

GAUTAM RANA

☎ +91-9664513886

✉ gautamelon@gmail.com

🌐 [linkedin.com/in/gautam-rana-rdx](https://www.linkedin.com/in/gautam-rana-rdx)
rdxtreme.dev

🐙 github.com/beingPro007



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, Go, C++, Rust, Dart, Javascript

Infrastructure: Docker, Google Cloud Platform (GCP), NVIDIA A100, Modal Labs, AWS Supabase, Linux

Experience

Propelius Technologies

Sep 2025 – Present

Junior AI and ML Engineer

Surat, Gujarat

- Architected an end-to-end multimodal product extraction pipeline that ingests product catalog PDFs, processes them using Qwen 2.5-VL on NVIDIA A100 GPUs, and converts unstructured catalogs into structured product metadata for automated e-commerce ingestion.
- Designed a trainable multilingual e-commerce retrieval system combining semantic search, personalized ranking, and a continuous feedback loop that improves retrieval models using historical search interactions, supporting Arabic and English catalogs.
- Replaced a Gemini-based product categorization pipeline with an on-device zero-shot CLIP classifier, reducing inference latency by 200× (10–20 s to 2s) while eliminating per-request API costs and enabling scalable, high-throughput product ingestion.
- Transformed a manual, operator-driven product enrichment workflow into a fully automated, embarrassingly parallel pipeline by eliminating human intervention across URL ingestion, scraping, and processing. Deployed the architecture on AWS ECS, optimized compute from t3.xlarge to t3.large, achieving over 20× higher throughput while reducing infrastructure costs by nearly 40 percent.

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 - Wallpaper App | *Flutter, Rust, Python, CLIP, RealESRGAN, ONNX*

Supabase, AWS S3, Amazon CloudFront, Modal Labs

Dec 2025 – Present

- Built and published a full-stack AI-powered Android wallpaper application on the Google Play Store, organically acquiring over 1,000 downloads and sustaining 1,500+ monthly active users.
- Built a fully automated GPU-accelerated ETL pipeline that ingests Pinterest boards, classifies images using zero-shot CLIP inference, enhances quality with RealESRGAN, and processes 500–1,000 wallpapers per execution.
- Optimized the CLIP text encoder for mobile by converting it to ONNX and deploying it through a Rust inference engine, enabling fully offline semantic search on Android with sub-200 ms embedding generation and zero server-side inference cost.

Multimodal AI Agent Platform | *Python, RAG, LiveKit, ANNOY, OpenAI API*

May 2025 – Jun 2025

- Architected a low-latency multimodal AI agent platform capable of processing real-time text, audio, and image inputs to automate enterprise workflows through conversational interactions.
- Engineered a multi-agent orchestration framework using OpenAI function calling, enabling specialized Legal and Sales agents to autonomously execute complex domain-specific workflows.

Livekit Go Agent SDK | *Go, Livekit, PCM*

Ongoing

- Developing an open-source Go SDK for LiveKit voice pipeline sessions, implementing agent session management, voice pipeline orchestration, and a modular plugin architecture for building production-ready AI voice agents.

Search Anything | *Python, Pytorch, Transformers, LoRA, Loop Engineering* | [Search Anything](#)

Ongoing

- Designed a trainable multilingual e-commerce retrieval system inspired by Amazon's *Graph-Based Multilingual Product Retrieval in E-Commerce Search*, combining Graph Convolutional Networks (GCNs), SPLADE sparse retrieval, learned ranking, and query understanding for dynamic search routing.
- Improved ranking quality by increasing nDCG from **0.70**→**0.80** over a BGE-M3 baseline using GCN-based retrieval and from **0.72**→**0.80** using SPLADE, while retrieving the top-30 products in approximately **800 ms**.

Education

Uka Tarsadia University

Surat, Gujarat

Bachelor of Science in Computer Science

May 2026