SUDHANSHU SRIVASTAVA

Senior AI Engineer | LLM Systems & RAG Specialist | Full Stack Developer

New Delhi, India • +91 85778 68240 • sri.sudhanshu1@gmail.com • LinkedIn • GitHub

PROFESSIONAL SUMMARY

Senior AI Engineer specializing in **production RAG systems, LLM infrastructure, and multi-agent orchestration** with 3+ years of end-to-end ownership across LLM integration, scalable backend architecture, and full-stack delivery. Proven track record: **35% inference latency reduction, 85%+ RAGAS faithfulness in production, 40% pipeline failure reduction, 25% LLM cost reduction, and 20% organic traffic growth.** Experienced driving AI feature delivery as the sole AI layer owner covering architecture, prompt engineering, and RAG pipeline tuning through deployment, evaluation, observability, and cost optimisation. Proficient with LangChain, LangGraph, LlamaIndex, OpenAI APIs, Hugging Face, LangSmith, and open-source LLMs (Llama 3, Mistral) via vLLM and Ollama.

TECHNICAL SKILLS

AI & LLM Engineering: RAG Pipelines, LangChain, LangGraph, LlamaIndex, OpenAI APIs (GPT-4, structured outputs, function calling, tool use), Hugging Face Transformers, Open-source LLMs (Llama 3, Mistral via vLLM/Ollama), Prompt Engineering, Embeddings, AI Agents, Multi-agent Orchestration, Multi-modal Generation, Fine-tuning (LoRA/QLoRA), RAGAS Evaluation, A/B Testing, Inference Optimization

LLM Observability & MLOps: LangSmith, Arize AI, Helicone, Model Monitoring, Guardrails, LLM Tracing, Experiment Tracking, MLOps Pipelines, CI/CD (GitHub Actions)

Vector Databases: Pinecone, ChromaDB, FAISS, pgvector

Backend & APIs: Node.js, Express, Python, FastAPI, REST APIs, Microservices, Async Job Queues, Distributed Systems, System Design, Scalable Architecture, SLA/SLO Ownership

Frontend: React, Next.js, TypeScript, Tailwind CSS, Core Web Vitals, SEO Optimisation

Cloud & DevOps: AWS (EC2, S3, Lambda, DynamoDB), Docker, Kubernetes (k8s), GitHub Actions

Databases: PostgreSQL, MySQL, MongoDB, DynamoDB, Redis

Tools & Collaboration: Git, GitHub, Bitbucket, Postman, Jira, Agile, Scrum, Technical Leadership, Cross-functional Delivery, Stakeholder Communication

PROFESSIONAL EXPERIENCE

Senior AI Engineer | Gaup Media Pvt Ltd

Feb 2026 – Present

New Delhi, India

- Owning end-to-end architecture of a **production-grade AI campaign reporting system** integrating Instagram and LinkedIn data pipelines via **multi-agent orchestration** (LangGraph-style agent graphs); architected ingestion → enrichment → reporting pipeline targeting sub-200ms p95 report generation.
- Driving **cross-functional AI delivery** as sole AI layer owner across data engineering and product teams including sprint planning, stakeholder reviews, and all technical decision-making for the LLM layer.
- Architected **distributed Node.js backend APIs** for automated analytics workflows with fault-tolerant async ingestion; reduced pipeline failure rate by **40%** in early production through robust error handling for API failures, rate limits, and schema drift.
- Implemented **LLM observability and tracing** using LangSmith for real-time monitoring of agent execution, token usage, latency, and failure modes across multi-agent workflows.

Senior LLM Engineer | Digitally Next

Apr 2024 – Feb 2026

New Delhi, India

- Led end-to-end delivery of **production RAG pipelines** using OpenAI GPT-4, LangChain, and vector search (Pinecone, ChromaDB) serving **~5K daily queries** across 3+ client products; drove RAGAS faithfulness from 62% to **85%+** through semantic chunking, cross-encoder reranking, and metadata filtering.
- Owned **LLM cost optimisation strategy** via intelligent request routing, response caching, and prompt compression reduced per-query API cost by **25%** at sustained traffic; full model monitoring and observability instrumentation via Arize AI.

- Designed and executed **A/B tests** on prompt variants and retrieval strategies; drove **15% increase in user engagement** and measurable session duration improvement using structured outputs and function calling patterns.
- Architected scalable REST APIs integrating ML models with PostgreSQL, MongoDB, and Next.js frontends; **defined SLA/SLO targets** and owned production incident response for AI feature layer.
- Reduced application load time by **30%** through React performance tuning, lazy loading, code-splitting, and optimised API contracts with direct improvement to Core Web Vitals scores.

Software Engineer | Happymonk AI
Mar 2023 – Apr 2024

Bengaluru, India (Full Time, promoted from internship)

- Promoted to full-time; **owned delivery of 4 client projects** spanning fintech and SaaS using React, Node.js, and AWS.
- Mentored 3 junior developers on React architecture and backend best practices; reduced PR review cycles by **25%** and elevated team code quality standards.
- Drove **20% growth in organic traffic** and **12% uplift in conversions** through structured data, lazy loading, and Lighthouse-based Core Web Vitals optimisation.

Software Engineer Intern | Happymonk AI
Dec 2022 – Mar 2023

Bengaluru, India (Internship)

- Delivered 15+ production web applications using React, Node.js, and AWS (EC2, S3, Lambda) across fintech, e-commerce, and SaaS clients.
- Built Express-based microservices with DynamoDB and S3 for scalable storage; implemented AI-driven automation pipelines for analytics and personalisation.
- Contributed to codebase quality through unit testing and code reviews within Agile sprint cycles; gained hands-on AWS deployment and CI/CD workflow experience.

KEY PROJECTS

ThreeZinc AI Platform | Multi-Agent Orchestration System |

- Owned architecture of an AI platform orchestrating multiple intelligent agents across Instagram and LinkedIn connectors for automated content generation and analytics using **LangGraph-style agent graphs**; architecture diagram, agent flow, and pipeline reliability metrics documented in GitHub README.
- Led backend connector design using REST APIs consolidating fragmented platform data into a centralised pipeline with robust handling for scraping edge cases, rate limits, and schema drift.
- Instrumented full LangSmith tracing across all agent nodes; documented failure modes and recovery strategies in repo.

Multi-modal AI Content Generation Platform

- Architected unified backend integrating image, video, and 3D generative AI models with async job queues for compute-heavy inference; reduced average inference latency by **35%** via provider-level optimisation and intelligent request routing.
- Achieved **25% cost reduction per 1,000 requests** through response caching and prompt compression; used A/B testing to validate provider routing strategies.

RAG-based AI Chat System | LangChain + OpenAI GPT-4 + Pinecone |

- Production RAG chatbot delivering **85%+ RAGAS faithfulness** (up from 62%) via semantic chunking, cross-encoder reranking, and metadata filtering; benchmark: **under 800ms p95 retrieval latency** at 10K documents, ~$0.003/query with caching.
- Implemented full RAGAS evaluation framework (context precision, faithfulness, answer relevancy) with documented eval runs, test cases, and a LoRA fine-tuning experiment for domain-specific QA available in evals/ folder in GitHub repo.
- Applied **inference optimisation** via quantization experiments (GGUF/Q4_K_M) on Mistral 7B for cost-sensitive deployments; results documented with methodology.

EDUCATION

B.Tech in Computer Science | Lovely Professional University
2020 – 2024

Punjab, India

CERTIFICATIONS

- Generative AI with Large Language Models | DeepLearning.AI & AWS
- Deep Learning Specialization | Coursera (deeplearning.ai)
- IBM Full Stack Application Development | edX / IBM
- Advanced React and GraphQL | Udemy