# JUAN ANDRES BERNAL GIL

Software Engineer — AI Systems | LLM Orchestration | Full-Stack | M.Sc. Computer Engineering

juan.bernal.2004.gil@gmail.com | linkedin.com/in/juan-andres-bernal | github.com/juanbernal13 | Bogota, Colombia (Open to Remote)

## PROFESSIONAL SUMMARY

Software Engineer specializing in AI systems design, LLM orchestration, and full-stack product delivery. Dual B.S. in Systems Engineering and Industrial Engineering, pursuing M.Sc. in Computer and Systems Engineering. Production experience building multi-model AI pipelines (Claude, GPT-4, Gemini), RAG architectures, Graph Neural Networks, and semantic search systems (Qdrant, Neo4j). Strong backend foundation in Java Spring Boot, Python (FastAPI), Node.js, and scalable cloud infrastructure (AWS, Docker, Kubernetes). Comfortable owning the full AI stack — from data ingestion and model integration to frontend dashboards and CI/CD deployment.

## PROFESSIONAL EXPERIENCE

### Software Engineer Associate | Scotiabank

*February 2026 - Present*

- Engineered mission-critical backend microservices (Java Spring Boot, Hexagonal Architecture) for international trade operations across Latin America, serving high-throughput financial workflows under strict SLA and compliance requirements.
- Integrated SWIFT GPI API enabling real-time worldwide transaction tracking via UETR references, adding semantic observability to distributed payment flows across correspondent banks.
- Built React-based interfaces for the Securities platform (equities, fixed income, bonds), handling millions of dollars in assets daily; implemented end-to-end test coverage with Playwright across complex multi-step financial journeys.
- Automated deployment pipelines through Jenkins and ArgoCD (CI/CD), reducing release cycles and improving environment parity across development, staging, and production.
- Containerized microservice evaluation environments with Docker, validating cross-service data consistency and debugging complex distributed-system logic before production releases.

**Tech Stack:** *Java, Spring Boot, React, Hexagonal Architecture, AWS (Lambda, S3, CDK), Docker, Jenkins, ArgoCD, GitHub Actions, REST APIs*

### AI Systems Engineer | Nebula Medical

*August 2025 - February 2026*

- Architected a multi-model LLM orchestration layer using LangChain, routing clinical queries across Claude, GPT-4, and Gemini based on task type, latency budget, and confidence thresholds — reducing hallucination rate by ~42% through iterative prompt engineering and structured output validation.
- Designed and deployed a RAG (Retrieval-Augmented Generation) pipeline over a multi-source medical knowledge base, combining vector search (Pinecone) with structured metadata filters to achieve precise, citation-grounded AI responses.
- Built automated AI evaluation pipelines (Python, pytest, async/await) benchmarking model output quality against 7 criteria including factual accuracy, latency (p99 < 500ms), consistency, and edge-case robustness.
- Developed React/Next.js dashboards visualizing AI model performance analytics, failure pattern distributions, and evaluation trends — enabling data-driven iteration on model selection and prompt strategies.
- Engineered HIPAA-compliant data access layer (Prisma ORM, PostgreSQL) with JWT/OAuth2 authentication, RBAC authorization, and end-to-end encryption across all patient data flows.

**Tech Stack:** *Python, LangChain, RAG Systems, Pinecone, Prompt Engineering, Next.js, Node.js, Prisma ORM, PostgreSQL, TypeScript, Docker, AWS Lambda*

### Software Developer I & Intern | Caseware

*June 2024 - July 2025*

- Engineered integrations within the imports/bindings module of the Caseware platfor. A mission-critical data ingestion layer used by 475,000+ professionals across 130 countries — ensuring schema-compliant, reliable data mapping at enterprise scale.
- Developed backend microservices in Java Spring Boot and Node.js; deployed serverless infrastructure on AWS via CDK, reducing operational costs and accelerating project launches from development to production.
- Led massive data migrations (500k+ records) ensuring transactional atomicity, improving data access latency by 40% through query optimization, indexing, and caching strategies.
- Built reactive Angular components with lazy loading and bundle optimization, cutting application load time by 30%; configured orchestrated Kubernetes deployments.
- Led Scrum ceremonies for a 7-developer team, maintaining continuous delivery across sprints with backlog refinement and stakeholder alignment.

**Tech Stack:** *Java, Spring Boot, Node.js, Angular, AWS CDK, Kubernetes, DynamoDB, S3, Docker, GitHub Actions, Jenkins, Agile/Scrum*

## EDUCATION

**M.Sc. Systems & Computing Engineering** | *Universidad de los Andes* | Bogota, Colombia | Next — August 2026 | Expected: December 2027

**Relevant Coursework:** *Advanced Algorithms, Reinforcement Learning, Machine Learning Techniques, Metaheuristics, Programming Paradigms*

**B.S. Systems and Computing Engineering** | *Universidad de los Andes* | Bogota, Colombia | Jan 2021 - June 2025 | GPA: 4.21/5.0

**B.S. Industrial Engineering** | *Universidad de los Andes* | Bogota, Colombia | Jan 2022 - June 2025 | GPA: 4.21/5.0

**Relevant Coursework:** *Advanced Optimization, Probabilistic Models, Stochastic Processes, Discrete Simulation, Machine Learning*

## KEY PROJECTS & RESEARCH

**Aura AI: Intelligent Talent Matching Platform** [link] -- End-to-end B2B AI recruitment platform that replaces keyword search with a hybrid architecture combining semantic vector search (Qdrant) and a Knowledge Graph (Neo4j). Implements Graph Neural Network models (PyTorch) for skill-gap analysis and an Explainable AI (XAI) ranking layer. Orchestrated async processing via Celery/Redis; full RESTful backend in FastAPI with a Next.js frontend.

*Tech: FastAPI, Next.js, PyTorch (GNN), Neo4j, Qdrant (Vector DB), PostgreSQL, MongoDB, Redis, Celery, LangChain, Docker*

**Multi-Model LLM Evaluation Framework** -- Python framework for systematically benchmarking Claude, GPT-4, and Gemini across complex reasoning scenarios. Measures hallucination rate, multi-step accuracy, and response consistency against 7 quality criteria. Includes async orchestration, structured scoring, and iterative prompt refinement feedback loops.

*Tech: Python, pytest, async/await, LangChain, Prompt Engineering, Docker, subprocess*

**Deep Learning Benchmark — Biological Pattern Prediction** [link] -- Rigorous ML pipeline with cross-validated model evaluation, PCA preprocessing, systematic hyperparameter search, and edge-case analysis achieving 90%+ accuracy on complex biological classification tasks. Published research output.

*Tech: Python, TensorFlow, Keras, scikit-learn, Pandas, NumPy, pytest*

**Real-Time Financial Alert System (WebSocket + Event-Driven)** [link] -- Full-stack event-driven system with NestJS backend, Redis pub/sub message brokering, PostgreSQL persistence, WebSocket real-time delivery, and React frontend — fully containerized with Docker and deployed via GitHub Actions CI/CD.

*Tech: TypeScript, NestJS, React, Redis, PostgreSQL, WebSocket, Docker, GitHub Actions*

## TECHNICAL EXPERTISE

**AI & LLM Engineering:** LangChain, Multi-Model Orchestration (Claude, GPT-4, Gemini), RAG Pipelines, Prompt Engineering, Vector Search (Pinecone, Qdrant), Graph Neural Networks (GNN), XAI, AI Evaluation & Benchmarking

**ML & Data Science:** PyTorch, TensorFlow, Keras, scikit-learn, XGBoost, LightGBM, NLP, Deep Learning, Feature Engineering, MLflow

**Backend:** Python (FastAPI, Django), Java (Spring Boot/Cloud), Node.js (NestJS/Express), TypeScript, REST APIs, GraphQL, Microservices, Hexagonal Architecture

**Frontend:** React, Next.js, Angular, TypeScript, JavaScript, HTML/CSS

**Data & Storage:** PostgreSQL, MongoDB, Neo4j, Qdrant, Redis, DynamoDB, Prisma ORM, Vector Databases

**Cloud & DevOps:** AWS (Lambda, S3, CDK, EC2, EKS, SQS), Docker, Kubernetes, Terraform, GitHub Actions, Jenkins, ArgoCD, CI/CD, Serverless

**Languages:** English (Advanced) | Spanish (Native)