

INÉS ARANA BACHELIS

+34 669 63 75 89 • inesaranabach@gmail.com • [linkedin.com/in/inés-arana-bachelis-ai](https://www.linkedin.com/in/inés-arana-bachelis-ai) • github.com/inesaranab

PROFESSIONAL SUMMARY

AI Engineer specializing in NLP/NLU and the deployment of agentic workflows with full LLMOps - alignment, observability, and evaluation. I build production AI systems oriented toward measurable business impact, not technical novelty: I automate what slows teams down and focus engineering effort where it delivers the most value. I design agentic architectures end-to-end - from orchestration and memory through guardrails, gateway, and evaluation - and implement them from scratch, ship fast, and navigate uncertainty. My background in international finance means I think in outcomes, not just outputs - evaluation, explainability, and transparency are non-negotiable.

WORK EXPERIENCE

Aily Labs

May 2026–Present

AI/ML Engineer

- Designed and built the feedback loop behind the main agent: retrieval over a Neo4j knowledge graph serving the semantic layer, PostgreSQL schemas for memory and feedback, and an Airflow DAG migrating feedback between the two.
- Build several skills that improve business function productivity such as powerpoint skill with agentic visual loop.
- Built memory module and memory MCP tool connected to the main agent and a custom Slack agent that stores and classifies user memories at multiple granularities - by tenant, module, user, and scope.
- Added multi-modality to all the MCP tools used by the frontend team so they could streamline the process of detecting bugs.
- Built an observability dashboard tracking time-to-first-token, latency and error rate across Bedrock models, surfacing provider-side degradation that was otherwise invisible to the teams consuming the gateway.
- Designed and implemented MCP governance: migrated MCP tools from local to remote (Kubernetes) behind three layers of security: Auth: user tokens authenticated through the LiteLLM gateway, minting a signed JWT and an M2M token for pod communication; Downstream services: a third trust layer for integrations such as Atlassian and Microsoft; Secrets: the pod assumes an IRSA role to retrieve secrets from AWS.
- GDPR Article 9 screening service (sole engineer, design to production): Removes special-category personal data from hiring assessments before any model sees it; Benchmarked five detectors on two labelled datasets, one built for the purpose. Self-hosted Gemma-4-31B won on measured F1 (0.786 vs 0.755); hybrid guardrail combining semantic and rule-based detection; Async request-reply on Azure Container Apps: API, queue-triggered worker, and a scale-to-zero GPU detector (vLLM, internal ingress) as three independently scaled services - €0.17/day idle; Assessment model routed through Portkey, OpenRouter, Gemini, so a provider swap is an env var, not a code change; DeepEval harness with LLM-as-judge on fixed goldens; built the Azure CI/CD pipeline from scratch.

dive.tech

Sep 2025–May 2026

AI Engineer

- Owned LLMOps across end-to-end agentic applications - observability and evaluation with LangSmith, Langfuse, and Arize.
- Built agentic automation with LangChain and LangGraph for SAP payments reconciliation at ISPD Group.
- Developed a LangGraph multi-agent research recommender to surface strategic product insights at Lida Research.
- Owned BBVA's annotation pipeline: experiments assessing brand reputation using Gemini structured outputs, with a versioned library for prompt and schema versioning.
- Automated the Los40 Principales 2026 Awards marketing campaign - a multi-modal (image + text) multi-agent system on Vertex AI. Implemented inpainting and semantic masking for brand-consistent asset generation, plus a custom MCP client/server for tool integration.
- Designed a few-shot visual retrieval system for logo identification in video (ISPD Group): trained a ResNet-34 with a custom Prototypical Loss for multi-modal embeddings, indexed in Qdrant for real-time similarity search. Curated a 2,000+ image dataset and built an evaluation pipeline reaching >90% retrieval accuracy.

Council Office of Laredo

Mar 2024–Sep 2025

AI Engineer

- Built a multi-modal, multi-agent RAG system with semantic search (OpenAI, LangChain, LangGraph) deployed on Railway; validated and evaluated with RAGAS.
- Fine-tuned models with RLHF to mitigate bias and improve accuracy.

- Developed a transcription + diarization pipeline to automate document retrieval and classification.
- Built a sentiment-analysis pipeline to gauge public perception of local policies.
- Designed a vector database and web interface for real-time AI-driven services; owned the end-to-end process from system design through deployment.

Colgate

Jun 2023–Sep 2023

Business Analyst Intern

- Analyzed data to identify key business trends; delivered insights to senior management via Power BI.
- Designed and led surveys of regional managers, improving the ticketing tool and its interface.
- Built and maintained BI databases (Microsoft Access, Google Sheets, Excel).

SESE

Jun 2021–Sep 2021

Imports & Exports Intern

- Coordinated and tracked import/export freight shipments, ensuring timely delivery and accurate documentation.
- Reviewed shipping documents, invoices, and purchase orders for compliance; managed freight inventories.
- Collaborated with software teams to test and improve freight-tracking and inventory systems.

PROJECTS

Multi-Agent Reasoning Research - Stanford AI Professional Program

Mar 2025–May 2025

- Designed a retrieval-augmented ReAct agent (DSPy + GPT-4o-mini) that externalizes all numeric computation to a dedicated financial calculator.
- Boosted end-to-end QA accuracy from 53.3% → 89.8% on the DOCFINQA dataset via a hybrid BM25 + dense-embedding retriever with a custom chunking strategy.
- Ran detailed error analysis, eliminating arithmetic errors (–36.5 pp) and improving transparency in high-stakes financial reasoning.

EDUCATION

UDIT

2025

MSc in Artificial Intelligence

Universidad Camilo José Cela

2024

BSc in Business & Technology (Specialty: International Finance), Graduated with Honors

CORE SKILLS

- Business & Finance
 - Agent Skills
 - Claude & Codex
 - Agents & Orchestration
 - LangGraph
 - LangChain
 - DSPy
 - n8n
 - MCP Client/Server
 - MCP Governance
 - Multi-Agent Systems
 - ReAct
 - Plan-and-Execute
 - Reflection
 - FastAPI
 - Pydantic
- Airflow
 - Retrieval & Memory
 - Qdrant
 - Neo4j
 - PostgreSQL
 - Hybrid BM25 Retrieval
 - LLM Gateways (LiteLLM, Portkey, OpenRouter)
 - Guardrails (Semantic + Rule-Based)
 - Security
 - CI/CD
 - Docker
 - Kubernetes
 - Evaluation Harness Engineering
 - DeepEval
 - Braintrust
 - Langfuse

CERTIFICATIONS

Generative AI Developer Professional, AWS; Certified AI Engineer, AI Makerspace; Agentic Bootcamp, Data Science Dojo; Machine Learning Specialization, Stanford University; Statistics for Data Science (Tech 105), Stanford University; Natural Language Understanding (XCS224U), Stanford University; Deep Learning Specialization, Coursera; Natural Language Processing Specialization, Coursera; Generative AI with Large Language Models, Coursera; AI for Good Specialization, DeepLearning.AI; Azure AI Fundamentals, Microsoft; Google Cloud Digital Leader, Google; Power BI, Microsoft

LANGUAGES

Russian (Native), Spanish (Native), English (Advanced)