Sukhmandeep Singh

AI/ML Engineer — RAG, LLM Systems & Backend

Pune, Maharashtra, India | +919915890547 | sukhnansaran@gmail.com | linkedin.com/in/sukhmand-singh | github.com/sukhmansaran

SUMMARY

AI/ML Engineer with 1+ year of experience building retrieval-augmented generation (RAG) and LLM-based systems — self-hosted LLM inference (Ollama), vector retrieval (Chroma, FAISS), reranking, citation grounding, and output guardrails — and FastAPI services for LLM tool-calling and agentic pipeline execution. Experienced shipping AI systems with Docker on AWS and HPC infrastructure, and collaborating with research and engineering teams.

PROFESSIONAL EXPERIENCE

AI Engineer — Nucleovir Therapeutics Pvt. Ltd. *03/2026 – Present | Pune, Maharashtra*

- Built a **FastAPI** service exposing an internal AI pipeline as callable tools, enabling **LLM-driven agentic invocation** and on-demand execution — with automated tests covering tool-call correctness and error handling.
- Built an **automated end-to-end AI pipeline** spanning data ingestion, model inference, candidate evaluation, and ranking, with validation, error handling, structured logging, and reproducible execution, in collaboration with computational biology and R&D teams.
- Converted CPU-bound protein structure-prediction workflows to **GPU-accelerated** pipelines using **batched inference** and **multi-worker CPU parallelism**, reducing processing latency and compute resource usage.
- Deployed and orchestrated pipelines across **AWS, SLURM, and Nextflow**, containerized with **Docker** and version-controlled with **Git**, enabling reproducible runs across local, HPC, and cloud environments.

Machine Learning Research Engineer (NDA Project) — NIT Warangal *08/2025 – 02/2026 | Warangal, Telangana*

- Deployed lightweight ML/DL models for fully offline classification and regression across **Android, Raspberry Pi, and NVIDIA Jetson Nano** devices; reduced model size **10MB → 3MB** via quantization, achieving **230–350ms** end-to-end inference latency.
- Built **Dockerized** inference pipelines and mentored **10+ interns** across ML, deep learning, IoT, and real-time sensor integration.

Data Science Intern — Evoastra Ventures Pvt. Ltd. *06/2024 – 08/2024 | Remote*

- Built end-to-end ML pipelines for fraud detection using **XGBoost and Random Forest**, achieving a **ROC-AUC of 0.76** through feature engineering, hyperparameter tuning, and cross-validation.

PROJECTS

Full-Stack RAG Platform (FastAPI, Next.js, Celery) *2026 – Present*

- Architected a full-stack platform — a **FastAPI** REST backend, a **Next.js/React/TypeScript** frontend, and a **PostgreSQL/Redis/Celery** data and task layer — covering authentication, chat, ingestion, and crawling services.
- Eliminated third-party LLM API dependency by building a fully self-hosted **RAG** pipeline — **Ollama** local inference, local embeddings, and **Chroma** vector retrieval — with a custom reranker weighting results by source authority, recency, and geographic match.
- Added **citation grounding**, algorithmic confidence scoring, and **input/output guardrails** — including validation of streaming responses — to control what reaches the model and what is returned to the user.
- Automated content ingestion and monitoring with asynchronous **Celery** crawler, ingestion, and change-detection pipelines, backed by an automated **pytest** suite.

Biomedical RAG Research Assistant — Alzheimer's Literature (FastAPI) *03/2026*

- Built a local, CPU-only **RAG** assistant over Alzheimer's research papers — **pdfplumber** ingestion, section-aware chunking, **sentence-transformers** embeddings, **FAISS** indexing, and local **phi3** inference via **Ollama** — served through a **FastAPI** API.
- Built a multi-stage retrieval flow — query enhancement, section-weighted reranking, embedding-based deduplication, and a per-paper diversity cap — with a filter that removes reference-list noise from chunks.
- Added a **hallucination guard** (key-term overlap with the source corpus), confidence scoring, and an **evaluation script** covering retrieval relevance, answer grounding, citations, and safety behavior.

SKILLS

Generative AI & LLM Systems: RAG Pipelines, LLM Tool-Calling / Agentic Workflows, Local LLM Inference (Ollama), Vector Search (Chroma, FAISS), Embeddings (sentence-transformers), Reranking, Guardrails & Citation Grounding, Prompt Engineering

Backend & Full-Stack: Python, FastAPI, REST APIs, Next.js, React, TypeScript, PostgreSQL, Redis, Celery, SQL, pytest

ML & Deep Learning: PyTorch, TensorFlow, Scikit-learn, XGBoost, Transformers, NumPy, Pandas

Cloud & Infrastructure: AWS, Docker, Git, GPU/HPC Computing (SLURM, Nextflow), Parallel Computing, ONNX, Model Quantization, Edge Deployment

EDUCATION

Bachelor of Technology in Information Technology

Guru Nanak Dev Engineering College, Ludhiana, India | 08/2020 – 06/2024