Gerardo Toboso

AI Backend Engineer | LLM Systems & Data Infrastructure

Buenos Aires, Argentina | gerardotoboso1909@gmail.com | 011 15-4045-6207

linkedin.com/in/gerardo-toboso-512a48290 | github.com/Gerardo1909

Summary

AI Backend Engineer specialized in designing and deploying production-ready LLM systems and scalable data infrastructures. Strong background in modular RAG architectures, LLM API integration, backend system design, and observability-aware engineering. Teaching Assistant in Bayesian Causal Inference with solid foundations in probabilistic modeling and uncertainty quantification. Focused on building robust, testable, and maintainable AI-driven backend services aligned with clean architecture principles.

Skills

Backend & API Development: Python (advanced), FastAPI (API-ready architecture), RESTful design, Dependency Injection, and Clean Architecture principles

LLM & AI Systems: RAG architectures, LLM API integration (Groq/OpenAI-compatible), Prompt engineering with contextual grounding, Vector databases (ChromaDB), Embedding pipelines (Sentence Transformers), and Similarity search and filtering strategies

Data Infrastructure: SQL (DuckDB, analytics modeling), NoSQL concepts, PySpark, Airflow orchestration, dbt-core, and AWS S3

DevOps & Testing: Docker & Docker Compose, Pytest with coverage, Structured project layout, and Environment configuration management

ML Foundations: Bayesian inference, Probabilistic modeling, MCMC & hierarchical models, and Uncertainty quantification

Experience

AI Backend Engineer, RAG Systems - Causalito AI Assistant (Personal Project) – GitHub Dec 2025 – Feb 2026

- Designed and implemented a modular RAG-based backend system indexing academic PDFs and generating grounded responses via LLM APIs.
- Architected full ingestion → indexing → retrieval → generation pipeline with strict separation of concerns.
- Integrated external LLM APIs (Groq/OpenAI-compatible) with custom prompt orchestration to reduce hallucinations.
- Implemented deterministic ID strategy for reproducible indexing and consistency.
- Developed similarity search infrastructure using vector embeddings (Sentence Transformers + ChromaDB).
- Built structured test suite (unit + integration) with ∼70–80% coverage for critical modules.

Undergraduate Teaching Assistant, Bayes Plurinacional – GitHub Dec 2025 – present

- Developed technical documentation and courseware on probabilistic modeling, MCMC, and hierarchical models.
- Translated advanced Bayesian theory into applied workflows for uncertainty quantification.
- Guided students in model comparison and credible interval interpretation.
- Simplified complex probabilistic systems into actionable engineering concepts.

Education

Universidad Nacional de San Martín (UNSAM), BSc in Data Science July 2022 – present

- 75% completed — GPA: 9.0 / 10