

Santiago Bolanos Vega

santiagbv@gmail.com | (412) 589-7464 | linkedin.com/in/9sntg | github.com/9sntg | sntg.me

EDUCATION

Carnegie Mellon University Master of Information Systems Management

- GPA: 4.0/4.0, Highest Distinction.
- Teaching Assistant: Distributed Systems (3 semesters, Head TA for 2).
- Relevant Coursework: Machine Learning in Production, Machine Learning, Deep Learning, Applications of Natural Language Processing.

Pittsburgh, PA
Aug 2024 – Dec 2025

Universidad de Los Andes B.Sc. Systems and Computer Engineering

Bogota, Colombia
Aug 2017 – Dec 2021

PROFESSIONAL EXPERIENCE

Groupon

AI Builder, CEO Office

Chicago, IL
May 2026 – Present

- Built Groupon's outbound lead generation engine end to end with AI-assisted development, using LLMs to write per-lead personalized emails across 2 verticals and 5 merchant categories, adding **8,000+** leads at **\$0.45** each. (**Apify, Apollo, Postgres, GCP**)
- Built a learning loop that scored each campaign on open, reply, and interest rates and lead volume, then automatically tuned subject lines, the winning email structure per category, and sender persona and tone. (**LLMs**)
- Scaled sending to **15,000** emails per week, **20x** a single sales rep, holding industry-benchmark open and reply rates through the inbox warmup and domain health infrastructure that kept deliverability viable at that volume. (**Smartlead, AgentMail, Inboxkit**)
- Worked under a CEO Office mandate to go from problem to shipped system in weeks, partnering directly with the CEO and C level commercial leadership to diagnose opportunity areas and build AI solutions against them.

Atria

AI Engineer Intern

Pittsburgh, PA
Jun 2025 – Aug 2025

- Cut new-client onboarding time **20x** with a context management system where granular, versioned instruction units stack dynamically into each agent's full context, so an instruction update propagates instantly across 100+ deployments. (**Node.js, React, MongoDB, OpenAI**)
- Replaced manual QA with a simulation agent that runs fully parallelized synthetic conversations, removing the human bottleneck and cutting QA time per conversation by **62.5%**. (**Node.js, OpenAI, OpenRouter**)
- Built a reviewer agent that scored each simulated conversation against configurable criteria and ran the pipeline on AWS Lambda, expanding QA from spot-checks to **100%** of production agents with continuous drift monitoring. (**Node.js, OpenAI, OpenRouter, AWS Lambda**)
- Cut product-inquiry response time from **2 hours to minutes** for a pharmaceutical machinery client with a RAG agent over a 2,000+ product catalog, replacing manual document search with semantic retrieval. (**OpenAI Embeddings, FAISS**)

Bip Consulting

IT Consultant

Bogota, Colombia
Oct 2021 – Jul 2024

- Delivered 11 client-facing technical projects across capital markets, banking, insurtech, microfinance, automotive, and infrastructure, and supported business development with 17 proposals at a **35%** win rate.
- Delivered a **zero-downtime** migration of 16,000+ customer, account, and card records in 3 hours with a **100%** reconciliation pass rate, using ETL pipelines with automated validation and compliance-ready reporting. (**Python, SQL, Core Banking Systems**)
- Designed an Azure data architecture with automated ingestion, transformation, and monitoring pipelines to unify fragmented data sources at Colombia's capital markets regulator, supporting surveillance of **10,200+** daily market operations. (**Azure, Databricks**)
- Defined an analytics roadmap for fraud detection and monitoring that the institution later reported cut false positives by **70%** after implementation, reducing manual review load. (**Python, pandas, scikit-learn, XGBoost, SQL, Power BI**)
- Ran technical due diligence on a Latin American insurtech platform for an investment fund, reviewing architecture, APIs, data models, and cloud readiness to surface scalability and resilience gaps for a potential **\$5M+** Series A. (**Azure, RabbitMQ, REST APIs**)
- Designed a digital transformation roadmap for a Colombian microfinance bank serving **220,000+** clients across a **\$320M+** portfolio, defining 4 initiatives including an AI-powered loan approval model, based on competitor benchmarking of digital maturity gaps.

PROJECTS

Dungeon Agents: Multi-Agent LLM Coordination System

Spring 2026

- Built a multi-agent dungeon simulation where two LLM agents coordinate under partial observability to reach a shared goal within a turn budget, with tool-use loops, inter-agent messaging, and action validation in a custom engine. (**LangChain, Claude, Python**)
- Instrumented the pipeline with Langfuse tracing and exported per-run artifacts (a human-readable report, JSON summaries, and a Streamlit dashboard) so a reviewer can audit any decision and trace coordination failures without re-running it.

Retrieval-Augmented Generation: Pipeline Experimentation and Domain Application

Fall 2025

- Developed and compared naive and enhanced RAG pipelines over 4,118 passages with query rewriting, confidence scoring, and adaptive retrieval, reaching a **5.6%** Exact Match gain and **62.5%** faithfulness on RAGAs. (**OpenAI, Embedding Models, Milvus, RAGAS**)
- Assembled a RAG-based financial learning platform over a curated investment corpus, serving source-cited responses through an interactive interface and achieving **73.7%** answer relevancy across 40 evaluation questions. (**LangChain, OpenAI, Milvus, Streamlit, RAGAS**)

End-to-End MLOps Platform for Real-Time Air Quality Forecasting

Fall 2025

- Built a containerized MLOps platform for real-time air quality forecasting, with Kafka for streaming, FastAPI for serving, MLflow for tracking and registry, and Evidently for drift monitoring, orchestrated via Docker Compose. (**Kafka, FastAPI, MLflow, Evidently, Docker**)
- Compared Random Forest, XGBoost, and ARIMA on air quality time series with temporal splits, promoting the best model to production via MLflow and validating pipeline integrity with drift detection. (**Python, XGBoost, MLflow, Scikit-learn**)

Production Movie Recommendation Service with MLOps and Continuous Experimentation

Spring 2025

- Built and operated a production movie recommendation service for **1M+** users across 20,000 movies, training an SVD collaborative-filtering model on ratings ingested via Kafka, with automated retraining. (**Python, Kafka, Surprise, PostgreSQL, MLflow, Flask**)
- Deployed a fully containerized inference stack with A/B testing, Prometheus and Grafana dashboards, and automated alerts for accuracy degradation and traffic spikes, supporting zero-downtime model updates. (**Docker, Prometheus, Grafana**)

SKILLS

- **AI & LLM Tooling:** OpenAI API, Anthropic Claude, LangChain, RAG Pipelines, Prompt Engineering, LLM Evaluation, FAISS, Milvus, Semantic Search, OpenRouter, MCP, Langfuse, Claude Code, Cursor
- **Machine Learning & MLOps:** scikit-learn, XGBoost, Random Forest, ARIMA, pandas, MLflow, Kafka, Evidently, Prometheus, Grafana, A/B Testing, Drift Monitoring, Time-Series Forecasting, Recommender Systems
- **Languages & Frameworks:** Python, SQL, JavaScript, TypeScript, Java, Node.js, React, FastAPI, Flask, REST APIs, Streamlit
- **Cloud, Infra & Data:** GCP, AWS (Lambda), Azure, Docker, Git, PostgreSQL, MongoDB, Supabase, Databricks, Power BI
- **Lead Generation:** Apollo, Smartlead, Salesforce

ACCOMPLISHMENTS

- Graduated with Highest Distinction and Academic Excellence Scholarship, Carnegie Mellon University (Dec 2025).
- COLFUTURO Loan-Scholarship, Academic Merit and Professional Excellence (May 2024).
- Top national score in Colombia's standardized graduation exam, recognized by the Ministry of Education (Dec 2021).