

# Boris Dev — AI Engineer

San Francisco • boris.dev@gmail.com • [github](#) • [linkedin](#)

## Interests

0-1 building RAGs • subjective AI quality evaluation • innovating human processes • LLM concept extraction • building cognitive empathy flywheels

## Stack

- **Data Science/LLM/ML:** Dspy, Pandas, Sklearn, Numpy, Plotly, PySpark, DataBricks, Jupyter, PyTorch, GroundTruth, SQLAlchemy, geo-spatial data and social science data, SageMaker, GroundTruth
- **BE Engineering:** Azure Search, FastAPI/Flask/Django, docker, Postgres, Mongo, Open Telemetry, Azure, AWS, Jenkins, Kafka, 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, 2025 - current, Consultant

Built initial evaluation framework to score the quality of an AI agent's supplier-error remediation. This saved developer time from manually validating side-effects and email quality after each prompt or agentic code change.

- **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 <https://nobsmed.com/> for users to search clinical studies and reddit posts for treatment and health insights.

- LLM extraction of clinical study treatment findings and reddit personal health experiences
- created [bertopic-easy](#) for clustering/topic modeling

### 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.

- Redesigned rubric and quality control pipeline → massive increase in training example quality resulting in launching a previously stuck feature
- Deployed 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
- Demo protocol → less panic before each sales demo
- Coordinated QA process with sales and engineering → better prioritization/triage
- Built company's first distributed tracing → simpler fire-fighting 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.

- 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 DataBricks → increased productivity on new product R&D

## **Developer at Urban Mapping, 2011-2013**

Urban Mapping provided geospatial analytics to Tableau.

- SRE: Built first performance regression gate → Reduced failed releases/customer complaints
- SRE: Built first observability → increased coding issues prioritization with new system performance metrics

## **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