Wanderson Gomes de Souza

Senior AI Engineer | Python Developer | GenAI, RAG & Agents

wandersonsouza.info@gmail.com +5583988579083 linkedin.com/in/wandersongsouza

wandersonsouza.com

PROFILE

Senior AI Engineer with 10+ years of experience designing, developing, and deploying AI-driven applications across Generative AI, NLP, RAG, Multi-Agent Systems, Computer Vision, and production machine learning systems. PhD with a strong background in applied machine learning, experimental design, data pipelines, and scalable AI solutions. Hands-on experience building Python applications using FastAPI, LangChain, LangGraph, CrewAI, OpenAI APIs, embeddings, vector databases, semantic retrieval, prompt engineering, and context orchestration. Strong ability to integrate LLMs with enterprise systems, optimize retrieval pipelines, and deliver maintainable AI solutions in cloud and production environments, collaborating with technical and business teams to solve complex problems.

Main Skills

Python Development · Generative AI · LLM Systems · RAG Architectures · Agentic Workflows · Multi-Agent Systems (CrewAI, LangGraph) · LangChain · OpenAI API · LlamaParse · Hugging Face Transformers · Prompt Engineering · Context Engineering · Vector Databases & Embeddings · FAISS · Semantic Search & Indexing · API Development (FastAPI, Flask) · SQL & NoSQL Databases · Data Pipelines · Cloud Platforms (Azure, AWS) · Docker · Kubernetes/AKS · Production AI Systems

PROFESSIONAL EXPERIENCE

Tech Lead / Senior AI Specialist 07/2025 – 05/2026

Motiva

- Acted as a hands-on technical leader, developing Python-based GenAI applications with RAG, semantic search, and agentic workflows.
- Built backend services and APIs with Python and FastAPI to integrate LLM components with enterprise systems and operational workflows.
- Engineered multi-agent workflows using LangGraph, enabling dynamic, context-driven reasoning and decision-making across complex AI pipelines.
- Implemented retrieval pipelines using embeddings, vector databases, and indexing strategies to improve relevance, grounding, and accuracy.
- Applied prompt engineering, context orchestration, and LLM evaluation strategies to improve consistency, reliability, and response quality.

Senior Machine Learning Engineer 04/2024 – 07/2025

Venturus

- Developed and architected an embedded LLM-based assistant for Smart TV environments, enabling contextual interaction over technical documentation.
- Designed modular RAG pipelines using LlamaParse for document parsing, with ingestion, chunking, semantic retrieval, context assembly, and response generation.
- Built Python-based integrations connecting LLM components, retrieval pipelines, and backend services for real-world AI applications.
- Implemented context management strategies to improve response quality under latency, memory, and limited context-window constraints.
- Evaluated LLM outputs by analyzing grounding, consistency, relevance, and response reliability in RAG-based conversational systems.

Senior Machine Learning Engineer 04/2023 – 04/2024

iTriad

- Developed production-grade Python services for real-time AI systems, supporting scalable deployment, backend integration, and inference workflows.
- Built FastAPI-based inference APIs to expose ML models through maintainable services for operational and real-time applications.
- Optimized inference pipelines, reducing latency by approximately 25% while preserving model performance and system reliability.
- Designed evaluation workflows using IoU, confusion matrices, threshold tuning, and error analysis to improve model robustness.
- Delivered computer vision models for head pose estimation, achieving ~9° MAE and improving robustness under pose variation.

Data Scientist 01/2021 – 04/2023

NeoPTO

- Developed large-scale NLP pipelines processing 11M+ patent documents, enabling text analysis, indexing, classification, and retrieval.
- Built Python and PySpark ETL workflows for ingestion, cleaning, normalization, and preparation of large unstructured text datasets.
- Designed retrieval-oriented data pipelines to structure patent titles, abstracts, claims, and descriptions for downstream NLP tasks.
- Worked with SQL and NoSQL databases to manage structured metadata and unstructured text used in scalable NLP and retrieval pipelines.

Machine Learning Researcher 03/2020 – 01/2021

GPICEEMA

- Led development of a computer vision system automating analog densimeter readings, enabling accurate measurements in industrial environments.
- Designed validation protocols improving robustness under illumination changes and optical distortions.
- Integrated computer vision and OCR into an end-to-end pipeline, enabling automated monitoring and data extraction.

Machine Learning Researcher 03/2015 – 03/2020

CBTU

- Led deployment of a real-time computer vision system for passenger counting and flow estimation on edge devices in production environments.
- Developed computer vision models robust to occlusion, perspective distortion and environmental variability.
- Built predictive models combining temporal and visual features to estimate passenger flow and forecast peak demand patterns.

🎓 EDUCATION

PhD in Mechanical Engineering 09/2019

Federal University of Paraíba (UFPB)

MSc in Computer Science 01/2014

Federal University of Paraíba (UFPB)

BSc in Computer Science 12/2009

State University of Paraíba (UEPB)

CERTIFICATIONS AND ADDITIONAL COURSES

AI Engineer Agentic Track: The Complete Agent & MCP Course 2025

Udemy

AI Engineer Core Track: LLM Engineering, RAG, QLoRA, Agents 2025
Udemy

Build Generative Adversarial Networks 2023
Coursera

Natural Language Processing with Classification and Vector Spaces 2020
Coursera

Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization 2020
Coursera

Math for Machine Learning 2020
Amazon

Scrum Fundamentals Certified (SFC™) 2020
SCRUMstudy

Exploring and Preparing Your Data With BigQuery 2020
Google

Deep Neural Networks with PyTorch 2020
Coursera

SKILLS

- **LLM Systems and Generative AI:** Design and development of LLM-based applications using GPT, LLaMA and related models for conversational AI, knowledge assistants and decision-support systems
- **RAG and Context Engineering:** Development of retrieval-augmented systems with document parsing using LlamaParse, ingestion, chunking, retrieval, context assembly and response generation to improve grounding and relevance
- **Agentic Workflows:** Design of agentic and multi-agent workflows using LangGraph and CrewAI, enabling task decomposition, tool orchestration and context-aware decision-making
- **Prompt Engineering and Prompt Management:** Creation of structured prompts, reusable prompt templates and context strategies to improve consistency, accuracy and response reliability
- **Python and Backend Development:** Development of maintainable Python applications using OOP, modular design, FastAPI, Flask, REST APIs, debugging and production integration.
- **Vector Databases and Indexing:** Implementation of semantic search and indexing pipelines using embeddings, FAISS, pgvector and Azure AI Search for efficient information retrieval
- **Data Pipelines and Databases:** Design of ETL workflows and data pipelines for ingestion, transformation and preparation of structured and unstructured data, including SQL and NoSQL databases
- **Cloud and Production AI Systems:** Deployment and integration of scalable AI systems in Azure environments, using Docker, Kubernetes/AKS, monitoring, performance optimization and production-oriented engineering practices

LANGUAGES

English ●●●●○ French ●○○○○

AWARDS

Semifinalist in the Santander 2K17 Entrepreneurial University Program 10/2017
The objective of the contest is to encourage entrepreneurial ideas and projects from all over Brazil, mobilizing three target audiences: university students, startups, and microentrepreneurs.
http://www.ufpb.br/antigo/content/professor-e-estudantes-da-ufpb-s%C3%A3o-semifinalistas-do-empreenda-santander-2k17