

RISHABH SAHU

+91-8965823672 | rishabhsahu1607@gmail.com | [linkedin.com/in/rishabh-sahu](https://www.linkedin.com/in/rishabh-sahu) | github.com/rishabh2023

Professional Summary

AI & Founding Engineer, 3+ years shipping products 0 to 1 — built and launched **3 live, in-market AI products** from scratch (meeting-intelligence, 22-language voice agents, and pharma RAG). Product-minded builder across RAG, LLM agents, and full-stack (React/Next.js, Python/FastAPI, AWS/GCP/Azure) — I decide what to build, then ship it fast.

Technical Skills

Gen AI & ML: LangChain, LangGraph, RAG Pipelines, Agentic Workflows & AI Agents, LLM APIs (OpenAI, Gemini, Anthropic), LLM Evals (RAGAS, LangSmith), Prompt Engineering, MCP (Model Context Protocol) Servers, Pinecone, pgvector, Vector & Hybrid Search, Cross-Encoder Reranking, OCR, Voice AI Agents (Pipecat)
Frontend: React.js, Next.js, TypeScript, JavaScript, Tailwind CSS, Redux
Backend: Python (FastAPI, Django, Flask), Node.js, TypeScript, REST APIs, Microservices, Celery
Cloud & DevOps: AWS, GCP, Azure, Docker, Kubernetes, CI/CD, GitLab, GitHub Actions
Databases: PostgreSQL, MySQL, Redis, Vector Databases

Experience

- BDIPlus** **Dec 2025 – Present**
Full-Stack Engineer *India*
- Architecting a scalable microservices platform for enterprise-grade systems to improve release velocity and maintainability
 - Building a standalone Testing Engine and a rule-based orchestrator engine that automate validation workflows and coordinate service-level execution flows
- Cufront Healthcare Technologies** **Jan 2024 – Dec 2025**
Founding Engineer *India*
- Built an EHR SaaS platform end-to-end (0 to 1) — patient records, billing, and scheduling — as the founding engineer, using React, FastAPI, and AWS, and took it from prototype to production for real clinics
 - Built GenAI features including an OCR prescription parser, a voice AI agent, and a drug-recommendation assistant using LangChain and LLM APIs, cutting manual data-entry time for staff
 - Owned deployment and reliability on AWS (Docker, CI/CD) and partnered with founders on the product roadmap, aligning engineering work with growth priorities
- Vegento Technologies Pvt Ltd** **Aug 2023 – Dec 2023**
Software Engineer Intern *Maharashtra, India*
- Containerized microservices with Docker and Kubernetes, improving deployment consistency across SaaS projects
 - Built and optimized backend APIs in Python (Django, FastAPI) with third-party service integrations
- Qureal AI** **Oct 2022 – Aug 2023**
Software Engineer Intern *Navi Mumbai, Maharashtra, India*
- Developed a drag-and-drop workflow builder in React for a SaaS automation platform, improving usability for non-technical users
 - Built reusable frontend components and contributed to backend services in Python (Django/FastAPI), improving consistency and reducing development time

Key Projects

- Whylegder — Meeting Intelligence Platform** | *Next.js, Python, LLMs, RAG, Vector Search* **2025 – Present**
- Built and shipped end-to-end (0 to 1) a live, fully usable AI platform (sign up and log in to use it, not a demo) that turns meetings into searchable memory, capturing decisions, commitments, and open questions with exact supporting quotes across Google Meet, Zoom, and Teams [\[Live — try it\]](#)
 - Implemented a RAG chat layer to query meeting history, auto-generated pre-meeting briefs and follow-ups, and multilingual transcription across 22+ Indian languages including mixed-language calls
- Bolora — AI Voice Agent Platform** | *Next.js, Python, Pipecat, LLMs, Sarvam AI, Redis* **2025 – Present**
- Building a production voice-AI platform that handles inbound and outbound business calls 24/7 across 22+ Indian languages, streaming real-time STT → LLM → TTS with CRM integrations and per-call analytics [\[Live\]](#)
 - Engineered a self-learning agent layer (post-call FAQ learning with a human review queue) and resilient entity-context caching that fails open, keeping LLM tool-calls reliable under transient infrastructure faults
- Klinavo — Pharma Intelligence Platform** | *Django, DRF, PostgreSQL, pgvector, Celery, React* **2025 – Present**
- Building a pharma-intelligence platform that aggregates clinical-trial data, industry news, and SEC regulatory filings into one searchable RAG interface with sub-2-second query latency
 - Designed a two-layer architecture (Celery ingestion pipeline feeding hybrid dense + sparse retrieval with RRF fusion and cross-encoder reranking over pgvector) plus a deterministic entity spine for grounded, cited answers
- Loan Support AI — RAG System** | *Python, LangChain, Pinecone, LLM APIs* **Sep 2025**
- Built a finance-domain RAG assistant with PDF ingestion, embeddings, and vector search for accurate, source-grounded answers [\[GitHub\]](#)

Education

Mansarovar Global University **Bhopal, India**
Bachelor of Technology in Computer Science *2021 – 2025*