

GAUTAM RANA

✉ gautamrana581@gmail.com

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

🐙 github.com/beingPro007



rdxtreme.dev

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

Communicate technical details to a non-technical audience.

Experience

Propelius Technologies

Sep 2025 – Mar 2026

SDE Intern – AI Team

Surat, Gujarat

- Developed an automated PDF product-catalog processing pipeline that converts unstructured multi-page catalogs into structured product data, including product names, model codes, descriptions, pricing, and associated images.
- Built an end-to-end vision-language extraction pipeline using PyMuPDF for page processing, Qwen2.5-VL for structured product extraction, and automated product-region detection and image cropping.
- Integrated Real-ESRGAN for product-image enhancement and Supabase for structured data and image storage, with GPU inference deployed on Modal using NVIDIA A100s for scalable batch processing.

Propelius Technologies

Mar 2026 – Present

SDE 1 – AI Team

Surat, Gujarat

- Automated a manual, operator-gated product pipeline (URL ingestion, scraping, enrichment) into an embarrassingly parallel fleet of independent AWS ECS Fargate tasks, removing all human intervention and a 2-job concurrency ceiling; cut end-to-end processing of a 4,000-product catalog from 2 days to 2 hours (over 20×) while reducing infrastructure cost by 49 percent via right-sizing compute from t3.xlarge to t3.large.
- Replaced Gemini-based product categorization with an on-device, LoRA-fine-tuned bge-base-en-v1.5 text classifier (title embedding matched against runtime category anchors) deployed on AWS ECS; improved top-1 accuracy from 0.45 to 0.66 (+21 points) on a 366-product held-out test set, cut categorization inference 4× by dropping the image encode and network fetch, and removed the Gemini API call on 70 percent of products.
- 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.

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.

Education

Uka Tarsadia University

Surat, Gujarat

Bachelor of Technology in Computer Science(CGPA:8.0)

May 2026