

# Ritesh Kumar Maurya

✉ riteshkumarmaurya7379@gmail.com | ☎ +91-9119217552 | LinkedIn | GitHub | Website

Machine Learning Engineer with 2.5+ years of experience building Generative AI systems across LLMs, diffusion models, and multimodal AI, with hands-on experience in model architecture, training, fine-tuning, optimization, and deployment.

## WORK EXPERIENCE

### Machine Learning Engineer-II

Zomato, Gurugram

GenAI Team | [Generative AI, Agentic AI]

June 2026 – Aug 2026

- **Multimodal RAG / Agentic AI:** Built production-grade multimodal RAG pipelines and MCP-based Agentic AI workflows using PostgreSQL vector search, multimodal embeddings, and tool orchestration.

### Machine Learning Engineer

Samsung R&D Institute India, Noida

AI Team | [Generative AI, Computer Vision]

Jan 2024 – June 2026

- **Led the end-to-end design, training, and optimization of a proprietary Flow Matching-based cinemagraph generation model:**
  - **Data Pipeline Engineering:** Built a video data pipeline for high-temporal-resolution data (150 frames @ 30fps) with aspect-ratio-preserving strided augmentation.
  - **Custom Architecture Design:** Designed a proprietary 3D Transformer with linear  $O(N)$  attention and trained a Spectral Autoencoder for efficient latent representations.
  - **Training & Distillation:** Implemented Flow Matching and step distillation, reducing inference from **20 to 2 steps** and achieving **10x latency reduction**.
  - **Infrastructure:** Utilized **DeepSpeed** with Accelerate for efficient multi-GPU training, optimizing memory usage for large-batch training.
- **Generative Model Development:** Built and fine-tuned diffusion models (SnapFusion/MobileDiffusion) for text-to-image generation, optimizing U-Net structures to reduce inference latency.
- **On-Device Transformer Optimization:** Engineered the deployment pipeline for T5-XXL (PixArt-Sigma), utilizing Post-Training Quantization (PTQ) and model partitioning to achieve a **4x model compression** using Qualcomm tools.
- **Edge Deployment:** Quantized and exported models using AI-Edge-Torch, AIMET, and QAIRT to TFLite and Qualcomm binaries for real-time mobile inference.

### Machine Learning Engineer Intern

Samsung R&D Institute India, Noida

AI Team | [Python, Deep Learning]

Jan 2023 – Dec 2023

- Developed Transformer and Inception-based models for time-series classification.
- Improved minority-class detection using Angular Loss and Focal Loss on production datasets.

## PROJECTS

### Qwen2.5-0.5B Tool Calling SFT

PyTorch, TRL, PEFT

- Fine-tuned **Qwen2.5-0.5B** with LoRA on **2K xLAM examples**, achieving **93.89% tool accuracy** and **81.58% argument accuracy** on 400 held-out samples.

### nanoFM – Flow Matching

PyTorch

- Implemented Flow Matching with a DiT-based velocity predictor from scratch, including training and sampling.

## PATENT

- **Co-inventor:** Patent filed for generative image editing and inpainting workflows.
- **Co-inventor:** Patent filed for gesture based cinemagraph generation.

## CORE SKILLS

Languages: Python, Java

**Frameworks:** PyTorch, TensorFlow, Hugging Face (Transformers, Diffusers, Accelerate), DeepSpeed

**Models:** LLMs, Diffusion Models, Flow Matching, Autoencoders

**Model Development:** Architecture Design, SFT, LoRA, PEFT, Distributed Training, Distillation, Quantization, Compression

**Deployment & Tools:** TensorFlow Lite, AIMET, QAIRT, Linux, Git

## EDUCATION

Bachelor of Technology in Computer Science & Engineering

July 2019 – June 2023

Indian Institute of Information Technology Kota