SATYAM SHIVAM

AI Engineer · LangChain · RAG · Multi-Agent Systems
shivamsatyam35@gmail.com · +91 9852015381 · New Delhi · LinkedIn · GitHub · Portfolio

Three production internships building AI applications that ship to enterprise users. Developed a production RAG chatbot (500+ daily queries, 99.2% uptime), cut PDF query latency 79% on a live pipeline, and built MCP-integrated agentic systems with CI/CD and Docker. LangChain, FastAPI, AWS, RAGAS. Graduating June 2026.

TECHNICAL SKILLS

AI/GenAI & LLMs: LangChain · LangGraph · LlamaIndex · RAG · Hybrid Search (BM25 + Dense) · Agentic AI · CrewAI
LLM Stack: OpenAI API · Groq · Hugging Face Transformers · sentence-transformers · Prompt Engineering · RAGAS · LLM Evaluation · MCP
Vector Databases: FAISS · ChromaDB · Pinecone · Neo4j · Semantic Search · Cross-Encoder Reranking
Backend & Cloud: FastAPI · Python · Node.js · AWS (EC2, S3) · Docker · GitHub Actions CI/CD · Async · Microservices
Other: PostgreSQL · React · Pydantic · pytest · JavaScript · TypeScript · OpenTelemetry

EXPERIENCE

Asvix · *AI Developer Intern* Jan 2026 – Apr 2026

- Shipped DigiLab, an enterprise AI assistant answering domain-specific queries on custom medical datasets, to production: **500+ daily queries at 99.2% uptime** using LangChain, FAISS, and Neo4j hybrid RAG with all-MiniLM embedding model
- Reduced hallucination rate **39% (18% to 11%)** using context-aware LangChain response modules and prompt engineering; improved medical query relevance by 23% with BM25 and dense retrieval
- Built a RAGAS evaluation harness tracking faithfulness, context recall, and context precision; scores fed two rounds of prompt refinement and became the team's quality gate for new model versions

Cloudily Scripts · *AI Chatbot Development Intern* Jun 2025 – Jul 2025

- Built a production RAG pipeline (FAISS IVF128 indexing, cross-encoder reranking, BM25 dense retrieval) that processed 100+ page PDFs at **91% accuracy** and cut support tickets 35%
- Cut query latency **79% (8.2s to 1.7s)** with parallel embedding and semantic chunking; Dockerized the stack and reduced image size 60% (2.1 GB to 840 MB)

IPtechhub · *Cloud Engineering Intern* May 2024 – Jul 2024

- Deployed containerised ML inference services on AWS EC2 with auto-scaling; handled 500+ daily requests at 99.5% uptime and cut infrastructure costs 32%
- Automated CI/CD via GitHub Actions; deployment time dropped **87.5% (2 hrs to 15 min)**; cold-start improved from 45s to 8s

PROJECTS

Production RAG Pipeline | *LangChain · FAISS · FastAPI · Docker · Groq · RAGAS · OpenTelemetry*

- 4-agent pipeline: Retriever, Guardrail Agent, Generator, Evaluator. RAGAS harness tracks faithfulness, context recall, and context precision. OpenTelemetry traces and structured logs throughout. Dockerized with GitHub Actions CI/CD and tenacity retry logic.
- OpenTelemetry observability layer for latency tracking, error rates, and agent performance monitoring
- Hybrid BM25 and dense retrieval with cross-encoder reranking and semantic chunking; **89% relevance vs 67% baseline**; 850ms mean latency

Multi-Agent Campaign Creator | *Python · CrewAI · LangGraph · Groq · Pydantic · pytest · GitHub Actions*

- 4-agent system (Research, Copywriter, Art Director, Manager) with a **LangGraph state machine** managing transitions and 4 specialised tools (TrendResearch, CompetitorAnalysis, CopyEvaluation, ImagePromptGenerator). Generates campaigns at 94% lower cost ($2,400 to $150) than manual production
- 92% test coverage across 25 tests; Pydantic models throughout; GitHub Actions CI runs lint, type-check, and tests on every push

HybridAI Syntax Error Detection | *Python · AST · scikit-learn · Gradient Boosting · FastAPI · Streamlit*

- Dual-mode detection across 5 languages: AST rule analysis runs first; the Gradient Boosting classifier handles the rest. **94.18% accuracy, Cohen's kappa 0.79, 1ms median inference**. The system degrades gracefully to rule-based mode when the ML model is unavailable. Targeting Q3 2026 IEEE submission.
- Three interfaces: FastAPI server, Streamlit app (omnisyntax.streamlit.app), CLI; 48+ commits with full test report and API docs in docs/

Customer Churn MLOps Pipeline | *XGBoost · FastAPI · Docker · scikit-learn · model-monitoring*

- End-to-end ML pipeline: EDA, feature engineering, 3-model benchmark (LR 0.849, RF 0.861, **XGBoost 0.868 ROC AUC**), FastAPI REST API, Docker. Batch prediction endpoint with data drift monitoring.
- Reproducible by design: synthetic dataset, Docker, and GitHub Actions CI mean any engineer can clone and run the full pipeline in under 10 minutes
- EDA and feature engineering on structured customer data; 3-model benchmark proving XGBoost superiority (0.868 ROC AUC) before production selection

EDUCATION

B.Tech Computer Science & Engineering, AI & ML Specialization 2022 to 2026
United Institute of Technology, Gandhinagar
Coursework: Data Structures & Algorithms, Machine Learning, Deep Learning, NLP, Database Systems, Cloud Computing