Julian Canales

832-931-1132 | juliancanales@utexas.edu | San Francisco, CA | GitHub | LinkedIn

Education

The University of Texas at Austin — M.S. Artificial Intelligence (online) Aug 2025 – May 2027
Coursework: Machine Learning, Online Learning & Optimization, Deep Learning

The University of Texas at Austin — B.S. Computer Science & Mathematics May 2025 | GPA: 3.65
Awards: QuestBridge Scholar, HSF Scholar, University Honors
Activities: Directed Reading Program, Texas Blockchain, CS Internal Transfer Program
Coursework: Principles of ML, Distributed Computing, Concurrency (Honors), Geometric Foundations of Data Science, Quantum Info Sci, Real Analysis I–II, Stochastic Processes I (Grad), Predictive Analytics, Game Theory

Experience

Avathon | Data Scientist Oct 2025 – Present

- Built AI-oriented microservices and agent workflows for internal infrastructure, focusing on reliability and extensibility.
- Built an AI agent fluent in the knowledge graph's query language (IQL) with tools for schema discovery, graph traversal, and embeddings-backed search; required deep research into the CKG codebase to surface undocumented capabilities and validate embedding pipelines.
- Worked full-stack on a transport management system, including resource-selection config and operational logic.
- Pitched and helped adopt DBOS for workflow durability and orchestration across services.
- Integrated Langfuse for AI observability; continuing work on AI infrastructure, frontend architecture, and DevOps.

IBM | Data Science Intern May – Aug 2023

- Built and deployed a Flask POC integrating Watson Assistant with Db2; helped close a client deal.
- Created a 1.5k-LOC Dash app backed by watsonx, Discovery, and a custom eval API for LLM comparisons.
- Owned backend API integration for a React + Flask onboarding MVP on IBM Cloud and OpenShift.

UT Austin | Undergraduate Researcher, Earthquake Modeling Jan – May 2024

- Re-engineered a legacy Mathematica SDOF solver into a multithreaded Python pipeline.
- Processed 29k+ NGA-West2 records; ran 64-vCPU GCP sweeps, cutting runtime from ≈3 yrs to ≈2 wks.
- Produced interactive Plotly dashboards and a seminar deck for faculty presentation.

Selected Projects

Crabs AI Assistant | *OpenClaw, LightRAG, CouchDB* 2026

- Self-hosted AI assistant running on a Hetzner VPS with CouchDB syncing an Obsidian vault across devices.
- Two-tiered memory system: short-term conversational recall via OpenClaw and long-term knowledge graph traversal via LightRAG; zero-guessing policy forces the agent to always search actual notes instead of relying on pre-trained weights.

Adaptive DQN Planner | *RL, DQN, CEM, Python* 2025

- Built an RL agent for autonomous driving in a roundabout simulator, pairing DQN with a CEM optimizer and k-NN error tracker to adapt uncertainty estimates during planning.
- Achieved 0% collisions and +17% safety margin over deterministic baselines across 20 evaluation runs.

CUDA K-Means / K-Means++ | *CUDA, C++* 2025

- Implemented 4 GPU-accelerated K-Means variants including shared memory and K-Means++ initialization; benchmarked on dual Quadro RTX 6000s, achieving 14x speedup on 1M×32D datasets.

GPU Fine-Grained Sync for EM | *CUDA, Distributed Systems* 2025

- Recreated and extended Wang et al. (ASPLOS '19) fine-grained GPU synchronization using a client-server design with shared-memory buffers; benchmarked on counters, hash tables, and EM for GMMs.

Fantasy Draft RL Agent | *Python, PyTorch, Gymnasium* 2024

- Built a custom Gymnasium environment modeling 12-team snake drafts with 300+ players, action masking, and roster-aware observations; trained Maskable-PPO across multiple NFL seasons, outperforming heuristic baselines by 26–42%.

Black-Scholes Derivation Paper | *Probability, Measure Theory, Stochastic Calculus* 2024

- Wrote a self-contained derivation of Black-Scholes for European call options from first principles, covering measure theory, martingales, Brownian motion, and connecting the PDE framework to practical asset pricing.

Skills

Languages: Python, SQL, C/C++, CUDA C, TypeScript, Bash | *Familiar:* Java, Go, Rust, R
AI & Agents: LLM APIs (OpenAI, Anthropic, Google), agent orchestration & tooling, prompt engineering, RAG pipelines, LiteLLM, ChromaDB, LightRAG, Langfuse, model routing & compatibility, streaming/thinking pipelines
ML & Data: PyTorch, scikit-learn, pandas, NumPy, SciPy, Gymnasium, Stable-Baselines3, Matplotlib/Plotly
Web & APIs: Flask, FastAPI, Django, React, Next.js, REST APIs, Dash, Tailwind CSS, SSE/WebSockets
Infra & Tools: AWS (EC2/S3), GCP, Docker, Git, Kubernetes, OpenMP/MPI, GPU profiling (nvprof/nsight)
Spoken: English (native), Spanish (native)