Boris Dev

AI Engineer

San Francisco • boris.dev@gmail.com • github • linkedin

Knowledge graphs • AI quality evaluation • Eliciting nuanced ground-truth from domain experts

Stack

- **AI / LLM:** DSPy, LangGraph, LangSmith, MCP, Pydantic, Jinja, Neo4j (GraphRAG), Azure Search (BM25)
- **ML / Data Science:** PyTorch, scikit-learn, Pandas, NumPy, Jupyter, SageMaker, AWS GroundTruth, Databricks / PySpark
- **Backend:** FastAPI, Flask, Django, SQLAlchemy, Postgres, Mongo, Docker, Temporal, Kafka, HTMX
- **Ops / Cloud:** AWS, Azure, OpenTelemetry, Jenkins, Splunk

Education

PhD in Quantitative Human Geography at SDSU and UCSB, 2015. Data science for location referenced social science problems. Dissertation: New Metrics for Assessing Inequality using Geographic Data

Experience

Sindri, Oct 2025 - Feb 2026, Consultant

Sindri is an early-stage startup applying AI to document management for large energy-industry construction projects.

Built the team's first AI evaluation framework, replacing engineer-driven manual QC/QA with automated checks and unblocking high-stakes customer demos by lifting AI email quality.

- Designed an SME-authored YAML expectations DSL (pre-run scenarios + post-run predicates) so domain experts — not just engineers — could specify what "correct" looks like for a Temporal workflow run
- Built a Temporal-aware test harness that snapshots post-run database side effects and activity outputs, then evaluates each expectation — became the team's foundational CI/CD for iterating on Temporal modules
- Built an LLM-as-judge pipeline that scores candidate prompts against synthetic test batches and emits a structured fault taxonomy (top faults, rationale, proposed prompt edits) to drive iteration

Nobsmed, 2024 - current, Founder

Nobsmed connects ChatGPT and Claude to clinical-trial findings that fit a user's specific situation — an auditable evidence layer addressing the "evidence-to-person fit" problem (e.g., a statin trial that excluded pregnant women being mis-applied to someone trying to conceive).

- Modeled a PICO-style ontology in Pydantic (`ParticipantGroup`, `StudyArm`, `OutcomeVariable`, with cross-reference integrity validators — defined once at paper level, referenced by id) and built it as a Neo4j knowledge graph queried with Cypher
- Exposed the graph as an MCP server (tools: `ask`, `decompose`, `resolve`, `evidence`, `filter_by_pertinence`, `concept_hierarchy`, `similar_concepts`) so agents compose multi-step graph queries — ontology-grounded GraphRAG, not vector-only retrieval
- Live demos (clickable): web UI answering *"OnabotulinumtoxinA vs sacral neuromodulation for urgency incontinence"*, and a public ChatGPT custom GPT (Clinical Trial Results) answering *"Show RCTs of non-metformin interventions for prediabetes"*
- Built an LLM extraction pipeline (Databricks / PySpark) over the PMC author-accepted-manuscript corpus that extracts structured findings per study arm (intervention, comparator, outcome, effect size, vs-baseline); ~250 papers ingested into the production graph to date
- Built an eval harness with subdomain competency-question YAMLs (gold questions across 11 clinical subdomains — prolapse, prediabetes, anxiety, infant sleep, etc.) plus per-paper extraction-error annotations; open-sourcing the IR + harness in progress

Smaller consulting gigs

- EcoR1, 2025 - LLM extraction of earning call calendar events
- Intuitive Systems, 2023, LLM extraction of AMD products from vendor receipts. LangSmith for evaluation.

AI Engineer consultant at Wolf Games, 2023-2024

Wolf Games is a murder mystery gaming company piloted by the producers of Law & Order.

- Fixed story generation to be consistent by building a DAG-based story composition engine that dynamically chained LLM prompts to maintain narrative coherence across overlapping multi-step workflows to ensure consistency in plot and in character MMOs (Means, Motive, Opportunity). Read Google AI showcase here

AI Engineer consultant at SimpleLegal, 2022-2023

SimpleLegal is a legal billing analytics company.

- Identified a poorly specified rubric as the root cause of low model quality on a stuck feature
- Designed a collaborative process for paralegals and lawyers to debate edge cases, build consensus, and elicit the nuanced expertise needed to refactor the rubric
- Built a quality-control annotation pipeline around the new rubric –> massive increase in training example quality and the launch of the previously stuck feature
- Deployed a PyTorch Small Language Model on SageMaker and the ML client into the Flask product app

Lead Analytic Endpoint Engineer at Sight Machine, 2018-2021

Sight Machine is a manufacturing analytics company.

- Built the backend engineering on biggest public facing analytic feature
- Implemented a pre-demo protocol between product and engineering –> less panic before each sales demo
- Coordinated QA process with sales and engineering –> better prioritization/triage
- Built company's first distributed tracing –> simpler firefighting for mid-level developers
- Containerized frontend build –> standardized team's setup & scaled testing to cloud

Lead Data Engineer at HiQ Labs, 2015-2018

HiQ Labs was a people analytics company.

- Taught data scientists how to refactor their pipeline code into microservices
- Refactored scraping system –> Established pipeline reliability
- Refactored data pipeline from a data science monolith to a micro-service paradigm –> Established release reliability
- Migrated the data science team from Mongo to PySpark/Databricks –> increased productivity on new product R&D

Developer at Urban Mapping, 2011-2013

Urban Mapping provided geospatial analytics to Tableau.

- Built developer tooling
- Built first performance regression gate –> Reduced failed releases/customer complaints
- Built first observability –> increased coding issues prioritization with new system performance metrics

Impactful projects

- Reduced Tableau customer complaints by building end-to-end regression tests for the top 100 geospatial queries, the company's first observability system, and CI/CD pre-commit performance gates
- Migrated a data science ETL monolith to microservices, reducing firefighting
- Revived a stuck AI feature by shifting the team's focus from training data quantity to quality
- Built a gaming company's first murder mystery story generator by chaining prompts to force consistency (post).

Papers & code	Non-tech fun
LLM-based taxonomy (topic modeling): bertopic-easy	Climbed Cotopaxi (21,000 ft)
Language AI Evaluation 101: Know your user	Bodyboarded Mexpipe
Langchain PR: Causal Program-aided Language (CPAL) — see Harrison Chase's tweet	Taught with students in Medellín, Colombia to make ClusterPy (open-source geo clustering library)
Work papers	Taught kids snowboarding as an instructor
Academic papers	Counseled severely emotionally disturbed children