Alejandro Garay
AI Solutions Architect
Buenos Aires, Argentina · alejandroa.garay.ag@gmail.com · +54 11 5660 5331
GitHub · LinkedIn

Summary

AI/NLP Solutions Architect with a linguistics background who turns ambiguous business goals into auditable AI services. I design and ship end-to-end systems: problem framing → pipeline → delivery; centered on LLM-based systems, information retrieval, ranking, context assembly, verification, and monitoring. I build maintainable prototypes hardened just enough for production handoff, with evaluation harnesses and safety/latency/cost guardrails. I map needs to proven patterns.

Flagship Architecture Work

Retail Copilot NL→SQL Architecture (PoC→MVP→Multi-tenant, GCP/Vertex)
Independent AI Solutions Architect (Challenge project, architecture/specification)

- Authored a 35+ page architecture SOW for a GCP-based Retail Copilot, converting natural language into validated SQL + VizSpec JSON over BigQuery, outlining a clear PoC→MVP→multi-tenant evolution path (blueprint/spec, not deployed).
- Defined an NL→intent→slots→template→validator chain with allowlists, tenant filters, dry-run bytes, and budget caps, ensuring safe-by-design query execution (incl. example planner JSON, SQL templates, and validation rules).
- Designed tenancy and blast-radius model from day 1: per-tenant datasets, logs, budgets, and service accounts; platform-shared orchestrator with tenant-scoped execution via service-account impersonation and RLS/CLS.
- Specified MLOps & evaluation surfaces for LLM-based systems: golden-set replay, structural checks, promotion gates, canary rollout, SLOs (latency, cost, faithfulness), rollback policies, and concrete evaluation metrics (SQL correctness, execution accuracy, robustness, faithfulness, contradiction rate).

Technical Skills

- **Architectural patterns:** Hybrid RAG (BM25 + dense), re-ranking (cross-encoders), context assembly, verification/attribution, offline/online eval (nDCG, faithfulness), data & index versioning, freshness policies, drift/safety filters.
- **Languages & Infra:** Python, SQL, Docker, GitHub Actions, MLflow, Ray, Flyte, Apache Beam, Spark, BigQuery/GCS.

- **NLP/LLM:** Transformers/SentenceTransformers, embeddings & FAISS/HNSW, FastAPI services, classification, semantic search.
- **Quality & Ops:** CI/CD pipelines, schema validation (Pydantic), config management (Hydra), experiment tracking, basic observability hooks.

Experience

Spotify — Data Scientist, Customer Experience & Privacy | Sep 2024 – Jul 2025

- Architected and delivered a privacy-compliant hybrid NLP classification pipeline for GDPR/CCPA workflows; translated regulatory constraints into taxonomy, data, and evaluation requirements. Full ownership from discovery to delivery.
- Delivered a semantic retrieval capability (transformer embeddings + FAISS) for internal knowledge search; defined indexing and freshness policies.
- Partnered with Legal & Ops to turn policy into system guarantees (labeling, redaction paths, safety thresholds).
- Prototyped a symbolic-LLM “agentic reviewer” for systematic agent performance evaluation and auditability at scale.

Spotify (via ModSquad) — Data Analyst | Jan 2022 – Sep 2024

- Built forecasting/monitoring tools for global CX operations; created dashboards for SLA and capacity management; facilitated cross-team delivery with Product/Data.

Invisible Technologies / MGM Linguistic Solutions — PM | 2020 – 2022

- Managed NLP annotation and linguistic QA workflows; automated data prep/validation for downstream models.

Engineering Portfolio (GitHub)

- **PCC**: Privacy-compliant content classifier + validation workflows. Pattern: policy-to-pipeline; evaluation gates before release. Stack: Flyte, BigQuery, MiniLM, Docker.
- **data-pipeline**: Synthetic data generation with automated schema validation and reproducible splits. Stack: Beam, Spark, Ray, Kafka.
- **model-training-pipeline**: Deterministic training with MLflow tracking and pinned configs. Stack: scikit-learn, MLflow, Python.
- **simple-model-api**: FastAPI service with containerized deploy and CI. Stack: FastAPI, Docker, GitHub Actions.
- **agentic-reviewer**: LLM-assisted semantic auditing and post-hoc evaluation. Stack: FAISS, Transformers, Python.

Education

B.Sc. in Technical-Scientific Translation · B.Ed. in English Language Teaching — Lenguas Vivas "Sofía E. Broquen de Spangenberg".

Certifications

IBM Data Science Professional Certificate · Google Data Analytics Professional Certificate · DataCamp (SQL, Python, NLP).