SAHIL KUMAR

sahilrathi.k@gmail.com Sahilkumar-r sahil-rathi-k +92-333-2530664

AI ENGINEER

PROFESSIONAL SUMMARY

AI Engineer focused on building production-ready LLM systems: RAG pipelines with hybrid retrieval, reranking, systematic evaluation, and guardrails. Experienced in LLM fine-tuning (LoRA/QLoRA, SFT, DPO) and knowledge distillation to build compact small language models (SLMs), both text-only and multimodal. Strong foundation in system design and API development (FastAPI, GraphQL), shipping scalable, observable AI services with CI/CD and LLMOps best practices.

EDUCATION

BS Data Science **2022 – 2026**
FAST (NUCES) University **CGPA: 3.84/4.00**

- **Key Courses:** Machine Learning, Deep Learning, Agentic AI, MLOps, Computer Networks, Big Data Analytics.

EXPERIENCE

Lab Demonstrator (DSA) & Teaching Assistant (Deep Learning, PDC) **Jan 2024 – Jun 2026**
FAST NUCES

- Conducted Data Structures & Algorithms labs on complexity analysis, trees, graphs, and dynamic programming with hands-on implementation.
- Deep Learning TA: mentored students on CNNs, RNNs, and Transformers in PyTorch, covering training, regularization, and tuning; graded projects.
- PDC TA: taught multithreading, MPI, OpenMP, and CUDA, and reviewed assignments on distributed training and scaling compute-heavy workloads.

PROJECTS

Distilled Small Language Model via QLoRA & Preference Alignment *PyTorch, PEFT, TRL, bitsandbytes, vLLM*

- Distilled a Llama-3.1-8B-Instruct teacher into a 1.5B-parameter student SLM, combining logit-level (KL-divergence) and sequence-level distillation on teacher-generated responses.
- Fine-tuned the student with QLoRA (4-bit NF4) in two stages: instruction tuning (SFT), then DPO alignment.
- Retained most teacher quality on MMLU, GSM8K, and LLM-as-judge evals at a fraction of the size; served via vLLM.

Production-Ready Multimodal RAG Chatbot *LlamaIndex, Qdrant, Whisper, FastAPI, Redis, RAGAS*

- Ingested thousands of sources (technical docs, PDFs, PPTs, audio, video) via Whisper and semantic chunking.
- Improved retrieval with hybrid search (BM25 + dense), query rewriting, HyDE, and cross-encoder reranking, raising context precision and faithfulness.
- Added RAGAS + LLM-as-judge evals in CI and guardrails for prompt injection, PII redaction, and hallucinations.
- Deployed as a Dockerized FastAPI service with streaming, semantic caching, and source citations.

Agentic Multimodal Document Intelligence System *Qwen2-VL, LangGraph, FastAPI, GraphQL, DVC, MLflow*

- Fine-tuned a small vision-language model with LoRA to extract structured data from invoices, forms, charts, and tables, outperforming OCR + LLM baselines at lower inference cost.
- Built a LangGraph multi-agent workflow (router, extractor, validator) with tool calling and human-in-the-loop review.
- Designed microservices exposing REST and GraphQL APIs with async job queues; versioned data and models with DVC/MLflow and automated deployment via GitHub Actions CI/CD.

TECHNICAL SKILLS

- **AI Engineering:** LLM Fine-Tuning (LoRA, QLoRA, SFT, DPO), Knowledge Distillation, SLMs & Multimodal Models, Advanced RAG (Hybrid Search, Reranking, GraphRAG, Vectorless RAG), Agentic AI, LLM Evals (RAGAS, DeepEval), Guardrails (NeMo, Llama Guard)
- **Frameworks & Libraries:** PyTorch, Hugging Face (PEFT, TRL), LangChain, LangGraph, LlamaIndex, vLLM
- **Vector Databases & Storage:** Qdrant, FAISS, Chroma, pgvector, PostgreSQL, NoSQL (MongoDB, Redis)
- **APIs & System Design:** FastAPI, GraphQL, REST, Microservices, Async Processing, Caching
- **MLOps & CI/CD:** Docker, Kubernetes, GitHub Actions, MLflow, Weights & Biases, Model Serving & Monitoring
- **Version Control:** Git, GitHub, Bitbucket, DVC
- **Languages:** Python, SQL, C++

ACHIEVEMENTS

- **Gold Medalist (3x)**