Gautam Rana

+91-9664513886 gautamelon@gmail.com linkedin.com/in/gautam-rana-rdx github.com/beingPro007 rdxtreme.vercel.app

Technical Skills

AI and ML: PyTorch, HuggingFace, VLLM, Retrieval-Augmented Generation (RAG), ONNX Runtime, CLIP, Zero-Shot Classification, RealESRGAN, Vision Transformer (ViT), Qwen 2.5-VL, OpenAI API
Languages: Python, Rust, Go, C++, SQL, Dart
Infrastructure: Docker, Google Cloud Platform (GCP), NVIDIA A100, Modal Labs, AWS S3, Amazon CloudFront, Supabase, Linux
Tools and Frameworks: Next.js, Node.js, Flutter, Git, GitHub, LiveKit, ANNOY Vector Index, IIS Server

Experience

Propelius Technologies **Jan 2025 – Present**
Junior AI and ML Engineer *Surat, Gujarat*

- Architected an end-to-end MLOps pipeline with fail-safe architecture that ingests product catalog PDFs, processes them through Qwen 2.5-VL hosted on NVIDIA A100, and outputs structured data, fully replacing third-party API dependencies.
- Implemented VLLM PagedAttention for model serving, reducing per-page inference latency from 5 minutes to 20 seconds, a 15x speedup, while maintaining output quality at production scale.
- Wrote production-grade Python across the full pipeline including PDF ingestion, model inference, post-processing, and structured output with error recovery and retry logic.
- Designed a Dockerized modular workflow for document extraction and image upscaling using RealESRGAN, ensuring consistent and reproducible deployments across staging and production environments.

Geeks For Geeks **Jul 2024 – Jan 2025**
Technical Content Writer *Remote*

- Authored 20+ technical articles on Git, JavaScript, and System Design, collectively reaching 8,000+ page views.

Projects

Wallee – AI Powered Wallpaper App | *Flutter, Rust, Python, CLIP, RealESRGAN, ONNX Supabase, AWS S3, Amazon CloudFront, Modal Labs* **Dec 2025 – Present**

- Built and published a full-stack Android wallpaper app on Google Play Store with 500+ downloads, featuring AI-powered image discovery and automated zero-shot classification.
- Engineered a Pinterest batch ETL pipeline processing 500 to 1000 images per run: extracts images from board URLs, runs GPU-accelerated zero-shot classification using CLIP on Modal Labs, and upscales assets via RealESRGAN, fully automated end to end.
- Served full-resolution assets from AWS S3 via Amazon CloudFront CDN with thumbnail variants for fast retrieval, and stored all metadata and category labels in Supabase.
- Extracted the CLIP text encoder, compiled it to ONNX Runtime, and deployed it via a Rust runtime for on-device inference, achieving sub-200ms semantic search embedding generation fully offline on any Android device at zero server cost.
- Published a technical article documenting the edge deployment architecture: How I Built a Lightning-Fast Text Encoder.

Multimodal AI Agent Platform | *Python, RAG, LiveKit, ANNOY, OpenAI API* **May 2025 – Jun 2025**

- Architected a low-latency voice-to-action AI platform handling real-time multimodal inputs including text, audio, and image for domain-specific enterprise tasks.
- Engineered a multi-agent orchestration layer using OpenAI function calling, enabling specialized autonomous agents for Legal and Sales domains to execute complex workflows independently.
- Built a high-performance RAG pipeline using ANNOY vector indexing, achieving sub-300ms document retrieval across a 5000+ document knowledge base.

Developer Portfolio | *Next.js, Vercel* | rdxtreme.vercel.app **Ongoing**

- Designed and deployed a personal portfolio serving as a live showcase of engineering projects, technical writing, and open-source work, built with Next.js and hosted on Vercel.

Education

Uka Tarsadia University **Surat, Gujarat**
Bachelor of Science in Computer Science *Expected: 2026*