DANIEL MORA

New York, NY | Authorized to work in the US
damorasoler@gmail.com | Personal Website | GitHub | LinkedIn

SUMMARY

Computer Science and AI graduate who builds and ships agent systems on Google Cloud, self-supervised models, GPU kernels and interpretability tools. Has worked in quantitative research, applied AI engineering, full-stack products and client projects. Reports results as they are, including the ones that fail.

EDUCATION

IE University **Madrid, Spain**
Bachelor in Computer Science and Artificial Intelligence Graduated: July 2026

- Thesis: *Can We Detect Shared Features Between Language Models and the Brain?*, supervised by Dr. Luciano Dyballa and written up as a paper.

EXPERIENCE

WhiteHole | *Hedge fund* **San Sebastián, Spain**
Quantitative Research Intern (AI Engineering Focus) Jun 2025 – Nov 2025

- Built a knowledge graph of company behavior with local LLMs. The team set the objective and data sources, and I designed the ETL pipeline, schema and extraction.
- Built an MCP server that connected the graph to an LLM so a non-technical team could query it in plain language. **The team adopted it** for its own work.
- Engineered NLP signals for market-volatility research, testing them in tree, linear and ridge models with Granger-causality analysis and regime checks.
- Built macroeconomic forecasting models in Python and presented the results to the finance team.

MAI-Book | *E-commerce start-up* **Remote**
Front-end Developer Intern Jul 2023 – Oct 2023

- **Cut client churn by about 20%** by analyzing user-testing data from 50+ UX surveys, talking directly with clients and refining the sales funnel.
- Built user-facing interfaces for the online library, simplified the checkout flow and ran A/B tests to guide product and conversion decisions.

PROJECTS

TC_ACC: agentic documentary studio | *Python, TypeScript, LangGraph, Pydantic, Remotion, ffmpeg* 2026

- Built a multi-agent pipeline that turns one click into a **finished 10-minute documentary with no human steps**: researched, written, narrated, illustrated, edited and rendered, for under $2 a run.
- Runs 28 typed stages with validated handoffs and 2 review gates that stop bad output. The test suite runs with every provider key removed, so a passing run checks the code itself.
- A reference run produced 128 shots from 760 reviewed assets at 1080p, and the episodes have drawn **400+ watch hours** on YouTube.

JEPA-Embedding Transformer (ArgusJEPA + CassaFormer) | *PyTorch, Triton, PostgreSQL* 2026

- Trained self-supervised JEPA encoders (up to 300M parameters) that learn by predicting an EMA teacher's latent of the full view from a half-masked one, then a separate forecaster for next-day moves of 500 US stocks; released the code and weights.
- Wrote Triton kernels A/B-tested on L4 and A100: a fused MoE expert MLP with GPU-side routing and deterministic gradients (**up to 2.4× full-model speedup** vs torch.compile) and a strict-FP32 matmul more accurate than cuBLAS. I deleted the kernels that lost.
- Built a point-in-time, survivorship-free data platform in PostgreSQL: daily prices, SEC EDGAR filings visible only after acceptance, FRED data by vintage and options data.

Squawk: private chat for coding agents | *Python, MCP, Starlette, FastAPI, GCP, Terraform* 2026

- Built a chat where Claude Code, Codex and other MCP agents on one project coordinate through DMs, channels, file claims and urgent interrupts, with a web app for accounts, projects and an admin dashboard.
- One pasted line lets an agent set up the project and lead it. In a live test, a Claude lead brought in Codex and a second Claude agent and split real work between them.
- Runs on Cloud Run with Firestore, Pub/Sub, Cloud Functions and a Dialogflow CX and ADK support assistant. Stores tokens only as hashes, offers end-to-end encryption and keeps dev and prod apart in Terraform.

Northwind Resolve: AI complaint handling for a retail bank | *Python, ADK, Dialogflow CX, GCP* 2026

- Built and deployed a multi-agent complaint system on Google Cloud. An ADK coordinator hands each case to intake, policy and case-status agents, and a Dialogflow CX webhook routes customer chats to it.
- Designed a startup check that runs everything on Vertex AI when the API is enabled, falls back to OpenAI, or runs **fully local on open-source models** with Ollama and SQLite, using the same agent code.
- Enforced Reg E and CFPB deadlines, logged every human override of AI triage, and streamed case events through Pub/Sub into a BigQuery audit trail.
- Secured it with Identity Platform roles, per-user data isolation and PII redaction, and defined all of it in Terraform with keyless deploys from GitHub Actions.

Brain–LLM sparse feature correspondence (BA thesis) | *PyTorch, SAEs, fMRI, GCP* 2026

- Designed a cross-description test of whether LLMs and the brain rely on the same SAE features: Narratives fMRI (59 participants, 200 parcels) and six SAE systems across Gemma and Llama models (1B–12B).
- Built the pipeline end to end: Whisper alignment, fMRIPrep preprocessing, causal feature extraction and ridge encoding models.
- When a stricter test removed the early effect, **measured the test's own power and reported a bound** instead of a claim; written up as a paper, with code and calibration released.

OLMo-3 sparse autoencoder release | *PyTorch, Hugging Face, SAELens, GCP* 2026

- Released seven SAELens-compatible TopK SAEs for OLMo-3-7B-Think (**100+ downloads**), streaming about 20M token activations per layer to disk on a single GPU.

Gemma 3 triplet extractor (SAE steering vs LoRA) | *PyTorch, PEFT, Hugging Face* 2025

- Compared SAE feature steering with LoRA fine-tuning for triplet extraction feeding GraphRAG, and **LoRA won clearly**. Released the model with a 4-bit GPTQ export (**300+ downloads**).

LearnMate: AI study planner (team of 7, solo rebuild) | *Next.js, TypeScript, FastAPI, Ollama* 2025–2026

- Built the full stack: a Next.js and TypeScript front end with auth, dashboards and study sessions on a FastAPI, SQLAlchemy and Alembic backend with hashed passwords.
- Rebuilt planning as a bounded orchestrator over five specialist agents on a local model, with a validated JSON action protocol, retries and keyword retrieval from uploaded material.

Autopost: automated content pipeline | *Python, GPT, ElevenLabs, Cloud Run* 2024–2025

- Built a pipeline that rewrote posts with GPT, voiced them with ElevenLabs, assembled video and posted to Instagram automatically on Cloud Run; ran it in production, then shut it down when the views didn't justify more time.

Cloud banking platform (team of 6) | *Vue, Flask, PostgreSQL, Azure* 2024

- One of two full-stack developers, doing the front-end work and some backend, on a test-driven banking app deployed to Azure with Bicep, GitHub Actions CI/CD and managed secrets.

Client websites | *React, Next.js, Tailwind, Vercel* 2025

- Took three client sites from scope to launch: a travel agency, the Reusa brand site and a medical practice.

Alien Invasion (team game) | *C++, SFML, CMake* 2024

- Built a C++ game with a team, with four enemy types, difficulty scaling, a persistent leaderboard and unit tests.

LEADERSHIP

- **Founder and President**, IE Dominican Republic Club: built IE's largest student-led diversity event (500+).
- **President**, IE Quant Finance Club.
- **Business Development Officer**, IE Investment Club: secured sponsorships with financial-sector firms.

SKILLS

- **Languages:** Python, TypeScript, JavaScript, C++, SQL. Bilingual in Spanish and English; French (B1)
- **ML:** PyTorch, Triton, Hugging Face, PEFT/LoRA, JEPA, MoE, sparse autoencoders, SAELens, vLLM
- **Agents:** Google ADK, Dialogflow CX, LangGraph, Pydantic, MCP, RAG, GraphRAG, OpenAI, Anthropic, Gemini
- **Google Cloud:** Cloud Run, Cloud Functions, Pub/Sub, Firestore, BigQuery, Cloud Storage, Vertex AI
- **Infrastructure:** Terraform, GitHub Actions, Workload Identity Federation, Docker, Azure, PostgreSQL, MongoDB, Neo4j, Redis
- **Web:** React, Next.js, Vue, FastAPI, Flask, Starlette, Tailwind, three.js