

Rohan Patil

Machine Learning Engineer | Production ML | Applied AI | Cloud Systems

hi@rohanpatil.ai | rohanpatil.ai | linkedin.com/in/knowrohan1 | github.com/knowrohan | Pune, MH

SUMMARY

ML engineer building production AI systems where model outputs are constrained by a deterministic core - schema- and integrity-validated before touching state. Currently architecting a multi-tenant AI publishing platform serving **8** publishers; previously built ML models and **500GB+** pipelines for nuclear-fusion research. MSc Data Science & Artificial Intelligence; BS Applied Physics.

TECHNICAL SKILLS

Production ML: Python, PyTorch, scikit-learn, Hugging Face, model evaluation

Applied AI & Agents: LangGraph, LangChain, LLM systems, tool-calling schemas, RAG, MCP, vector search

Data & Backend: FastAPI, Pydantic, PostgreSQL, SQL, async job runtimes, ETL pipelines

Cloud & Delivery: Azure, AWS, Cloudflare Workers, Docker, GitHub Actions/OIDC, observability

Analytical: Regression, hypothesis testing, time series, experimental design, Monte Carlo, R

EXPERIENCE

Machine Learning Engineer - PANTA OS

Jan 2026 - Present

ILD - Multi-tenant AI editorial layout platform (copy + images to print-ready magazine spreads)

- Architected a multi-tenant AI publishing platform (React 18/TypeScript SPA; FastAPI/PostgreSQL on Azure) that composes print-ready magazine spreads from unstructured copy and image sets, with production PDF and InDesign IDML export across **40+** services; **8** publisher tenants live.
- Built the AI layer around a **deterministic core**: vision-based image tagging, ranking and assignment with per-brand scoring; natural-language brand guidelines compiled into validated JSON patches; and embedding-based semantic media search. Every model output clears schema and integrity validation before mutating state; geometry stays math.
- Shipped **ILD Lotse**, a conversational page-editing assistant: Claude over SSE with a **20-tool** server-executed schema for typography, image swaps, object moves and body-chain resizing; stylesheet, object-ID and chain-continuity validation before apply; confirmation gating for persistent operations; full undo.
- Built the deterministic layout-selection and copyfit engine: capacity-targeted template search under spread-parity, page-side, reuse and fixed-page-count constraints with asymmetric overflow penalties. Ran the frontend's own text-measurement code server-side in headless Chromium so preview and print agree to the character.
- Implemented **VJSON**, the canonical layout representation, and the bidirectional VJSON ↔ Viva XML conversion layer. One format drives preview, PDF/JPG export, IDML handoff, designer round-trip, canvas editing and the AI assistant, so new formats ship without renderer changes.
- Built the async composition runtime: job state machine with concurrency limiting; original-input snapshotting so retries replay the user's request rather than mutated state; scheduled stuck-job recovery; and GitHub Actions/OIDC image promotion (dev-tested artifact to production), eliminating deployment drift.

Computational Analyst - SapientAI, Austin, TX

Nov 2022 - Dec 2024

- Developed and deployed Python ML models improving the fidelity of nuclear-fusion plasma simulations - a **15%** gain in prediction accuracy for downstream research analysis and client reporting.
- Engineered data pre-processing pipelines across diverse **500GB+** datasets, cutting computational runtime **40%**.
- Maintained **3** production research applications and presented findings to cross-functional stakeholders to guide technical decisions.

Freelance Full-Stack Developer - Self-Employed

2024 - Present

- Architect high-performance, edge-first web applications on the Cloudflare stack - Workers, Pages, D1 and R2 - delivering Next.js and Astro builds for agency, hospitality and wellness clients.

Teaching Assistant - University of Massachusetts Amherst

2019 - 2022

- Led Physics & Astronomy labs and discussion sections for **120+** undergraduates; graded and gave feedback on **1,200+** assignments.

SELECTED PROJECTS

- 2048 Master - AlphaZero RL** | Python, PyTorch, MCTS, ResNet, bitboards. Self-play agent with **zero** human gameplay heuristics: 1D bitboard engine at approximately **1,000 simulations/sec**, MCTS over a ResNet policy/value network, **8x** board-symmetry augmentation and a reproducible train/evaluation pipeline.
- Autonomous Agents - LangGraph Due Diligence** | LangGraph, LangChain, RAG, vector search. Mandate planner to specialist agents (filings, news and market data), retrieval, cross-verification and sourced synthesis; compresses a days-long research cycle into minutes, with every claim carrying primary/secondary sources and a confidence score.
- Product Ingredient Analyzer** | FastAPI, Pydantic, PostgreSQL, Cloudflare Workers, Gemini API. Turns one photo of an ingredient label into a structured safety brief, executed at the edge with drag-and-drop upload, image processing and graceful error handling.
- More:** FinBERT SEC sentiment, YouTube sentiment, clean-news aggregator and Monte Carlo muon mass - rohanpatil.ai/projects

EDUCATION

MSc Data Science & Artificial Intelligence - COEP Technological University, Pune

2025 - 2026

BS Applied Physics - University of Massachusetts Amherst

2018 - 2022

Minor: Resource Economics