Shivang Kumar

Email - shivangkumar1857@gmail.com
Phone - +91-8923207202
Website - www.shivangkumar.com
Linkedin - @kumar-shivang
Github - @kumar-shivang

PROFESSIONAL SUMMARY

AI System Engineer specializing in LLM-based systems, agentic pipelines, and real-world AI deployments. Experienced in building end-to-end solutions including reasoning pipelines, OCR post-processing with SLMs, and real-time simulation engines. Focused on shipping scalable AI products with measurable impact.

EXPERIENCE

System Engineer, VaultIQ.ai, Vadodara, GUJARAT

June 2026 - Present

- Building full-stack systems for an AI-driven knowledge retrieval and RAG platform.
- Developing scalable document ingestion, chunking, embedding, indexing, and retrieval pipelines.
- Implementing contextual search, reranking, and knowledge graph-based retrieval workflows.
- Designing backend APIs and frontend features for enterprise knowledge management use cases.
- Integrating generative AI models into production-grade retrieval and automation workflows.
- Profiling and optimizing API performance, retrieval latency, and data pipeline scalability.

Data Scientist, IIT Mandi. Mandi, HIMACHAL PRADESH

Dec. 2025 - May 2026

- Designed data pipelines and backend for an educational app integrating personalization features and recommendation engine.
- Increased the data retrieval accuracy by 21% by using knowledge-graph to detect relationships between various subjects.
- Built a simulation generation engine based on 3js powered by an LLM to generate STEM simulations on the go using a custom simulation engine built using an agentic pipeline.
- Built AI observability infrastructure on Langfuse, Grifana and Prometheus to observe AI agent performance in real time and get alerts on bottlenecks.

AI Engineer Intern, Read-Ink Technologies Private Limited. Bangalore Rural, KARNATAKA

August. 2025 - November. 2025

- Built a heuristic-based OCR post-processing pipeline to improve OCR accuracy for handwritten documents.
- Used Small Language Models (SLMs) to correct OCR errors in handwritten text.

AI Engineer Intern, Livup.ai. Bangalore, KARNATAKA

January. 2025 - August. 2025

- Built pipelines for multi-turn reasoning for large language models (LLMs) to solve software bugs (to compete on SWEBench Benchmark).
- Analyzed solution trajectories to identify LLM failures and improve performance.
- Built internal tools for monitoring and evaluating LLM performance on various benchmarks.

EDUCATION

Indian Institute of Technology, Madras, Chennai

Bachelor of Science, Data Science and Applications, September. 2025

SKILLS

Large Language Models - Deployment, Observability and Finetuning, Data Analysis and Feature Engineering, System Design, Model Optimization and Training, Retrieval Augmentation Generation (vanilla and GraphRAG)

TECH STACK

- Python, PyTorch, Huggingface
- Typescript, Bun/Node, React/Vue
- LLMs (OpenAI, VLLM, OpenRouter)
- LangChain, LangGraph, LangFuse, MCP
- FastAPI, Docker
- Postgres (Pgvector for VectorDB, Apache AGE for graph database), MongoDB

LANGUAGES

Hindi - Native Proficiency
English - Professional Proficiency
German - Beginner