

Divanshu Sharma

Applied AI / ML Systems Engineer · LLM evaluation, inference & reliability ·

Founding Engineer @ Uniiq · Mumbai, India

divyanshu74.80@gmail.com · div90.vercel.app · github.com/sdivyanshu90 ·
linkedin.com/in/divsha22

● Open to AI engineering & founding roles

SUMMARY

Founding engineer who ships reliable AI products – currently leading the technical turnaround of Uniiq's AWS-deployed AI advising platform: 40+ critical vulnerabilities resolved, LLM sanitization and recoverable AI failures, LCP cut from 3.3s to 0.7s. Before that: privacy-preserving deep learning research at Yale and three years of quantitative research at WorldQuant BRAIN.

EXPERIENCE

Founding Engineer – Uniiq May 2026 – present

- Led the technical turnaround of inherited legacy code, resolving 40+ critical vulnerabilities and adding prompt and response sanitization against injection and data-leak risks.
- Reworked student onboarding into a three-phase AI conversation with follow-up questions, persisted transcripts, structured validation, rate limits, and recoverable AI-failure handling.
- Implemented role-aware profile-completion enforcement and international phone capture across protected APIs, sign-up, setup, and profile flows.
- Expanded institution and opportunity administration with deep-partial updates, nested Mongo normalization, validation, and external-ID uniqueness safeguards; added 1,199 lines of automated tests across 12 modules.
- Raised the web performance score from 73 to 93, cut LCP from 3.3s to 0.7s, and reduced INP to 130ms after code-splitting 21 routes and optimizing delivery.

ML Researcher – Yale University Oct 2025 – Mar 2026

- 88.08% ChestMNIST test accuracy in 5 epochs at batch size 512 – trained under MPC.
- No external autograd: layers, backprop and optimizers derived by hand.
- CNN stack proposed upstream to 0xTCG/sequire across four pull requests.

Research Consultant – WorldQuant BRAIN Nov 2022 – Jul 2025

- Developed and backtested 15+ alpha signals; average Sharpe ratio 1.8.
- GOLD LEVEL (Top 1%) and Stage 1 Qualifier, Global Alphathon 2022.

SELECTED WORK

Uniiq – AI Student-Advising Platform Founding engineering · Full-stack AI product · 2026
40+ critical vulnerabilities – resolved in inherited legacy code, including prompt-injection and data-leak risks; 73 → 93 performance score – LCP cut from 3.3s to 0.7s; INP reduced to 130ms.

Privacy-Preserving Deep Learning via MPC ML research · Yale University · 2025–26
88.08% test accuracy – ChestMNIST, strided-convolution CNN, 5 epochs @ batch 512; 0 external autograd deps – backprop, conv & pooling layers derived and implemented by hand.

Indic ASR Optimization & Consensus Evaluation Speech · Evaluation research · 2025
>48% absolute WER reduction – FLEURS Hindi benchmark vs. baseline Whisper-small; 177,000 unique terms – normalized via Devanagari engine + reverse transliteration.

RenAIssance OCR – Historical Document Pipeline Computer vision · NLP · 2026
Word accuracy 97.96% → 98.17% – full corpus, after LLM post-processing; CER 3.44% → 0.00% – on the held-out 183-page validation set.

Paged KV-Cache Engine (vLLM-style)

LLM inference systems · 2026

Copy-on-write forking – ref-counted blocks shared across forked sequences; Radix-tree prefix cache – shared prompt prefixes served without recompute.

Graph Data Explorer AI

Applied LLM product · 2026

765 nodes / 877 edges – typed knowledge graph from 19 source tables; Two-stage LLM query engine – NL → traversal → NL, fully sandboxed.

Deep Learning for CMS Detector Physics

Scientific ML · ML4SCI / CERN CMS · 2024

80% AUC – electron-vs-photon classification, custom ResNet-15; -30% training time – GCN matching the GAT baseline for momentum regression.

DIV-1 – this portfolio, built as a production AI system

LLM systems · Evaluation · Full-stack · 2026

72/72 golden questions – routed correctly, plus 37/37 prompt-injection cases – enforced in CI on every push; \$0 for most answers – presets, honest absences and cached repeats need no model; a narrated answer costs about \$0.0002.

OPEN SOURCE

Mastra ([mastra-ai/mastra](#)) 11 merged PRs – 29 PRs and 18 bug reports to the TypeScript agent framework (Aug–Sep 2026). 11 merged directly – codemod and CLI fixes for its migration tooling – and all 18 bugs he found and patched are fixed upstream, one merged with him as co-author. #58 of 466 contributors by commits.

EleutherAI lm-evaluation-harness 4 merged PRs – 25 PRs to the standard LLM evaluation harness (Aug–Sep 2026): 4 merged – CLI argument parsing, request-cache robustness, and a restored benchmark task – and 21 in review.

Experiential ([experientiallabs](#)) 2 merged PRs – 20 PRs (Sep 2026): 2 merged model-provider fixes – paginated Anthropic model discovery and Gemini page-token handling – and 18 in review.

Sequire ([0xTCG](#)) proposal, in review – 4 PRs bringing the Yale work upstream – CNN layers and an MPC ChestMNIST training pipeline (#41 alone: +3.5k lines across 54 files). Not merged yet; #42 is open.

p5.js Web Editor ([Processing Foundation](#)) 4 merged PRs – 10 PRs (2023–24), 4 merged: a keyboard shortcut for adding files, the Korean translation of keyboard shortcuts, and two UI fixes.

OpenCode ([anomalyco/opencode](#)) 21 in review – 22 fix PRs to the open-source AI coding agent (210k★), Aug–Sep 2026 – 21 open, awaiting maintainer review.

Also merged: [vulnerablecode](#), [Music Blocks](#), [PyNN](#) 3 merged PRs – A license-URL fix to [aboutcode-org/vulnerablecode](#), a close-button fix to [Sugar Labs' Music Blocks](#), and a docs-layout fix to [NeuralEnsemble's PyNN](#).

SKILLS

LANGUAGES	Python (primary) · C++ · TypeScript / JavaScript · Rust · SQL · Bash · Java
MODELING & TRAINING	PyTorch · TensorFlow · Keras · scikit-learn · Hugging Face Transformers · PyTorch Geometric · CNNs / GNNs · Whisper · DeBERTa / RoBERTa
LLM SYSTEMS	Gemini · structured AI outputs · RAG · SFT · LangChain / LangGraph · Model Context Protocol · prompt engineering · Qdrant · Chroma · Ollama
SERVING & MLOPS	FastAPI · Docker · ONNX · GitHub Actions CI/CD · MLflow · Airflow · AWS (SageMaker, EC2) · GCP · Azure · Streamlit
DATA	NumPy · Pandas · Polars · CatBoost · PostgreSQL · MySQL · MongoDB · SQLite
WEB & PRODUCT	React · Next.js · Node.js / Express · Django / Flask · Tailwind CSS · server-sent events · system design

EDUCATION, CERTIFICATIONS & HONORS

B.E. Computer Science – University of Mumbai

CGPA 8/10 · 2021 – Jun 2025

Machine Learning Specialization (DeepLearning.AI, 2023) · Deep Learning Specialization (DeepLearning.AI, 2023) · TensorFlow Developer Certificate (DeepLearning.AI, 2024) · GANs Specialization (DeepLearning.AI, 2024) · Meta Front-End Developer (Coursera, 2023) · Postman API Fundamentals Student Expert (Postman, 2023)

- GOLD LEVEL (Top 1%) & Stage 1 Qualifier – WorldQuant BRAIN Global Alphathon 2022
- Kaggle 2× Expert
- LeetCode contest rating 1,821 (top 7.5%) · CodeChef 3★ (1662) · 4,400+ problems solved
- GitHub: Pull Shark ×3 · Starstruck ×2 · YOLO · Quickdraw
- 195 public repositories · 337 stars