

JAI RAJSINGH

Backend & AI Engineer | Node.js • RAG Systems • ML
Distributed Systems | LLM Applications | Cloud (AWS)

✉ jairajsingh28july@gmail.com | 💻 [JAIRAJ28](#) | 🌐 [/jai-raj-singh-1491](#)

☎ 6291508824

📍 Bandel, West Bengal, 712104

TECHNICAL SKILLS

Core Stack: JavaScript, TypeScript, Python, Node.js, React, Next.js, Zustand, Redux Toolkit, Tailwind CSS, REST APIs, Microservices, Socket.IO, JWT/OAuth, MongoDB, PostgreSQL, Redis, FastAPI.

GenAI & ML: RAG, LangChain, LangGraph, Embeddings, Scikit-learn, Cross-Encoders, Linear Regression, Machine Learning, NVIDIA - (NeMo Guardrails), RAGAS

Databases: MongoDB, PostgreSQL, Redis, ChromaDB, Pgvector, Milvus

DevOps & Tools: Docker, AWS (EC2, S3, Lambda), Nginx, PM2, CI/CD, Linux, Git, GitHub, Jira, Kubernetes

WORK EXPERIENCE

SOFTWARE DEVELOPER I – AI/BACKEND | NUVO AI, MERIL GROUP

| APR 2026 – PRESENT

Tech: Python, FastAPI, PostgreSQL, pgvector, Redis, Kafka, LangGraph, LangChain, PyMuPDF, Tesseract OCR, Sentence Transformers, Cross-Encoder, RAG, LLMs, Amazon Bedrock, Gemini, OpenAI-Compatible APIs, Docker, Kubernetes, pytest

- **Built a production-scale Agentic AI** document intelligence platform using Python, FastAPI, PostgreSQL/pgvector, Redis, Kafka, LangGraph, and RAG for **PDF research, semantic search, claim validation, grounded Q&A, citation tracing, and evidence-backed reports.**
- **Engineered an asynchronous, coordinate-aware document ingestion pipeline** using PyMuPDF and Tesseract OCR for extraction/OCR fallback, context-aware chunking, embedding generation, and pgvector indexing while preserving **page geometry, bounding boxes, metadata**, and source coordinates for precise highlighting and evidence traceability.
- **Designed a multi-stage hybrid RAG retrieval pipeline** combining Sentence Transformers, dense embeddings, keyword/BM25 scoring, pgvector similarity search, and **Cross-Encoder reranking**, improving **retrieval precision, evidence relevance, and LLM grounding** across long technical documents.
- **Developed LangGraph/LangChain agentic research workflows** with planner, search, evidence, contradiction, citation, judge, and reviewer agents, supporting **conditional routing, parallel execution, retries, stateful orchestration, and Human-in-the-Loop approval.**
- **Built a provider-agnostic LLM inference, guardrail, and evaluation layer** across Amazon Bedrock, Gemini, OpenAI-Compatible APIs, and local models with **structured outputs, model fallback/routing, LLM-as-Judge, citation validation, hallucination/unsupported-claim detection, unanswerable handling, schema checks, required-fact coverage, and retrieval/reranking evaluation.**
- **Scaled distributed AI workloads** with Kafka workers, backpressure, retries/DLQs, async fan-out, and Redis caching, rate limiting, and idempotency.
- **Optimized AI inference** with batch embeddings, query caching, concurrency controls, model tiering, and context limits, improving **throughput, latency, token usage, and cost.**
- **Productionized AI services** with Docker, Kubernetes, logging, correlation IDs, monitoring, pytest, and CI/CD.

BACKEND & GENAI ENGINEER – SUPERCEUTICALS PVT. LTD

| AUG 2023 – MAR 2026

Tech: Node.js, Express.js, Python, FastAPI, MongoDB, Mongoose, Redis, Socket.IO, REST APIs, JWT, OAuth 2.0, RBAC, AES-256-GCM, React.js, Next.js, Redux Toolkit, Zustand, Tailwind CSS, LangChain, RAG, ChromaDB, Hugging Face Transformers, PyTorch, Scikit-learn, Ollama, Docker, Nginx, PM2, AWS EC2, WhatsApp API, SMS, PhonePe

- **Engineered scalable distributed backend and microservices** with Node.js, Express.js, MongoDB, and Redis, supporting 10,000+ DAU, 99.9% uptime, and 65% performance improvement through **caching, API optimization, and scalable system design.**
- Built **production-grade GenAI/RAG systems** using Python, FastAPI, LangChain, ChromaDB, Hugging Face, PyTorch, and Ollama, implementing **document ingestion, embeddings, vector search, semantic retrieval, re-ranking, intent classification, prompt orchestration, guardrails, and LLM inference.**
- **Integrated AI-driven healthcare workflows** for grounded prescription, medication, test, and diagnostic-report explanations using **RAG, contextual retrieval, safety validation, and hallucination-control guardrails.**

- **Architected secure, real-time APIs and event-driven systems** using REST, Socket.IO, Redis, JWT, OAuth 2.0, RBAC, AES-256-GCM, and rate limiting for live **queues, appointments, consultations, notifications, and multi-role workflows.**
- **Productionized and deployed AI, backend, and frontend services on AWS EC2** using Docker, Nginx, and PM2, enabling high availability, automated deployments, **reverse proxying, monitoring, process management, and zero-downtime releases.**
- **Led and mentored a 15-member engineering intern team**, driving scalable architecture, code quality, performance optimization, reusable components, state management, and seamless **AI-backend-frontend integration.**

EDUCATION & CERTIFICATIONS

MASTER OF COMPUTER APPLICATIONS (MCA) LOVELY PROFESSIONAL UNIVERSITY, PUNJAB	2024 -PRESENT
BACHELOR OF BUSINESS ADMINISTRATION (BBA) TECHNO INDIA UNIVERSITY, WEST BENGAL	2018 -2021
FULL STACK WEB DEVELOPMENT MASAI SCHOOL (CERTIFICATE)	2022 -2023