Rehan Alam

+91-7281988705 rehanalam5090@gmail.com LinkedIn GFG

Professional Summary

Generative AI Engineer with 1.8+ years of experience specializing in LLM-driven NLP systems for healthcare. Designed and deployed RAG-based medical assistants, fine-tuned large language models using SFT and DPO, and built secure multi-tenant AI platforms used in production. Strong focus on system design, LLM evaluation, latency optimization, and safety-critical AI development.

Technical Skills

LLM & NLP (Applied): Retrieval-Augmented Generation (RAG), Prompt Engineering, Embeddings, Semantic Search, SFT, DPO, LLM Evaluation (faithfulness, relevance, grounding), NLP Pipelines
LLM (Conceptual / Exposure): RLHF, PPO, Preference Optimization Pipelines, Model Alignment Concepts
Frameworks & Libraries: LangChain, Hugging Face (Transformers, PEFT), PyTorch, TensorFlow, BentoML, FastAPI
Databases & Search: Milvus, Pinecone, Elasticsearch, FAISS, PostgreSQL, MySQL, Redis, Neo4j
Security & Reliability: Prompt Injection Mitigation, Input Validation, RBAC, JWT Authentication, Secure API Design

Professional Experience

Meril Life Sciences — June 2024 – Present

Generative AI Engineer — *Vapi, Gujarat*

- **OrthoAI (Surgical Decision Support):** Designed and deployed an end-to-end AI platform for orthopedic fracture detection, leveraging **Faster R-CNN (ResNet-50 FPN)** and **Vision Transformers (ViT)** to achieve clinically reliable recall for femoral and tibial shaft fractures.
- Built scalable data engineering pipelines to process and augment **200K+ medical images** using **Albumentations**, implementing automated train/validation/test splits to ensure robust model generalization.
- Owned the **LLM and NLP layer** for medical document understanding and decision support, selecting **Retrieval-Augmented Generation (RAG)** over full fine-tuning to improve factual reliability and traceability.
- Fine-tuned domain-specific LLMs using **Supervised Fine-Tuning (SFT)** and **Direct Preference Optimization (DPO)**, improving medical document processing accuracy by **40%** and reducing clinician review effort.
- Designed semantic retrieval pipelines using **custom chunking strategies** and **Milvus/Elasticsearch hybrid search**, enabling scalable retrieval across large medical document corpora.
- Developed **NuvoBot**, a **task-oriented multi-agent architecture** separating retrieval, reasoning, and action execution to improve reliability, debuggability, and controlled tool usage.
- Built and deployed **FastAPI-based LLM services** with asynchronous request handling, achieving **99.9% service uptime** and reducing end-to-end query latency by **35%**.

Coding Ninja — August 2023 – June 2024

Full Stack Developer — *Gurgaon, Haryana*

- Engineered large-scale web applications using JavaScript and Python, supporting 50,000+ concurrent users with high-availability architecture.
- Built ML-based monitoring system for transaction anomalies, reducing fraudulent activities by 60% and improving financial security.
- Optimized MySQL and Redis data pipelines, reducing response time by 35% in high-traffic workflows with WCAG-compliant interfaces.
- Led Agile sprints, conducted code reviews, and collaborated cross-functionally to ship features and resolve production bugs.

Professional Achievements

- Solved 700+ problems on GeeksforGeeks, 400+ on HackerRank, and 160+ on LeetCode, demonstrating strong algorithmic problem-solving skills.
- Ranked 3rd globally in Coding Ninja Codekaze 2023, competing against 250,000+ participants.
- Secured 5th global rank in the ServiceNow Hire-Thon 2023, showcasing exceptional technical prowess.
- Achieved 2nd global rank on GeeksforGeeks, outperforming 80,000+ participants in competitive programming challenges.
- Successfully contributed security fixes to Open Interpreter (65K+ stars), addressing critical vulnerabilities across multiple Python modules.

Education

RGPV University, Madhya Pradesh — May 2018 – Jun 2022

B.Tech in Computer Science — *CGPA: 7.9*