Aman Verma

AI Engineer | Machine Learning | GenAI Specialist

LinkedIn | GitHub

sunnyaman331@gmail.com

+91 7236947401

Bangalore, India

Summary

AI / GenAI Engineer with 3 years of experience building production-grade LLM systems for Fortune 500 clients, delivering 80% cost reduction and supporting 100K+ users. Specialized in **LLM fine-tuning, hybrid RAG, agentic workflows**, and secure full-stack delivery on Kubernetes/OpenShift, achieving up to 45% retrieval gains and strong improvements in model reliability.

Experience

- **IBM Watsonx** — Bangalore, India
  AI Engineer — *Aug 2023 – Present*
  - **Platform Architecture**: Designed and deployed production-grade GenAI platforms across healthcare, telecom, and public sector, supporting 100K+ users and reducing operational costs by 80% through enterprise automation.
  - **LLM Fine-Tuning & Optimization**: Enhanced LLM performance by fine-tuning IBM Granite 13B, improving NER gains by 35%, reduced GPU memory footprint by 25%, and packaged models for OpenShift deployments.
  - **RAG & Agentic Pipelines**: Built hybrid RAG systems (BM25 + dense vectors + cross-encoder reranking) over Milvus and Elasticsearch with 1M+ embeddings, improving retrieval precision by 30–45% for complex queries.
  - **Multimodal Automation**: Implemented Whisper STT + OCR + LLM pipelines processing 10K+ forms monthly for 100+ field officers, reducing manual operations by 70% and automating document + speech workflows across departments.
  - **Enterprise Deployment & Reliability**: Orchestrated GPU-backed LLM microservices on Red Hat OpenShift with CI/CD, autoscaling, monitoring, and hardened security, sustaining 99.9% uptime in air-gapped environments.
  - **Governance, Evaluation & Client Impact**: Established PDP guardrails and GDPR/SOC 2 compliance with 98% DLP precision; executed 5,000+ evaluation tests improving reliability by 20%; delivered $2M+ ROI.
- **IBM Expert Labs** — Bangalore, India
  Software Developer Intern — *Jan 2023 – Jul 2023*
  - Built an ML pipeline for fraud detection, enabling review of 700K+ insurance claims monthly with 99.99% availability.
  - Pioneered the implementation of an OCR and fuzzy-matching workflow, boosting reconciliation accuracy by 40% and slashing manual intervention by 10+ weekly hours for the entire reconciliation team.
  - Managed CI/CD, observability, and integrations for a semantic-similarity model, enabling 4x faster deployments.

Technical Skills

- **LLMs & GenAI**: GPT, Claude, Llama, Mistral, Granite; RAG, agentic systems, prompt engineering; LangChain, LangGraph, LlamaIndex, CrewAI; LoRA/QLoRA; Python, FastAPI, React, TypeScript, Node.js, async microservices.
- **AI/ML & Data**: ML, DL, NLP, CV; PyTorch, HuggingFace; Milvus, Elasticsearch, PostgreSQL, MongoDB, Redis; BM25, hybrid retrieval; Pandas, SQL.
- **MLOps & Cloud**: Docker, Kubernetes, OpenShift; CI/CD; MLflow; Prometheus, Grafana; AWS, GCP.
- **Security**: OAuth2, JWT, RBAC; PII masking, DLP; GDPR, SOC 2, PCI-DSS compliance.

Education

- **Indian Institute of Information Technology (IIIT) Ranchi** — Ranchi, India
  B.Tech (Hons.) in Computer Science and Technology — *Aug 2019 – Jun 2023*

Projects

- **CrewAI Resume Analyst | CrewAI, FastAPI, Gemini API**: Developed a GenAI agent that improved resume parsing accuracy from 70% to 90%, helping recruiters focus on candidate engagement instead of manual data entry.
- **AI Portfolio Advisor | FastAPI, Gemini API, Docling**: Extracts portfolio holdings from PDFs/screenshots, normalizes data, benchmarks returns, and generates insights via LLM APIs.
- **HackType | ReactJS, Node.js**: Gamified typing simulator with leaderboard and real-time feedback.
- **MovieFlix / TVFlix | ReactJS**: Movie discovery app using the TMDB API with caching and lazy loading for faster browsing.

Achievements

- Star of the Month – IBM (Dec 2024): Recognized for leading ASEAN GenAI pilots.
- Excelled in Google Hash Code 2022, securing a position within the top 2% globally
- Qualified for Meta HackerCup 2021 Round 2 (Top 1,500 globally).
- Solved 800+ DSA problems on platforms like LeetCode and GFG,enhancing algorithm design skillset