

ANIRUDDH ATREY

LLM Engineer | Full-Stack GenAI | RAG, Agentic Systems & Production AI

atreyaniruddh@gmail.com | +91 93197 88556 | Noida, India | aniruddhatrey.com | linkedin.com/in/aniruddhatrey | github.com/AndrousStark | ORCID

EXPERIENCE

GenAI Full-Stack Engineer | [SaveLIFE Foundation](#)

New Delhi, IN | 10 Nov 2025 – Present

- Engineered a multi-stage RAG system (Haystack v2 + e5-base-v2 768-dim) over 3 MongoDB Atlas vector+text index pairs spanning corridor, district and EDAR collections, with entity-first query routing, Reciprocal Rank Fusion and citation grounding, cutting analyst lookup time ~70% across 100+ Zero-Fatality-Corridor reports per week.
- Productionised a bilingual (EN/HI) WhatsApp conversational AI on a 12-state FSM (TypeScript, Express 5, BullMQ 5.x, ioredis, lru-cache) with HMAC-SHA256 timing-safe verification, Redis SET-NX idempotency, dual LRU+Redis caches and a circuit-breaking Cloud API client behind a rate-limited outbound queue at 70 msg/s with exponential-backoff retry — serving 114 videos and 18 graded milestone exams across 99 learning journeys.
- Built a computer-vision telemetry pipeline (Python, FastAPI, OpenCV, EXIF, GeoPandas, Shapely, QGIS) over 500+ hours of dashcam footage at 95% precision, mapping violations across 200+ km of national highway and reducing manual audit effort by 85%.
- Deployed predictive ETL and ML dashboards on AWS (Lambda, S3, RDS, EventBridge, React, Recharts) identifying 15+ accident black-spots, at 99.9% service availability.

VAPT & Cybersecurity Engineer | INNEFU Labs (DRDO partner)

New Delhi, IN | 21 Mar 2025 – 8 Nov 2025

- Built the Phantom security framework (10+ VAPT tools) automating discovery, scanning, validation and remediation.
- Built agentic alert triage (Splunk + LangGraph + Claude 3.5 Sonnet) improving SLA to ~90% and analyst throughput by 3×.
- Implemented an evaluation framework with real-time metrics (F1, MTTR, false-positive monitoring), DPO-finetuned fallbacks, KV-cache optimisation and cost-aware routing, maintaining 99.95% reliability in a compliance-critical environment.
- Secured 50+ government web assets for the Ministry of Defence, India through structured reconnaissance, SQL-injection testing, subdomain enumeration and directory brute-forcing, reducing exposure to critical CVEs by 65%.

DevOps Engineer | Arlo Technologies

Carlsbad, CA, USA | 7 Jul 2024 – 7 Dec 2024

- Optimised AWS ETL pipelines (Lambda, SQS, DynamoDB, S3) reducing DynamoDB load by 30% and stabilising fleet-wide latency via predictive auto-scaling.
- Hardened CI/CD across 150+ deployments (GitHub Actions, Docker, CodeDeploy, Jenkins) lifting success rate from 95% to 99% with automated rollback and distributed tracing.
- Built multi-model LLM serving (GPT-4o, Claude Opus, Mistral) with quantisation, KV-cache optimisation and cost-routing, reducing inference costs by 35%.

AI Developer Team Lead | Microsoft-funded startup incubator

Gurugram, HR, IN | 31 May 2022 – 31 Aug 2022

- Engineered a two-stage semantic ranking system (BERT cross-encoder fine-tuned on university-fit labels + scikit-learn ensembling) for program matching across 5,000+ institutions, achieving 92% ranking precision and shipping 0→1 to 15,000+ users in 16 weeks with 3,200+ successful transactions and 92% user satisfaction.
- Led a cross-functional 5-person team through end-to-end ML delivery (data labeling → feature engineering → model training → A/B testing → canary deployment), driving 40% MoM growth and a 260% traffic lift.
- Uncovered market insight via geospatial analysis: 78% of conversions concentrated in tier-2/tier-3 Indian cities, enabling targeted expansion and geographic feature encoding.

FLAGSHIP PRODUCTS

CourtNetra + NyayaLens — Legal-AI SaaS over an 18.8M+ judgment Indian case-law corpus courtnetra.com | [Play Store](#)

- Built and operate a corpus of 18,863,754 Indian judgments (39,158 Supreme Court + 18,824,596 High Court across 25 High Courts, 1950–2026) on self-hosted PostgreSQL + pgvector, alongside 1,807 statute sections, 6,315 judge profiles and ~165,000 citation-network edges.
- Embedded 687,289 chunks with Qwen3-Embedding-4B at halfvec(2560) over HNSW; hybrid retrieval fuses dense pgvector search with tsvector BM25 through Reciprocal Rank Fusion, with citation grounding and a weekly automated hallucination-eval gate in CI.
- Shipped a 3-repository multi-tenant platform — Express.js + Prisma backend (268 endpoints), Next.js 14 web (41 pages) and an Expo/React Native app (37 screens, live on Google Play) — auto-fetching cases from 19,660 courts, with 8-language i18n and Razorpay billing.

OutBrew — Enterprise LLM outreach platform, 500+ teams; ~138K Python + ~96K TypeScript LOC

[case study](#)

- Implemented a 9-layer LLM extraction pipeline (sentence-transformers e5 + FAISS, Instructor-validated JSON extraction, Anthropic Batch API for 98% cost reduction, Claude 3.5 Sonnet + Mistral fallback) with PII redaction, content-safety gates and prompt versioning for A/B testing.

- Scaled to 50,000 emails/day at 95% deliverability and 3× industry reply rate, sustaining 99.9% uptime on Next.js 14, FastAPI, PostgreSQL, Redis and ClickHouse behind a multi-tenant dispatcher with per-domain throttling.
- Built the deliverability and reply-intelligence stack: domain warm-up, DKIM/SPF/DMARC alignment, multi-ESP rotation, bounce and complaint feedback loops, and an LLM reply classifier auto-routing qualified conversations to CRM with real-time conversion analytics.

VelaReach — Enterprise AI ad platform — 10 ad platforms, 11 attribution models

velareach.com

- Architected 9 Dockerized microservices with 10 platform adapters and 14 AI features across 122 API endpoints, covered by 807 tests — consolidating the roles of Keitaro, Dolphin, Triple Whale and Northbeam into one system.
- Implemented 11 attribution models including Markov chains, Shapley value, Bayesian inference and LSTM sequence modelling.

UNJYNX — AI productivity OS — Flutter + Next.js + Hono, omni-channel delivery

unjynx.me

- Built 3 repositories and 6 services: a Hono + Drizzle + PostgreSQL backend (152 endpoints, 21 modules, 1,102 tests), a Next.js 15 SSR client, Refine admin/dev portals and a Flutter monorepo of 24 packages with 61 mobile screens, watchOS and Wear OS targets.
- Delivers reminders across 8 notification channels (WhatsApp, Telegram, Instagram, SMS, Discord and more); shipped to Play Store and App Store via Codemagic with Shorebird OTA updates.

Badhiya AI — AI business OS for 60M+ Indian MSMEs — Hindi-first, ONDC-integrated

[case study](#)

- Delivered 3 repositories with 175 API endpoints, 62 mobile screens and 33 web pages on a 7-layer AI pipeline, with ONDC Beckn protocol integration and a Hindi-first voice interface.

SaveLIFE Super-App & Surakshak — Mobile super-app and bilingual WhatsApp e-learning bot for crash-investigation training

- Contributed to a ~10-service road-safety platform (Node/Express, Python Flask, MongoDB Atlas, PM2, nginx, GitHub Actions CD) including the SaveLIFE mobile super-app and the Surakshak WhatsApp e-learning bot for the Margdarshak crash-investigation programme.

Savidhi — Production engineering & security audit for a live temple-services platform (via MetaMinds)

savidhi.com

- Delivered a consolidated technical, financial and risk audit synthesising 51 underlying reports, every claim cited to file:line — surfacing 4 internet-exploitable critical vulnerabilities (forgeable identity headers trusted by 4 of 7 services, open admin self-registration, a seeded default credential, and unauthenticated SMS/OTP/push endpoints).
- Led the platform off GCP (GKE Autopilot + Cloud SQL + Memorystore) onto a single Hetzner instance, then took over ongoing production engineering across ~60 deployments.

OPEN-SOURCE & PUBLIC REPOSITORIES

- [FormWright](#) (Node 20, TypeScript, Playwright) — an agent that fills multi-page insurance forms it has never seen from one applicant JSON, with no hardcoded selectors: it resolves each control's accessible name the way a screen reader would, then maps it through a cost-ordered cascade (exact → concept/alias → fuzzy → embeddings → bounded LLM classifier). Filled 24/24, 31/32 and 28/29 fields across three forms with 0 fabricated values, every fill confirmed by DOM read-back — all three runs fully deterministic (0 LLM calls, ~6–10s, \$0). Abstains and flags rather than guessing when confidence is below threshold.
- [Medical Record Policy Evidence Extractor](#) (Python 3.11) — grounded clinical evidence extraction where hallucination is impossible by construction: the model returns an integer ID into a code-built candidate inventory, never a path or offset, and code re-slices every emitted snippet from the untouched record; a final assertion re-resolves the rendered path before output and treats a violation as a P0 fault.
- [StealthWeb Enumerator](#) — web-security reconnaissance tool for DNS and directory enumeration, virtual-host discovery and web-technology fingerprinting.
- [Network Intrusion Detection](#) (ML-based IDS) | [kitsune-py](#) | [Gator Taxi](#) (red-black tree + min-heap dispatch) | [port-scanner](#)
- [Savidhi](#) | [MetaMinds Store](#) | aniruddhatrey.com — Astro + React + Three.js portfolio scoring 100/100 Lighthouse SEO.

EDUCATION

M.S. Computer Science | [University of Florida](#)

Jan 2023 – Dec 2024

B.Tech Computer Science & Engineering | [Amity University](#)

Jul 2019 – Apr 2023

PUBLICATIONS

- Atrey, A. et al. Urban Digital Twins for Sustainability: Singapore & Indian Smart Cities. IEEE 2nd Global AI Summit, 2025.
- Atrey, A. et al. Real-Time Temperature-Based Food Recommendation Using AI. IEEE 14th ICCCNT, IIT Delhi, 2023.
- Atrey, A. et al. Cybersecurity & Ethical Hacking with Vulnerability Assessment Tools. Book chapter, Big Data Analytics Framework for Smart Grids, 2023.

SELECT CERTIFICATIONS

Stanford Machine Learning Specialization | Google Cybersecurity Professional | IBM Data Science | Cisco CCNA (Enterprise, Security & Automation) | EC-Council Ethical Hacking | ISO/IEC 27001

TECHNICAL SKILLS

LLM / AI / ML: Hugging Face Transformers, sentence-transformers (e5, BAAI/bge, Qwen3), PyTorch, TensorFlow, scikit-learn, XGBoost, spaCy, NLTK, OpenCV, Haystack v2, LangChain, LangGraph, LlamaIndex, Anthropic Claude SDK, OpenAI API, Ollama, vLLM, ONNX, MCP, RAG, CRAG, Reciprocal Rank Fusion, bge-reranker, Cohere Rerank, Pinecone, FAISS, ChromaDB, MongoDB Atlas Vector, prompt engineering, prompt versioning, safe fallbacks, structured prompting, JSON-mode tool calling, multi-LLM routing, cost-aware cascades, context management, autonomous agents, task decomposition, shared memory, tool-augmented reasoning, human-in-loop evaluation, hallucination and groundedness gating, content-safety guardrails, PII redaction, latency-and-cost optimisation, caching, batching.

Backend, Data & Storage: Python (FastAPI, Flask, Django, Pydantic, pytest, gunicorn, Celery), Node.js (Express, NestJS, Hono, BullMQ, Pino, ioredis), TypeScript (Zod), PostgreSQL + pgvector (HNSW, halfvec), Redis, MongoDB Atlas, ClickHouse, Snowflake, BigQuery, DuckDB, Elasticsearch/OpenSearch, Kafka, RabbitMQ, Apache Spark, Airflow, Prefect, Dagster, dbt, Pandas, NumPy, Polars, SQLAlchemy, Alembic, Prisma ORM, Drizzle ORM, gRPC, REST, GraphQL, WebSockets, Server-Sent Events.

Cloud, DevOps & Observability: AWS (EC2, S3, Lambda, RDS, DynamoDB, SQS, EventBridge, CloudFront, IAM, ECR), GCP, Azure (Functions, Cognitive Services), Docker (multi-stage, GPU images), Kubernetes, Helm, Argo CD, Terraform, AWS CDK, Pulumi, GitHub Actions, GitLab CI, Jenkins, Vercel, Hetzner, DigitalOcean Spaces, nginx, PM2, OpenTelemetry, Sentry, Datadog, Grafana, Prometheus, ELK stack, Loki, Jaeger, HashiCorp Vault, pre-commit.

Security & Adversarial: OWASP Top 10, MITRE ATT&CK, Burp Suite Pro, Nmap, Metasploit, SQLMap, OWASP ZAP, Nuclei, Wireshark, tcpdump, Splunk SIEM, Snyk, Semgrep, SonarQube, Bandit, Trivy, AWS Security Hub, GuardDuty, STRIDE/PASTA threat modeling, SOC 2, PCI DSS, NIST CSF, CIS Benchmarks, TLS 1.3, AES-256, RSA/ECDSA, SAML, OIDC, JWT, OAuth 2.0, HMAC-SHA256.

Computer Vision, Spatial & Analytics: FastAPI, OpenCV, YOLOv8, EXIF, GeoPandas, Shapely, QGIS, Google MyMaps, spatial reasoning, route mapping, image-processing pipelines, ground-truth labelling, precision/recall gates, React, Next.js, Flutter, React Native, Tailwind, Recharts, Three.js.

Additional strengths: Production system design, retrieval evaluation, AI workflow automation, agentic pipelines, scalable backend engineering, cross-functional leadership, and shipping reliable products in high-stakes environments.