Shivam Kumar Srivastava

+91-7497988948 – thisisshivam18@gmail.com – github

Professional Experience

Software Engineer Intern **Feb 2025 - Jul 2025**

Euron (Remote)

Tech stack: Python, FastAPI, OpenAI, LangChain, FAISS, Docker, AWS

- Built and maintained FinSight, an LLM-powered finance analytics tool used by 8K+ monthly users with 85% retention.
- Implemented a RAG workflow with FAISS for semantic search over 1M+ docs with automated deployment pipelines.
- Designed LangChain pipelines with prompt templates and conversation memory, reducing bad responses by 35%.
- Secured APIs with JWT/RBAC, structured logging, and caching, optimized preprocessing for 30% faster data handling.

Key Projects

SmartPrep AI: Adaptive E-Learning SaaS Github **Mar 2025 - Jul 2025**

Python, FastAPI, Groq, Lang Chain, FAISS, GCP, Docker, Kubernetes

- Implemented adaptive RAG with metadata-aware prompts and per-user context, FAISS powered conversational memory and fast recall of prior turns for quiz personalization.
- Built retrieval-augmented chat with prompt guards and output validation that reduced off-topic drift by 35%; deployed on Cloud Run with lightweight telemetry for latency and hit-rate tracking.

Interactive Story Generator (AdventureEscape) GitHub **Aug 2025 - Sep 2025**

Python, FastAPI, LangChain, LLMs, Image APIs, Analytics, PostgreSQL

- Engineered a Python/LangChain pipeline to leverage external LLM, generating dynamic, branching narratives associated with visual assets per node, enabling scalable AI-driven content creation and adaptive gameplay features.
- Integrated a real-time analytics system to capture player choices and calculate critical business metrics (e.g unfinished, completion and winning rates), establishing a foundation for player retention modeling and personalization.

Flipkart Product Recommender GitHub **Jan 2025 - Mar 2025**

Python, Hugging Face, LangChain, Astra DB, Groq, Flask, GCP, Prometheus, Grafana, Docker

- Ingested product reviews, applied transformer-based chunking/embeddings, and indexed vectors in Astra DB for low-latency vector search, Agentic RAG on Groq produced explainable, source-based outputs with cited snippets.
- Exposed a Flask inference service with query routing and instrumented Prometheus metrics (latency, token usage) along with Grafana dashboards, containarized via Docker and deployed on GCP with automated monitoring.

NetworkGuard ML: Real-Time Phishing Detection GitHub **Nov 2024 - Jan 2025**

Python, Apache Kafka, Scikit-learn, FastAPI, AWS, MLflow, Docker, CI/CD

- Built Kafka streaming ETL processing 50K+ events/hour with 99.5% data quality, ensembles reached 94% precision/92% recall with p95 inference upto 100ms, monitored in MLflow.
- Implemented real-time phishing detection with Kafka messaging, 30-feature extraction, containerized deployment on AWS with automated CI/CD pipelines and observability stack.

Technical Skills

Programming:	Python, C++, SQL
Backend/APIs:	Flask, FastAPI, REST API, OpenAI, Claude, Groq
DevOps/Cloud:	Docker, Kubernetes, AWS, GCP, CI/CD(Github Actions)
Databases:	SQL, MongoDB, Vector Databases (ChromaDB, Astra DB, FAISS)
Monitoring/QA:	Prometheus, Grafana, MLflow, pytest, Postman
AI/Ml Frameworks:	PyTorch, Keras, Hugging Face Transformers, LangChain, LangGraph, Scikit-learn
Generative AI:	LLMs, Retrieval Augmented Generation, Fine-tuning, MCP, A2A Protocol
Tools:	Git, n8n, Streamlit, Kafka

Education

B.Tech., Computer Science Engineering (AI & Data Science) **CGPA: 8.0/10**

Bennett University, Greater Noida 09/2022 – 05/2026

Certifications/Achievements

- **3rd Place, GFG Hackathon (SVIET, Chandigarh).**
- **Top 5 finalist (BU), Smart India Hackathon 2024.**
- Solved **300+ DSA problems on GeekforGeeks**, Active Kaggle Contributor.
- **Certifications:** DeepLearning Data Structure Object Oriented Design Machine Learning
- **Head of Tech:** CodeChef Chapter: organized 15+ contests; mentored 100+ students.