

RISHAV RAJ

AI Engineer

Patna, Bihar, India | s17992307@gmail.com | +91 7903153411

[LinkedIn](#) | [GitHub](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

AI Engineer building practical systems across RAG, agentic workflows, applied machine learning, and AI-native software. Experienced in retrieval and grounding pipelines, governed AI execution, evidence-based investigation, and Python/TypeScript backends, with a focus on security, testing, and reliable system behavior.

TECHNICAL SKILLS

Programming: Python, SQL, JavaScript, TypeScript

AI / ML: Generative AI, LLM Applications, Gemini, OpenAI, Anthropic, Applied Machine Learning, LightGBM

RAG / Retrieval: Vector Search, Semantic Search, Hybrid Retrieval, Reciprocal Rank Fusion (RRF), FlashRank, Citation Validation, Grounding Validation

Backend / APIs: FastAPI, Node.js, Next.js, REST APIs, Webhooks, Background Workers, Socket.IO

Frontend: React 19, Next.js 16, TypeScript, Tailwind CSS, Zustand, TanStack Query, Vite

Databases / Data: PostgreSQL, pgvector, SQLite, Prisma, SQLAlchemy, Data Validation

Security / Reliability: RBAC, Multi-Tenancy, Tenant Isolation, Prompt Injection Defense, HMAC Verification, Idempotency, Audit Logging

Testing / DevOps: pytest, Vitest, Docker, Git, GitHub Actions, CI/CD, Vercel

SELECTED AI & SOFTWARE PROJECTS

01. Secure Enterprise RAG | *Generative AI / Retrieval System* | [GitHub](#)

Stack: Python, FastAPI, React 19, TypeScript, PostgreSQL 16, pgvector, RRF, FlashRank, Gemini

- Engineered a multi-tenant RAG backend using Python, FastAPI, and PostgreSQL 16, implementing database-level authorization and strict document-level RBAC.
- Integrated hybrid retrieval combining pgvector and PostgreSQL full-text search, enhanced by Reciprocal Rank Fusion (RRF) and FlashRank cross-encoder reranking.
- Designed prompt-injection defenses, untrusted-document delimiters, and grounding validation with explicit refusal behavior for insufficient context.
- Evaluated lexical, vector, hybrid, and reranked retrieval on a curated 10-query/7-chunk benchmark, using results to document retrieval and scaling boundaries.

02. OWNARA-AI | *Governed Agentic AI Platform* | [GitHub](#)

Stack: Next.js 16, React 19, TypeScript, PostgreSQL, Prisma, Node.js, Gemini, OpenAI, Anthropic

- Built a deterministic agentic platform using Next.js 16, TypeScript, and Prisma to orchestrate an Observe -> Reason -> Approve -> Execute -> Audit workflow.
- Integrated a multi-provider LLM gateway with a human-in-the-loop approval gate to validate AI outputs before execution and enforce authority boundaries.
- Developed a Node.js background worker for CSV ingestion and execution processing, maintaining a hash-chained audit trail of execution contracts.
- Implemented idempotent execution and transactional boundaries to prevent duplicate actions and keep AI operations within approved limits.

03. RepoPilot | *AI Software Engineering Intelligence* | [GitHub](#)

Stack: Python, FastAPI, React 19, TypeScript, Tree-sitter, SQLite FTS5, Vector Search, Docker, GitHub Actions

- Architected a modular intelligence backend using Tree-sitter for deterministic AST parsing of GitHub repositories and codebase symbol representation.
- Engineered hybrid retrieval using SQLite FTS5 and vector search, applying deterministic code-aware reranking to prioritize relevant semantic context.
- Enforced grounded RAG with strict citation verification and insufficient-evidence fallbacks; validated the backend with 138 tests.
- Packaged the application with Docker and established automated CI/CD pipelines through GitHub Actions.

04. Enterprise AI Investigation | *AI Orchestration & Analytics* | [GitHub](#)

Stack: Python, FastAPI, SQLAlchemy, React 19, TypeScript, Pydantic, Docker

- Orchestrated a deterministic investigation planner using FastAPI and SQLAlchemy, executing multi-step workflows through a controlled tool registry.
- Integrated read-only SQL tools and document retrieval with an EvidenceStore that records per-run artifacts using SHA-256 hashes and citation identifiers.
- Tested fail-closed root-cause behavior and prompt-injection handling with an offline MockLLMProvider to keep evaluation deterministic and reproducible.
- Validated the system with 207 backend tests (206 passed, 1 skipped), 32 frontend tests, and 6 golden evaluation scenarios.

05. Real-Time Demand & Risk Intelligence Engine | *Applied ML Data Pipeline* | [GitHub](#)

Stack: Python, FastAPI, SQLAlchemy, SQLite, pandas, LightGBM, Streamlit

- Developed an applied ML pipeline using Python, FastAPI, and SQLAlchemy to forecast inventory risks and simulate material runouts.
- Implemented a 40-check data-quality gate that filters and validates relational inputs before analytical processing.
- Trained LightGBM forecasting models and statistical baselines, using chronological splits and leakage-aware evaluation.
- Engineered decision-support endpoints for root-cause attribution, risk analysis, and DC transfer recommendations.

06. Meridian Market | *AI-Enhanced E-Commerce Platform* | [GitHub](#)

Stack: Next.js 16, React 19, TypeScript, Prisma, SQLite, Socket.IO, Zustand, TanStack Query, Tailwind CSS 4, Vitest

- Engineered a multi-vendor marketplace using Next.js 16, React 19, and Prisma with transactional checkout and atomic inventory reservation.
- Secured workflows with HMAC-SHA256 webhook verification, replay protection, and role-based access control.
- Integrated a catalog-grounded AI assistant with deterministic fallback and real-time Socket.IO updates backed by TanStack Query state management.
- Validated application state machines and transactional logic with 75/75 Vitest tests.

ADDITIONAL PROJECTS

- Bhashini Voice-to-Action Gateway - FastAPI/LangGraph backend integrating Bhashini ASR, Claude NLU, PostgreSQL row-locking, and a human confirmation gate.
- Sahayak - FastAPI/React multimodal application integrating Whisper and LLM extraction, deployed via Docker.
- HRIS Import Preview - Django technical exercise for CSV ingestion, identity validation, manager resolution, and O(n) cycle detection.

PROFESSIONAL EXPERIENCE

Business Development Associate | *Achintya Softech Private Limited* | May 2026

- Collaborated with technical and non-technical stakeholders to clarify project requirements and translate them into accurate business documentation.
- Facilitated client communication, stakeholder follow-ups, and structured operational records to keep project objectives aligned.

EDUCATION

B.Tech - Computer Science Engineering (Artificial Intelligence) | Rungta College of Engineering and Technology | 2026

CGPA: 7.0 / 10

CERTIFICATIONS

- GenAI Job Simulation - Forage
- Deloitte Australia Data Analytics Job Simulation - Deloitte
- MERN Full Stack Development - Coding Spoon