SAURAV KUMAR

AI / LLM Engineer | Production RAG & Retrieval Systems | IIT Madras

Hyderabad, India | sauravkumar585@gmail.com | +91-7042610302 | linkedin.com/in/sauravkumar1503 | github.com/SK-15 | saurav-portfolio.in

PROFESSIONAL SUMMARY

LLM engineer with 4+ years building production AI systems across legal-tech, fintech, and consumer AI. Built the core search and retrieval infrastructure for DocuSwift, an NVIDIA Inception-backed legal AI platform indexing 500,000+ Indian court judgments, and a near-real-time personalization engine for Kitab, a consumer book-tech platform. Specializes in RAG pipeline architecture (LangChain, LlamaIndex, LangGraph), hybrid vector search (HNSW/IVFFlat), and cost-aware full-cycle deployment on Azure/AWS. Outside full-time roles, independently builds AI systems from first principles, including a hand-built GPT implementation and Koda, a knowledge-oriented developer agent owning everything from architecture to deployment.

TECHNICAL SKILLS

- **LLM & Generative AI:** RAG, Prompt Engineering, Fine-Tuning, LangChain, LlamaIndex, LangGraph, OpenAI Agents SDK, Text Embeddings, Re-ranking Strategies
- **ML & Modeling:** Scikit-Learn, PyTorch, TensorFlow, Model Evaluation (BLEU, F1, ROC-AUC, top-k recall), Supervised & Unsupervised Learning, Hyperparameter Tuning
- **Cloud & Deployment:** Azure (AI Studio, Cognitive Search, Functions, DevOps), AWS (App Runner, Lambda), Docker, FastAPI, Serverless Architecture, CI/CD
- **Databases & Vector Search:** PostgreSQL (HNSW/IVFFlat, RPC Functions), Supabase, Vector Databases, SQL, MongoDB
- **Programming:** Python (primary), JavaScript, Node.js

PROFESSIONAL EXPERIENCE

ML/LLM Engineer — Kitab

Jul 2025 – Present

Remote

- Owned end-to-end design of a near-real-time recommendation engine using Supabase; engineered a dynamic taste-profile system with SQL triggers that update user embeddings via weighted moving averages for personalization that adapts to evolving reading preferences.
- Built and benchmarked a multi-stage LLM summarization pipeline transforming long-form books into wisdom-centric summaries; iteratively optimized prompts against competitor outputs to maximize insight density and engagement.
- Architected the full database schema with custom PostgreSQL RPC functions and hybrid HNSW/IVFFlat vector indexing, sustaining search performance as the content library scales.
- Integrated ElevenLabs TTS to generate AI voiceovers for book summaries, tuning prosody and delivery for a premium, natural-sounding listening experience.

Generative AI Developer — DocuSwift

May 2024 – Jun 2025

Remote · docuswift.in

- Designed Case Chat, a RAG-based chat interface letting lawyers upload a judgment and query it directly auto-generating structured insight sections (facts, case background, issues, final judgment) via Azure Cognitive Search and Azure Functions before opening the document to open-ended natural-language Q&A, with one-click Word export of results.
- Solved cross-document retrieval leakage by indexing user_id and document_id as filterable fields at chunk time, scoping every Cognitive Search query to only the requesting user's own uploaded document; cut perceived response time by prioritizing the three most-viewed sections (facts, issues, case background) and parallelizing both LLM insight generation and document chunking across sections instead of running them sequentially.
- Led development of the Case Search RAG pipeline on Azure Cognitive Search, combining BM25 keyword retrieval with semantic re-ranking to surface jurisdiction-relevant judgments from 500,000+ Indian court cases with high precision and low latency.
- Shipped serverless APIs on Azure Functions with Azure DevOps CI/CD, enabling zero-downtime deployments for a Microsoft for Startups- and NVIDIA Inception-backed legal AI platform.
- Implemented chat-history management, message-feedback persistence, and session lifecycle controls, improving data governance and UX platform-wide.

Data Scientist — Soothsayer Analytics

Dec 2023 – May 2024

- Designed and deployed ML models for client business challenges; tuned hyperparameters to improve prediction accuracy and computational efficiency.
- Conducted EDA on structured datasets; built dashboards and visualizations to communicate model findings to non-technical stakeholders.

- Collaborated cross-functionally with product managers and engineers to align ML outputs with measurable business goals.
- Monitored deployed models for drift and degradation, proactively retraining to preserve production performance.

Data Scientist — Agastya Data Solutions *Jun 2022 – Nov 2023*

- Developed quantitative investment strategies using ML techniques on technical and fundamental data, achieving 12–15% strategy returns.
- Optimized core data-processing algorithms, reducing pipeline execution time by 50% for faster, more reliable analytical outputs.
- Built interactive dashboards and fintech feature prototypes to improve data-visualization workflows for fund managers and investors.
- Fine-tuned LLM prompts for consistent, high-quality outputs in downstream financial-analysis applications.

INDEPENDENT PROJECTS & VENTURES

Koda — *Knowledge-Oriented Developer Agent — coding agent*

- Built a self-hosted, multi-tenant AI coding agent (REST + WebSocket API) that reads, writes, and reasons about a codebase architected as a LangGraph agent loop (agent → tools → agent) with automatic context summarization, Postgres-checkpointed durable threads, and human-in-the-loop approval gates for risky operations (shell commands, code execution, HTTP requests).
- Designed a coordinator mode that decomposes a task and runs it across parallel worker agents, multi-provider LLM routing (Anthropic, OpenAI, Gemini, DeepSeek, Kimi, Ollama) with per-run cost tracking and budget limits, and a file-based memory system that persists context across sessions without a vector store.

github.com/SK-15/Koda

GPT From Scratch — *decoder-only Transformer implemented by hand*

- Built a character-level, decoder-only Transformer in PyTorch from first principles no nn.Transformer or Hugging Face abstractions progressing from a bigram baseline to a full 10.8M-parameter GPT with multi-head self-attention, residual connections, and LayerNorm; trained to generate Shakespeare-style text on consumer GPU hardware.

github.com/SK-15/gpt

EDUCATION

Diploma in Computer Programming — Indian Institute of Technology (IIT), Madras 2021 – 2022
B.Sc. (Hons) in Mathematics — Shaheed Bhagat Singh College, University of Delhi 2016 – 2019