Boris Dev — AI Evaluation Engineer

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

Stack

- **Data Science:** Pandas, Sklearn, Numpy, Plotly, PySpark, DataBricks, Jupyter, SageMaker, PyTorch, GroundTruth, SQLAlchemy
- **Web BE**: FastAPI/Flask/Django, docker-compose, kubectl, Postgres, Mongo, Prometheus, Grafana, Splunk, ELK, OTEL
- **Cloud**: Azure, AWS
- **Pipelines:** Jenkins, Kafka
- **LLM tooling**: Instructor, Claude code, Goose, Cursor, Copilot

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, 2025 - current, Consultant

Built Temporal AI agent evaluation CLI framework

- **Temporal AI workflow evaluation steps:**
  1. Author an expectations yaml containing post-run predicates and pre-run scenarios
  2. Run temporal workflow and collect snapshot (post-run db side-effects and activity outputs)
  3. Test each expectation and report failures
- **LLM-as-judge and prompt fine-tuning steps:**
  1. Patch the AI system with a candidate prompt
  2. Build a batch of input test examples synthetically
  3. Using a Jinja prompt template, instruct the LLM to identify major and minor faults per test example
  4. Instruct LLM to summarize the aggregation of faults per prompt: score, score rationale, top faults, propose prompt changes.
  5. iterate
- **Design doc for SME-authored executable expectations** [in progress]

Nobsmed, 2024 - current, Founder

Built LLM based website, https://nobsmed.com/, for users to search for biohacking experiences found in clinical studies and Reddit comments.

- ETL: Language AI parsing of 100,000 studies and 1,000,000 Reddit comments.
- Topic modeling: Extended Bertopic using LLM for naming and classification of outliers, bertopic-easy.
- Customer discovery and user interviews.
- Experimented with fine-tuning opensource embeddings model and ML classification.
- Prompt engineering using InstructorAI.
- Embedding DBs for search (Azure Search, Weaviate, OpenSearch).

Smaller consulting gigs

- The Program Labs, 2025 - Building LLM AI automation tooling and a Google Ads Experimentation platform for a non-profit accelerator that helps startups in disadvantaged communities.
- EcoR1, 2025 - Built LLM AI scraper to extract earnings calls
- Wolf Games, 2023 - Built new graph prompting for story generation
- Intuitive Systems, 2023, NED (Name Entity Disambiguation) of vendor receipts fed into AMD's sales analytic pipeline. Used LangSmith for evaluation.

AI Engineer consultant at SimpleLegal, 2022-2023

SimpleLegal is a legal billing analytics company.

- Training examples curation: two SME lawyers and five human annotators
- Deployed PyTorch SLM on SageMaker and its client in to the product.

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
- Coordinated QA process with sales and engineering
- Built company's first distributed tracing
- Built the containerized frontend engineering development environment

Lead Data Engineer at HiQ Labs, 2015-2018

HiQ Labs was a people analytics company.

- Built the scraping system and observability (Splunk)
- Led the migration from a monolith to a micro-service paradigm
- Migrated the data science team to DataBricks PySpark and microservices

Developer at Urban Mapping, 2011-2013

Urban Mapping provided geospatial analytics to Tableau.

- Formulated first map rendering and query performance metrics integrated into our CI/CD pipeline

Impactful projects

- Revived a stuck AI feature by shifting the team's focus from training data quantity to quality
- Reduced Tableau customer complaints by building a new observability system and CI/CD pre-commit metrics
- Reduced the data science team's firefighting by building a microservice architecture
- Built a gaming company's first murder mystery story generator by chaining to prompts to force consistency (post).

Papers and code

- LLM based taxonomy (topic modeling): bertopic-easy.
- Language AI Evaluation 101: Know your user
- Langchain PR: Causal Program-aided Language (CPAL). See Harrison Chase's Tweet.
- Work papers
- Academic papers

Non-tech fun points

- Climbed Cotopaxi (21,000 ft).
- Bodyboarded Mexpipe.
- Taught with students in Medellín, Columbia to make an open-source geo clustering library (ClusterPy)
- Taught kids snowboarding, as an instructor
- Managed service workers, as a restaurant assistant manager
- Counseled severely emotionally disturbed children