Federico Ortega Riba

federico.ortegariba@fulbrightmail.org | linkedin.com/in/federico-ortega-riba | github.com/fede-ortega

AI/ML Engineer specialized in agentic systems and production pipelines, with experience in LLM integration, retrieval evaluation, and observability. MS in CS from CU Boulder (USA), Fulbright scholar.

Experience

AI Engineer Aug 2026 – Present
Sabio Group *Spain, Remote*

- Built **real-time voice agents** for the financial sector, orchestrating **Deepgram (STT)**, **ElevenLabs (TTS)** and **Twilio** telephony, and cut **p50 TTFA from 2s to 300ms**.
- Deployed **self-hosted open-weight LLMs** (DeepSeek-V4, GLM-5.2) to production with **vLLM on GKE (GCP)**, backed by **AWS S3** for storage.
- Developed **guardrails and evaluation frameworks** for text and voice agents with **Presidio** and **NeuralTrust** for safety, and **MLflow** for experiment tracking.

AI Engineer (Contractor) May 2025 – Jul 2026
Overwatch AI *Spain, Remote*

- Designed and orchestrated **multi-agent LangGraph systems** for a pre-seed startup, fetching live data from external APIs via **custom MCP servers** and coordinating specialized agents over **A2A**.
- Built a production **RAG system** with embeddings in **Qdrant**, deployed on **AWS** via **Terraform**, powered by closed- and open-weight models (**Claude, GPT, Gemini, GLM-5.2, Nemotron-Ultra, Llama**).
- Established **LLM evaluation** (DeepEval, Ragas, TruLens) and **OpenTelemetry observability** via Phoenix (Arize), cutting p50 latency from 12s to 7s and tokens from 12k to 8k.

AI/Machine Learning Engineer Feb 2025 – Apr 2026
University of Colorado Boulder *Colorado, USA*

- **Designed and deployed end-to-end retrieval pipelines** (LangChain/LlamaIndex + Azure OpenAI) with vector search, improving **Recall@5 from 85% to 90%** and answer accuracy by 7 pts via chunking and evaluation.
- Built scalable PyTorch/Hugging Face models for tool-use agents with **RLHF-based reward modeling**, improving task success rate from **74% to 85%** on a 400-task evaluation set.
- Refactored production LLM serving codebase (Jigsaw Agent + Mistral) for the **NSF National AI Institute for Student-AI Teaming**, cutting inference cost by **40%** via modular prompts and Langfuse observability.

Education

MS in Computer Science (Natural Language Processing) Aug 2024 – May 2026
University of Colorado Boulder *USA*

- GPA: 4.0; fully funded by the **Fulbright Scholarship** (USDOS).
- Coursework: *Machine Learning*; *Deep Learning & Neural Networks*; *Advanced NLP*.

MA in Computational Linguistics Sep 2023 – Jul 2024
Complutense University of Madrid *Spain*

- GPA: 3.5; Coursework: *Language & Speech Technologies*; *Advanced Programming*; *Statistics with R*.

Technical Skills

Languages & AI: Python, SQL, TypeScript, LangGraph, LangChain, LlamaIndex, ReAct, RAG, prompt eng
Cloud & MLOps: AWS, Terraform, Docker, CI/CD, Qdrant, FAISS, REST APIs, microservices
ML & Evaluation: PyTorch, Hugging Face, BERT/T5, DeepEval, Ragas, OpenTelemetry, Phoenix (Arize)
Spoken Languages: Spanish (native), English (C2), French (C1), Chinese (HSK5)

Projects & Publications

1. CUCLASIC at SemEval-2026 Task 5: LLM Prompting Strategies for Rating Ambiguous Word Senses *Proceedings of SemEval-2026 (ACL).* Paper — GitHub Repo 2026

2. Complex Word Identification for Lexical Simplification in Spanish Texts for Patients *Spanish Society for Natural Language Processing.* Paper — GitHub Repo 2025

3. CLARA-MeD Tool: A System to Help Patients Understand Clinical Trial Announcements and Consent Forms in Spanish *Studies in Health Technology and Informatics.* Paper — GitHub Repo 2024

Awards

Fulbright Spain Graduate Scholarship (USDOS) — one of 13 Spanish graduate scholars 2024 – 2026