

Nick Lemoff

San Francisco Bay Area / Remote | nlemoff@gmail.com | github.com/nlemoff | linkedin.com/in/nicklemoff | nlemoff.com

Founding engineer focused on applied AI research and full-stack product engineering. I turn papers, open-source systems, and fast prototypes into production agents, retrieval systems, eval harnesses, and user-facing software.

EXPERIENCE

FullProof — Founding Engineer

San Francisco, CA / Remote | Mar 2025 - Present

- Architected Evie, FullProof's real-time deposition agent, from a demo spread across disconnected services into a unified GCP-backed production system for live transcript processing, document-grounded reasoning, attorney-facing suggestions, and deposition-prep workflows.
- Owned FullProof's applied AI research loop: read papers, studied open-source repos and technical blogs, tracked model releases, and built hundreds of prototypes across agent memory, real-time agents, hybrid RAG, retrieval/reranking, tool use, long-context workflows, and eval methodology.
- Productionized the strongest research ideas into a Python/FastAPI agent service with streaming endpoints for live transcript lines, chat/RAG, document extraction, strategy generation, outline generation, tracker summaries, persistent case memory, and final-round deposition workflows.
- Designed the real-time agent orchestration layer that records transcript state, routes each line through research / quick-check / suppression paths, and returns citation-backed tactical suggestions with timing, route, usage, and tool-call traces.
- Built document-grounded retrieval over case files with Vertex AI Search / Discovery Engine, case-scoped filtering, dynamic exhibit mapping, caching, concurrency controls, source-identity observability, and fail-closed behavior for unsupported or cross-case evidence.
- Created replay/evaluation infrastructure for AI reliability: transcript replays, browser-projected agent output, service health snapshots, event logs, line-level lifecycle traces, scoring, and regression coverage.
- Hardened the agent's evidence chain by synchronizing generated text, grounded documents, citation bundles, UI citations, deduping/threading behavior, and unsupported-output suppression so user-facing suggestions stayed tied to source material.
- Shipped across React/Vite, TypeScript, Node/Express, Python, FastAPI, PostgreSQL, Prisma, Redis, Socket.IO, BullMQ, Vertex AI Search, and Google Cloud.

Formula SAE Electric at Berkeley — Simulations Team Lead

Berkeley, CA | May 2024 - Jan 2025

- Led a 25+ person simulations team across 7 vehicle-modeling projects for UC Berkeley's Formula SAE electric racecar team.
- Developed a lap-time simulator using a two-track bicycle model, introduced Git/GitHub workflows, built CAN-to-simulation analysis pipelines, and containerized simulation environments with Docker.

National Institute of Standards and Technology — Research Intern

Remote | Jan 2024 - Aug 2024

- Implemented optimized graph edit distance code in Python and C++ for inorganic materials datasets after reviewing core algorithmic approaches.
- Contributed to a long-running materials-science project embedding 167,500 ICSD entries into a structural-similarity space to study experimental discovery, computed materials proposals, and synthesizability priors.
- Reviewed and merged multiple PRs to make the latest research code reproducible and runnable on other machines ahead of the group's Nature submission.

Lawrence Berkeley National Laboratory — Machine Learning Student Assistant

Berkeley, CA | Sep 2023 - Aug 2024

- Developed Faster R-CNN and YOLOv8 models to detect oil and gas wells in historical maps, reaching 0.98 mAP in under 5 minutes of training.
- Reviewed 20+ CNN architectures and co-authored an abstract submitted to AGU 2024 on the object-detection approach.

Mojo Vision — Data Science and Data Management Intern

Saratoga, CA | Jul 2023 - Jun 2024

- Designed Python/MySQL data pipelines for quantum dot life tests that became the company's standard processing workflow.
- Built Tableau dashboards and Python/SQLAlchemy monitoring code for validation data and internal database health.

Formula SAE Electric at Berkeley — Simulations Engineer / Data Scientist

Berkeley, CA | Feb 2024 - May 2024

- Developed a tire model using Tire Testing Consortium data, pandas, and scikit-learn; created a battery-pack sizing process balancing sizing constraints, weight, and lap-time impact.

SELECTED PROJECTS

Evie — FullProof's real-time deposition agent. GCP-backed real-time agent with FastAPI streaming endpoints, hybrid RAG over case documents, agent memory, live transcript orchestration, citation-grounded outputs, and replay/eval infrastructure.

SfRB — Straightforward Resume Builder. Local-first TypeScript resume design tool with installable Node CLI, Vite browser editor, canonical JSON model, Zod validation, templates, shared print/PDF export, overflow checks, and optional AI layout repair proposals behind explicit acceptance. github.com/nlemoff/SfRB

FEBSim — Formula Electric Berkeley lap simulation. Racecar performance simulation project. Led work around lap-time modeling, vehicle dynamics, CAN/simulation analysis, Dockerized contributor workflows, and team engineering practices. github.com/Formula-Electric-Berkeley/FEBSim

SKILLS

Languages: Python, TypeScript, JavaScript, SQL, C/C++, Java, Go, R

AI / ML / research: LLM agents, real-time agents, agent memory, hybrid RAG, retrieval/reranking, Vertex AI Search, eval harnesses, replay testing, prompt/system design, PyTorch, TensorFlow, scikit-learn, computer vision, NLP, literature review, rapid prototyping

Product engineering: React, Vite, Next.js, Node.js, Express, FastAPI, PostgreSQL, Redis, Socket.IO, BullMQ, Prisma, Tailwind CSS

Infrastructure / workflow: Google Cloud, Docker, Git/GitHub, CI/CD, Playwright, Jupyter, Tableau, observability

EDUCATION

University of California, Berkeley — B.A. Data Science | Aug 2021 - Dec 2024