Ciro Villafraz

Buenos Aires, Argentina
Phone number: +54 1133031402
E-mail: contact@cirov.com
LinkedIn: https://linkedin.com/in/ciro-villafraz
GitHub: https://github.com/cvillafraz/

Machine Learning Engineer

I am a Machine Learning Engineer with 5+ years of experience in the tech industry, currently working in chatbots, NLP and LLMs. I have worked on multiple AI and Machine Learning projects with this stack of technologies: **Python, SQL, Docker, LangChain, AWS.** What I enjoy the most is building products using the power of **AI** with a focus on **AI Safety**.

Work Experience

AI Engineer at Harvard Business Publishing October 2024 to date

- Improved a multilingual Retrieval-Augmented Generation (RAG) chatbot supporting queries across 20K+ articles, with fine-grained retrieval over structured assets (articles, videos, podcasts) and enhanced source relevance ranking.
- Optimized prompt templates for query translation, classification, and evaluation, enhancing chatbot decision-making on retrieval invocation and parameter selection to improve response relevance.
- Designed and trained a custom content moderation model using pre-trained embeddings and a Bagging MLP classifier on 4.6K real and synthetic samples; achieved 89% balanced accuracy in predicting sensitive queries, improving system safety.

Main Technologies: Microsoft 365 Agents Toolkit, Python, FastAPI, Amazon Bedrock, Anthropic, Postgres (RDS), S3, Docker, Jenkins

AI Engineer at SimSkills August 2023 to October 2024

A role-playing simulator that uses AI to teach soft skills

- Built 90+ role play simulations to teach soft skills using prompt engineering, OpenAI and Anthropic models with LangChain. Out of +350 attempted simulations, users found them valuable 81% of the time. In addition, 8 of those simulations were built for the Madrid Community employment office (Oficina de empleo, Comunidad de Madrid)
- Created POCs for implementation of DSPy and agentic RAG with LangChain and Pinecone
- Created an LLM based scoring and feedback system to judge the performance of users after each simulation attempt.
- Developed a Cloud Function to export the user’s conversations (anonymously) to Google Sheets for further analysis and improvement.
- Implemented LangChain and LangSmith in production to improve LLM call observability, tracing

and debugging.

Main Technologies: Python, Numpy, Pandas, LangChain, DSPy, OpenAI, Anthropic, Node.JS, Firebase, GCP, Pinecone

Machine Learning Engineer at Anyone AI Jan 2023 to June 2023

- **Financial Advisor Chatbot**: Led a four-member team towards building a Question-Answering chatbot focused on Nasdaq companies. The chatbot can receive a question on the financial information of a certain company, and answer with factual information. Processed a +30 GB dataset, consisting of +9800 PDF documents.
- **Movie Review Sentiment Analysis:** Implemented sentiment classification on +50000 movie reviews, using techniques such as TF-IDF, Embeddings, and Logistic Regression. Achieved a 0.93 ROC AUC score on testing.
- **Vehicle image classifier**: Performed image classification on a dataset of vehicle images for 25 different make-models, achieving +85% testing accuracy by training a CNN using transfer learning with ResNet and EfficientNetB0 models.
- **Image classification API:** Created an API service for image classification using Flask and Docker, utilizing a ResNet50 model. Conducted load testing on the API using Locust.
- **Home Credit Risk Analysis**: Conducted risk analysis for home credit applicants, pre-processed and manipulated a +240000 rows long dataset. Achieved a +0.74 ROC AUC using supervised models such as Logistic Regression, Random Forest, and LightGBM.

Main Technologies: Python, Scikit-learn, Keras, LangChain, Haystack, Docker, ElasticSearch, HuggingFace, Pinecone.

Writing Contributor at Platzi April 2022 to July 2022

An online courses platform

- I wrote articles for the video lessons of +10 courses in Platzi. In those articles I transcribed, explained and expanded the concepts from each lesson.

Backend Developer at 321 Ignition Mar 2020 to Aug 2020

A website platform for car dealerships

- Created an end-to-end testing package for +10 websites using Cypress. Those websites are generated using a complex, custom bundler, and they share plenty of components. Furthermore, the team deploys a new website every 1-2 months. Hence, I had to create a package with tests that worked for the websites that were already in production, as well as those to be deployed in the future.
- Created a GitHub package that is used to share code across multiple backend services. Saved backend developers an uncountable amount of time to be wasted in copying such code.

Main Technologies: NodeJS, MongoDB, Docker, GitHub, Cypress.

Fullstack Developer as Freelancer at Talktomira May 2019 to Jun 2019

- Integrated the front end with the blog using the WordPress API
- Designed and developed the blog and article pages
- Implemented a blog carousel and a “Book Now” section on the landing page

Main Technologies: *React.Js, AWS, WordPress.*

Projects

The Ciro-verse Journal Mar 2022 to date

- Created 4 data science projects and wrote at least an article for each of them at cirov.com. Featured project: analysis of house rentals in Mexico City.

Main Technologies: *Python, Postgres, GitHub, Tableau.*

Skills

Tech Skills: Python, LangChain, DSPy, LLMs, SQL, JavaScript, AWS, Docker, *Cursor.sh*, SCRUM.
Languages: Spanish (native), English, French, Port.