SUNNY RAJ

📍 Hamirpur, U.P| ✉ rajputsny22@gmail.com | 📞 +91 9198834159 🌐 [Portfolio](#) |🔗 [LinkedIn](#) | 💻 [GitHub](#)

AI Engineer | Full-Stack Developer

Engineer skilled in building **production-grade AI systems** and **real-time full-stack applications**. Experienced in designing end-to-end ML pipelines, developing scalable web platforms, and deploying containerized microservices on AWS EC2. Passionate about building cloud-ready products that combine performance, usability, and intelligence.

Skills

- **AI / ML:** RAG Systems • LLM Integration (Gemini API) • TensorFlow • Keras • OpenCV • Scikit-learn • Vector Databases • Embeddings • Prompt Engineering
- **Full-Stack Development:** React • Next.js • TypeScript • Tailwind CSS • Node.js • Express.js • PostgreSQL • MongoDB • Prisma • WebSockets • WebRTC • REST APIs
- **DevOps & Tools:** Docker • GitHub Actions (CI/CD) • AWS EC2 • NGINX • Postman • Linux

Education

Kamla Nehru Institute of Technology, Sultanpur

B.Tech, Information Technology (CGPA: 8.0)   **Sep 2023-Jun 2027**

Projects

[RAG-Based AI Assistant](#)   **Aug 2024 – Sep 2024**

- Designed and implemented a **Retrieval-Augmented Generation (RAG)** pipeline to answer queries from long-form video/audio content with high contextual accuracy.
- Built a complete data flow: **speech-to-text → embeddings → vector search → context-aware prompt generation**, integrated with Gemini API.
- Developed modular pipelines for **data ingestion, indexing, semantic retrieval, and interactive Q&A**, achieving **<2 sec average response time**.
- Optimized embedding search latency and integrated dynamic prompt generation for improved relevancy.

[Metaspace](#)   **June 2024 – Sep 2024**

- Engineered a **browser-based 2D metaverse platform** enabling **10,000+ concurrent users** for real-time collaboration, chat, and video calls.
- Added **Gemini AI-powered avatar reactions** and context-based assistants, boosting engagement by **40%**.
- Deployed **containerized microservices** (Docker + PostgreSQL) with automated **CI/CD on AWS EC2** (99.9% uptime).
- Designed a scalable, fault-tolerant communication architecture leveraging **HTTP** and **WebSocket** protocols, enabling **low-latency** (<200ms) streaming and ensuring high availability across platforms

Position of Responsibility

Placement Cell Member, KNIT Sultanpur 2024 – Present

- **Coordinated** campus placement activities, organized 10+ recruitment drives, and guided 500+ students in interview preparation and resume building, improving successful placements by 20%.