RAJ CHHAPARIYA

AI / Data Engineer · Data & LLM Systems · Machine Learning

Bengaluru, India · rajchhapariya8@gmail.com · Portfolio ↗ · GitHub ↗ · X ↗ · LinkedIn ↗

PROFESSIONAL SUMMARY

AI / Data Engineer building and evaluating data pipelines, retrieval-augmented LLM systems and reproducible benchmarks. Python, SQL and TypeScript, with SQL query guardrails, evidence-gated RAG, and streaming and federated-learning benchmarks. Full-stack work across PostgreSQL, DuckDB, MongoDB and Next.js.

TECHNICAL SKILLS

Data & ML:	Python, SQL, DuckDB, Pandas, NumPy, Scikit-learn, PyTorch, Parquet, SMOTE
LLM Systems:	Retrieval-Augmented Generation (RAG), OpenAI API, BM25, RRF, Pydantic, Evaluation Design
Pipelines & Infra:	Kafka API (Redpanda), Terraform static analysis, PostgreSQL, Supabase (RLS), MongoDB, REST APIs, pytest
Full-Stack:	TypeScript, Next.js, React, Node.js, Streamlit, Git

FEATURED AI & DATA SYSTEMS

Agentic RAG Knowledge Assistant `[Python, OpenAI API, BM25, NumPy, Pydantic, RRF]` GitHub ↗ — Aug 2026 – Present

- Built a RAG pipeline that plans queries, retrieves, audits evidence sufficiency, and answers with citations. Weak evidence triggers bounded retries or refusal. Dense retrieval and BM25 are fused with **Reciprocal Rank Fusion (RRF)**.
- Evaluated 8 configurations on a **32-question held-out test set**: 24 answerable, 4 unanswerable and 4 false-premise questions. Answer accuracy is calculated over the 24 answerable questions; false-premise results are not reported here. These are **frozen pre-fix** results at portfolio scale, not a large-sample benchmark.
- **Full agentic system:** 11.0/24 (45.8%) accuracy; refused 4/4 unanswerable questions; 3/24 false refusals; 55.2% citation precision; 17.69 s average latency.
- **Naive dense baseline:** 14.0/24 (58.3%); refused 0/4; 0/24 false refusals; 40.6% citation precision; 1.87 s.
- **Hybrid baseline:** 14.5/24 (60.4%); refused 0/4; 0/24 false refusals; 43.8% citation precision; 1.86 s.
- 90 test functions, 87 run offline by default.

AI Data Analyst Agent `[Python, DuckDB, Streamlit]` GitHub ↗ — Aug 2026 – Present

- Built an agent that answers plain-English dataset questions through four fixed tools (query, chart, statistics, clarify). SQL is limited to single SELECT/WITH statements by regex keyword guardrails (35 blocked keywords) on in-memory DuckDB, and a deterministic number-matching check compares the figures in each answer with the tool output.
- Selected the correct tool in all 20 questions in a 20-question benchmark; answer correctness was 17/20. Single small benchmark.
- Streamlit UI with CSV upload, automatic profiling, charts and report export.

Taxi Stream Benchmark `[Python, Redpanda (Kafka API), Parquet, DuckDB]` GitHub ↗

- Benchmarked at-least-once against **effectively-once** delivery into a Parquet lake queried with DuckDB, on 2.87M cleaned NYC Yellow Taxi rows (Jan 2024). Effectively-once is not Kafka exactly-once, because Kafka transactions cannot coordinate with a filesystem sink.
- Under injected SIGKILLs (3 seeds, 10 kills per run, 3 runs per configuration; 30,000-row emulator runs and a 6-partition Redpanda run), at-least-once left a mean of 93–3,531 duplicate rows by configuration. Effectively-once left 0 duplicates and 0 missing rows. This is a single-machine benchmark, not a production guarantee.
- Offline compaction cut 803 files to 1; queries ran 14.3× faster (emulator) and 9.3× faster (Redpanda).

EdgeFed-Bench `[Python, PyTorch, DuckDB, Parquet]` GitHub ↗

- Benchmarked Multi-Krum, Coordinate Median, Trimmed Mean and FedAvg with INT8 gradient quantization on Fashion-MNIST (20 clients, 50 rounds, 5 seeds, 280 robustness runs).
- INT8 cut per-client uplink by 75% (397.5 to 99.4 KB). Multi-Krum accuracy was unchanged under 20% sign-flip (84.90% vs 84.89% in float32); FedAvg fell to 10.0% at 30% sign-flip. Single host, small MLP.

tf-misconfig-scanner `[Python, python-hcl2, SARIF]` GitHub ↗

- Built a static Terraform scanner for AWS: 12 rules, Table/JSON/SARIF output, UNKNOWN reported instead of guessing.
- 33-item independent TerraGoat evaluation: 100% precision and recall, against Checkov's 100% and 66.7%. Small sample, and TerraGoat is from Checkov's vendor.
- Separately, 73 scored items from locally authored fixtures, written alongside the scanner and so weak evidence, also scored 100%; 24 UNKNOWN cases are excluded from precision and recall.

RAJ CHHAPARIYA

AI / Data Engineer · rajchhapariya8@gmail.com · Portfolio ↗ · GitHub ↗

APPLIED WEB SYSTEMS

Founder & Independent Developer — Resume Roaster — Jul 2026 – Present

`[Next.js, TypeScript, Supabase (PostgreSQL, RLS), Razorpay]` Live ↗

- Independently designed, developed and deployed a full-stack AI-powered resume platform, including frontend and backend systems, database architecture, AI integration, payment processing and PDF/DOCX generation.
- Codebase implements PDF resume analysis (pdf-parse plus layout heuristics), AI bullet rewriting under a no-repeated-action-verb constraint, Razorpay checkout and webhooks, Upstash Redis rate limiting, and Resend email; availability of individual features to users is not claimed.

JanDarshan `[Next.js, TypeScript, MongoDB, Three.js]` Live ↗ · GitHub ↗ — Apr 2026 – Jul 2026

- Civic information platform for Indian legislators, with a representatives directory (role, party, state and constituency filters), party registries and state/UT dashboards.
- 3D Lok Sabha/Rajya Sabha chamber view, SVG state map, side-by-side comparison and fuzzy multi-token search. Not an official government portal.

Founder & Developer — Truly `[Next.js, Panda CSS, Neon PostgreSQL, Drizzle ORM]` — Apr 2026 – Present

- Built a social video platform backend organised as domain services (auth, bookings, calls, payments, connections).
- Reserves call-slot capacity with an atomic conditional UPDATE (booked count below capacity) inside a database transaction, with a partial unique index against duplicate active bookings, UUIDv7 identifiers and an outbox table.

MACHINE LEARNING

Customer Churn Prediction `[Python, Scikit-learn, SMOTE, Random Forest]` GitHub ↗ — Oct 2025

- Telco customer dataset (7,043 subscribers, 21 attributes), SMOTE plus a Random Forest. The split was 80/20 stratified, giving 1,409 test accounts.
- At a 0.5 threshold: accuracy 77.9%, ROC-AUC 0.823, churn recall 59.0%. Gini feature importance.

SUPPORTING SYSTEMS

• **MandateStack:** subscription billing platform with Razorpay UPI AutoPay and webhook processing, built on Next.js and Supabase (RLS).

• **VyaparStack:** GST invoicing with a CGST/SGST/IGST tax engine, Neon PostgreSQL and pdf-lib. Live ↗

EDUCATION

Bachelor of Technology (B.Tech) in Information Technology
Dr. A.P.J. Abdul Kalam Technical University
2020 – 2024

CERTIFICATIONS

• **Google Data Analytics Professional Certificate** (May 2023)

• **Cloud Computing Certification**, Honeywell (Grade A, Oct 2023)