Alejandro Garay

AI Engineer (NLP / RAG / Agentic Systems)

Buenos Aires, Argentina | alejandroa.garay.ag@gmail.com | +54 11 5660 5331 | GitHub: github.com/naaas94 | LinkedIn: linkedin.com/in/alejandro-garay-frontini

SUMMARY

AI/NLP engineer with a linguistics background and systems-first approach. Builds end-to-end NLP services and retrieval workflows (classification, semantic search, local-first RAG), with an emphasis on auditability, evaluation harnesses, and production patterns (CI/CD, reproducibility, observability).

CORE SKILLS

- LLM/NLP: Transformers, embeddings, semantic search, RAG (dense + lexical), re-ranking, prompt/tool workflows, FAISS/HNSW
- Backend/Serving: FastAPI, REST, Pydantic, Docker, basic security guardrails (validation, rate/timeout control)
- Data/Orchestration: BigQuery/GCS, Flyte, Ray, Apache Beam, Spark; schema contracts and data validation
- MLOps/Quality: GitHub Actions CI, MLflow experiment tracking, deterministic runs, offline eval harnesses, structured logging/metrics

EXPERIENCE

Spotify — Data Scientist, Customer Experience & Privacy

Remote / Buenos Aires | Sep 2024 – Jul 2025

- Designed and delivered an NLP classification pipeline supporting GDPR/CCPA compliance workflows for privacy-sensitive content categorization.
- Built an internal semantic retrieval capability (transformer embeddings + FAISS) for knowledge base search; defined indexing and freshness policies.
- Partnered with Legal and Operations on taxonomy and labeling strategy, translating regulatory requirements into system constraints and evaluation checks.
- Prototyped a symbolic-LLM hybrid “agentic reviewer” to audit model/agent behavior and generate traceable review artifacts.

Spotify (via ModSquad) — Workforce Management Analyst (Data/Analytics)

Buenos Aires | Jan 2022 – Sep 2024

- Owned analytical reporting and forecasting for global customer support operations (SLA tracking, capacity planning, performance monitoring).
- Built internal dashboards and operational tooling; collaborated cross-functionally with Customer Experience, Data, and Product teams.

Invisible Technologies / MGM Linguistic Solutions — Assistant Project Manager (NLP Annotation & Linguistic QA)

Remote | 2020 – 2022

- Managed NLP annotation workflows and linguistic QA for data labeling projects; enforced quality guidelines and review loops.
- Automated language processing tasks for data preparation and validation pipelines.

SELECTED ENGINEERING PROJECTS

Open-source/portfolio evidence: github.com/naaas94

Local-First RAG System (PoC, portfolio-grade)

- Hybrid retrieval (dense + BM25) with Reciprocal Rank Fusion (RRF), stable document/chunk IDs, and offline generation via Ollama.
- Security hardening (signed indices), resource limits, timeout/retry logic, multi-level caching, and structured JSONL tracing.
- Offline evaluation harness (HitRate@K) with reproducible datasets.

Agentic Reviewer (classification auditing demo)

- LLM-assisted semantic auditing loop that reviews predictions, proposes corrections, and writes traceable artifacts (CSV/JSON/Markdown).
- Output validation, prompt-injection detection, caching/pooling, and reproducible benchmark mode.

QuickCapture: Symbolic Ingestion Layer (SNR ecosystem)

- Local capture UI + FastAPI “brain” with Ollama parsing, semantic validation, hybrid SQLite + FAISS storage, and offline fallbacks.
- Observability modules (health, metrics, drift detection) for production-style debugging.

EDUCATION

B.Sc. Technical-Scientific Translation; B.Ed. English Language Teaching — Lenguas Vivas “Sofía E. Broquen de Spangenberg”

CERTIFICATIONS

- IBM Data Science Professional Certificate (Coursera)
- Google Data Analytics Professional Certificate (Coursera)
- DataCamp: SQL, Python, NLP track