems with tool-augmented retrieval, stateful memory, and graceful degradation for autonomous enterprise knowledge assistants (Azure OpenAI, LlamaIndex ReAct workflows, PostgreSQL/pgvector).
- Built research-assistant and ranking workflows with dynamic question generation, document-scoped search, hybrid retrieval modes, rank fusion, and source-aware answers for autonomous answer quality.
- Hardened the retrieval layer with safe vector-store initialization, advisory-lock-based schema setup, embedding-dimension safeguards, and non-destructive defaults for live indexed data.
- Designed autonomous fallback paths and regression QA workflows — CLI probes and Playwright-backed browser validation — so assistants remained usable when orchestration dependencies or credentials were unavailable.
- Delivered full-stack agentic applications with FastAPI backend and React frontend, deployed to Azure Container Apps with Grafana observability and Docker Compose orchestration.
- Delivered unsupervised and decision-science solutions including clustering/segmentation, anomaly detection with SHAP-based explanations, experimentation/A-B testing, and hybrid NLP pipelines combining TF-IDF, BM25, embeddings, UMAP, HDBSCAN, and vector search.
- Architected production ML pipelines with PySpark, SQL, Azure, Databricks Jobs, Azure DevOps CI/CD, MLflow model registry, containerized train/predict workflows, and scheduled retraining or model-renewal loops.

AI Engineer

SingularityAI (Part-Time), Remote, Brazil

Jun 2024

Nov 2023– Jun 2024

- Led end-to-end development of AI voice assistant with real-time transcription, TTS generation, and persona-aware RAG responses.
- Built YouTube→ChromaDB pipeline with AssemblyAI transcription, speaker diarization, and metadata-rich chunking for knowledge-base extraction.
- Implemented LangGraph agents with PostgreSQL checkpointing for persistent conversation state across sessions.
- Integrated ElevenLabs TTS, AssemblyAI STT, and Ultravox for real-time voice conversation with audio event handling (interruptions, pauses).
- Designed PCA-based scope analysis and concurrent 22-attribute persona extraction for personalized, personality-adjusted responses.
- Built MinIO-synced ChromaDB vector store with Redis caching for production-ready retrieval infrastructure.

LLM Consultant

Vortigo (Part-Time), Remote, Brazil

Jun 2024

Nov 2023– Jun 2024

- Built internal assistant chatbots grounded on proprietary code and spreadsheet knowledge sources.
- Delivered retrieval and prompt pipelines with production guardrails for repeatable workflows.

Senior Data Scientist

Condati, Remote, USA/Brazil

Dec 2023

Nov 2021– Dec 2023

- Led redesign of bid-optimization and forecasting ML systems for digital marketing workflows.
- Built Python and SQL pipelines for production forecasting, anomaly detection, and decision support on Snowflake and Aurora-backed marketing data.
- Partnered with business stakeholders to align model behavior with campaign KPIs and operational constraints.

ML / Computer Vision Consultant

ConstructIN, Remote, Brazil

Apr 2023

Mar 2022– Apr 2023

- Led production computer-vision delivery for construction monitoring and analytics.
- Coordinated model development, validation, and delivery workflows with client teams.

Senior AI Researcher (ML for Health)

Samsung Research Brazil, Campinas, Brazil

Nov 2021

Mar 2020– Nov 2021

- Led wearable-health ML feature development deployed globally in Galaxy Watch products.
- Delivered memory-efficient inference designs for constrained devices and real-time operation.
- Coordinated cross-functional research-to-product delivery with HQ stakeholders.

Data Scientist

Poatek IT Consulting, Porto Alegre, Brazil

Mar 2020

Jun 2019– Mar 2020

- Delivered DS/ML projects across optimization, NLP, computer vision, and risk modeling.

Education

PhD in Computer Science

Aug 2015 – Jul 2019

Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

Research focus: machine learning, AI ethics, graph deep learning, and combinatorial optimization.

Selected publications: AAAI 2019, IJCAI 2020/2021, NCA 2020.

Selected Impact

Built and shipped production AI systems across recruiting, research-assistant, and enterprise-grounding use cases using OpenAI APIs, Azure OpenAI, LlamaIndex, retrieval/reranking, and tool-driven workflows. Backed GenAI delivery with production ML experience in evaluation, experiment tracking, and deployment reliability.