Contact

+55 41 988502908 (Mobile)
ruanchaves.ai@gmail.com

www.linkedin.com/in/ruanchaves (LinkedIn)

Top Skills

Natural Language Processing (NLP)
Machine Learning
Python

Languages

English (Full Professional)
Portuguese (Native or Bilingual)

Certifications

From Data to Insights with Google Cloud
NVIDIA DLI Certificate – Fundamentals of Accelerated Computing with CUDA C/C++
Modernizing Data Lakes and Date Warehouses with Google Cloud

Honors-Awards

1st place at the II Evaluation of Semantic Textual Similarity and Textual Inference in Portuguese
University Council's Certificate of Honours
1st place at the Competition on Legal Information Extraction/ Entailment (COLIEE) 2021
1st place at the Aspect-Based Sentiment Analysis in Portuguese (ABSAPT @ IberLEF 2022)

Publications

Construção de Datasets para Segmentação Automática de Hashtags
Yes, BM25 is a Strong Baseline for Legal Case Retrieval

Ruan Chaves Rodrigues

Senior AI Engineer | Generative AI | Agents | RAG | LLMs | NLP | Python
Brazil

Summary

I architect and deploy production-grade Generative AI, Multimodal AI, and NLP systems. Most of my time goes into complex Agentic Systems and RAG architectures. I design multi-agent frameworks and autonomous agents, fine-tune and integrate LLMs, and own the full ML pipeline from prototype to production.

I hold a double Master's in AI & NLP (University of Malta / University of the Basque Country) and a Bachelor's in Computer Science. I've been working remote-first with distributed teams across the US, Europe, and LATAM for years, and I consistently deliver high-impact projects on schedule.

What I work with:

Agentic AI & Orchestration: Multi-agent frameworks, autonomous agents (LangGraph, LangChain, LlamaIndex), workflow automation (n8n)

Generative AI & LLMs: Fine-tuning, function/tool use, MCP, prompt engineering, Graph RAG, Multimodal AI, Transformers, Hugging Face, spaCy, Haystack

MLOps & Engineering: Docker, Kubernetes, Kubeflow, MLflow, A/B testing, model evaluation (LangSmith)

Cloud & Data: AWS (SageMaker, Bedrock), GCP (Vertex AI, BigQuery), Azure, Apache Spark, SQL, vector databases (Pinecone, Weaviate)

Core Stack: Python, PyTorch, TensorFlow, Keras, Scikit-learn, FastAPI, Node.js, React

Multilingual Transformer Ensembles for Portuguese Natural Language Tasks

Domain Adaptation of Transformers for English Word Segmentation

Deep Learning Brasil at ABSAPT 2022: Portuguese Transformer Ensemble Approaches

Feel free to reach out if you're working on agentic AI, RAG at scale, or getting GenAI into production.

Experience

Xseed Solutions
Senior AI Engineer
September 2025 - Present (7 months)
United States

Qive
Senior AI Engineer
February 2025 - August 2025 (7 months)
Brazil

- Led the development of a product search engine over proprietary data, reporting directly to the company’s founders in a fast-paced startup environment.

- Reduced RAG indexing time from several hours to minutes without compromising retrieval quality by uncovering reliable product groupings through data analytics and targeted indexing strategies.

- Increased RAG retrieval efficiency by achieving over 90% recall with only 20 candidates (down from 100), through a redesigned vector retrieval system using GCP tools and Gemini-based data augmentation.

- Conducted technical interviews for software engineering candidates, assessing coding skills, system design, and cultural fit.

- Mentored team developers to improve code quality, strengthen technical expertise, and support professional growth.

C6 Bank
1 year 5 months

Senior AI Engineer
May 2024 - February 2025 (10 months)
Brazil

- Led AI initiatives within a cross-functional team at one of Brazil’s largest digital banks, serving over 30 million clients; collaborated closely with backend, MLOps, data science, sales, and marketing teams to develop scalable, high-

impact AI solutions that drove business growth and enhanced customer experience.

- Built and deployed a Generative AI sales assistant using a multi-agent architecture with LangChain and LangGraph. The solution reduced both customer acquisition costs by a factor of six, improving sales team productivity.

- Led the full development lifecycle of AI agent systems, including stakeholder alignment, design, testing, and production rollout.

- Applied a range of Agentic AI techniques such as tool use, custom workflows, secure execution guardrails, few-shot learning, personas, model selection, and REACT strategies to enhance system performance and reliability.

Data Scientist
October 2023 - May 2024 (8 months)
Brazil

- Optimized a retrieval-augmented generation (RAG) system on Google Cloud Platform using advanced prompt engineering techniques and research into new GCP features, reducing inference costs by 90% and improving response accuracy beyond top competing LLM alternatives.

- Enhanced customer segmentation insights and enabled targeted marketing campaigns by developing time series forecasting and spending profile clustering models using BigQuery ML, resulting in more precise customer profiles.

Argilla
Data Science Intern
November 2021 - May 2022 (7 months)
Spain

- Worked closely with early members in a fast-paced startup (later acquired by Hugging Face).

- Pioneered the first production-ready implementation of state-of-the-art embedding-based annotation methods in the Argilla open-source library, contributing 68K+ lines of code.

- Translated cutting-edge academic research into scalable, robust features deployed in real-world NLP workflows.

- Authored tutorials and documentation that boosted user adoption.

Centro de Excelência em Inteligência Artificial (CExIA) & Deep Learning Brasil

AI Engineer

September 2020 - July 2021 (11 months)

Brazil

- Collaborated in a research team bridging academia and industry at CExIA, contributing to early-stage Transformer and language model research during the initial release of these technologies.

- Achieved 1st place twice in major NLP competitions open to both academic and industry participants:
• ASSIN 2 (2019): Led a stacking ensemble of multiple Transformer models for textual entailment, surpassing the runner-up by 0.7% in F1-score. Competition organized by the Brazilian Computer Society (SBC).
• ABSAPT (2022): Secured top accuracy in Aspect-Based Sentiment Analysis in Portuguese through meticulous hyperparameter tuning of Transformer models, beating the second-place team by 3.5%. Organized by the Spanish Society for Natural Language Processing (SEPLN).

- Delivered proof-of-concept NLP solutions to Brazilian enterprises including Copel (Forbes Global 2000), applying cutting-edge techniques in intent detection, named-entity recognition, sentiment analysis, entity linking, and early generative chatbot models.

NeuralMind

Data Scientist

February 2021 - May 2021 (4 months)

Brazil

Won 1st place out of 7 teams in the COLIEE 2021 textual entailment competition by applying zero-shot Transformer models without domain-specific adaptations, surpassing the runner-up by 6% in F1 score.

The Competition on Legal Information Extraction/Entailment (COLIEE) 2021 was organized by the Alberta Machine Intelligence Institute (Amii) at the University of Alberta.

Education

University of Malta

Master of Science in Human Language Science and Technology, Artificial Intelligence · (August 2021 - September 2023)

Universidad del País Vasco/Euskal Herriko Unibertsitatea

Master in Language and Communication Technologies, Artificial Intelligence · (August 2021 - September 2023)

Universidade Federal de Goiás

Bachelor's degree, Computer Science · (January 2017 - June 2021)