

SAURAV THAKUR

AI Engineer | Agentic AI & RAG Specialist | Full-Stack

[portfolio - sauravthakur.in](#) | Phone +91 6377939533 | sauravthakur6310@gmail.com | [in LinkedIn](#) | [GitHub](#)

PROFESSIONAL SUMMARY

AI Engineer and Full-Stack Developer specializing in production grade **Agentic AI**, RAG architectures, and LLM integrations. Built and scaled AI systems **serving 37,000+ active users** across EdTech, LegalTech, Finance, Healthcare, Travel, and government domains, with end-to-end ownership from architecture to production deployment.

- Built and scaled **46+ production** AI systems and **21+ mobile** applications across **13+ industries**, cutting operational costs by up to 63%, from sovereign government infrastructure to leading AI startups based in **New York, USA**.
 - Implemented a zero-data-egress, **sovereign AI infrastructure** for **government procurement** running isolated local LLMs inside Docker for strict data sovereignty.
-

CORE SKILLS

- **Agentic AI & Multi-Agent Orchestration:** LangGraph, OpenClaw, Claude Code, Codex, CrewAI, AutoGen, PydanticAI, LlamaIndex, LangChain, N8N, Mem0, MCP, Haystack, Browser Use, Playwright
 - **GenAI, LLMs, RAG & LLMOps:** OpenAI, Claude, Grok, Gemini, GraphRAG, Agentic RAG, vLLM, Pinecone, Qdrant, Weaviate, LangSmith, RAGAS, DeepEval, Guardrails AI, Vercel AI SDK, semantic/hybrid retrieval, embeddings, Ollama, AWS Bedrock, Prompt Engineering (System Prompts, Structured Outputs, Function/Tool Calling, Context Engineering)
 - **Programming & Full-Stack Development:** Python, TypeScript, JavaScript, React.js, Next.js, FastAPI, Node.js, Tailwind CSS, WebSockets, REST APIs, GraphQL, Microservices
 - **Multimodal AI:** ElevenLabs, OpenCV, FFmpeg, Whisper, LLaVA
 - **Machine Learning & NLP:** TensorFlow, PyTorch, Keras, HuggingFace, SpaCy, pandas, NumPy, Scikit-learn
 - **Cloud, Databases & DevOps:** PostgreSQL, Redis, MongoDB, Firebase, Supabase, AWS (EC2, S3, Lambda), GCP, Azure, Docker, Kubernetes, CI/CD, GitHub Actions
-

PROFESSIONAL EXPERIENCE

alfred_ | Product Design & Frontend Engineer | New York, USA

Jun 2026 – Aug 2026

- Directly engaged by founder to own the complete redesign of marketing site and brand experience for an AI executive assistant startup serving professionals handling **100+ emails/day**
- Engineered the complete visual system and all conversion-critical surfaces in **Next.js/React**, every design decision built around one principle, if the product removes cognitive load, the brand cannot add to it.
- Delivered comprehensive growth strategy, competitive analysis of **15+ competitor analysis** (Fyxr, Superhuman, Motion), market sizing (\$11.87B to \$43.9B by 2034), ICP scoring, SEO/AEO engine design, 30 ranked growth experiments, and 30/60/90-day execution roadmap.

NEXBOOK | Founder & Lead AI Engineer | Google for Startups '26 | DoT 5G National Selection '26

Aug 2025 – May 2026

- Built and scaled **India's most comprehensive AI powered campus ecosystem**, uniquely integrating PYQs, Notes, Timetables, Results, a peer to peer Book Marketplace & **250,000+** certified free courses (Harvard, MIT, Stanford, Google) into one ecosystem, serving **37,000+ active users** with **85% retention & 98% transaction success**.
- **Selected: Google for Startups Immersion '26 (Google & Antler India)** Google Cloud & Gemini AI startup immersion.
- **Selected: DoT 5G Innovation '26 (Govt. of India, C-DOT)** National Level Presentation Based Evaluation.
- **Synapse AI:** Agentic classroom (LangGraph) AI Teacher + Classmates autonomously teach & debate on live whiteboard, PDF to full lesson (slides, quizzes, HTML5) in <60s; 8+ LLMs, zero API cost via Docker.

- **Thinkly AI:** NLP + RAG learning engine personalized study plans, document analysis, Q&A generation & AI powered PYQ prediction (~87% accuracy across 10+ years data); scaled to 98% uptime via Redis, Firebase, CDN & CI/CD.

GeM AI | Lead AI Engineer | Sovereign Multi-Agent Procurement Engine

Jan 2026 – May 2026

- Architected an **8 Agent autonomous procurement system** (Next.js 14 + FastAPI + pgvector) covering full lifecycle, Tender Parsing, Fraud Detection, Audit, deployable in **<15 min** via Docker Compose.
- Built **dual-engine multimodal OCR** (LLaVA + MinerU, confidence scoring 0–100) solving India's 'blurry smartphone photo' problem, auto triggers forgery flags on cross validation discrepancies.
- Engineered RAG tender parser handling **50–200 page** docs with no context window limits, Pydantic v2 + self-healing retry loop eliminates hallucinations, extracts GFR 2017/CVC rules into validated Master Checklist.
- Deployed **SHA-256 hash-chained immutable audit log**, replaced **2-3 week manual cycles**; achieved **70-85% fraud** detection vs **<10% manually**. Same day RTI/CAG export API.
- Instrumented LangSmith + RAGAS eval pipelines tracking retrieval precision and hallucination rates. **Zero data egress** all LLM inference runs locally inside Docker.

True Trips | Lead AI Engineer & Full Stack Dev | AI Travel Agent

May 2025 – Jan 2026

- Architected a full-stack AI Travel Agent App (10+ modules) - Multimodal RAG Concierge (Gemini/Grok/GCP) generating 7 day itineraries in <4s with budget breakdown, 360° place previews, interactive maps, live hotel/flight booking & YouTube to itinerary parsing via Next.js 14 + Mapbox.
- Built RAG powered travel intelligence engine, crawls & synthesizes global web into verified destination guides real time place cards with Google ratings, weather & Wikipedia, Activities Marketplace with direct booking flow.
- Developed a Collaborative WebSockets/Stripe ecosystem for **500+** users, scaling via an offline-first **PWA** with a **75% latency reduction**, enabling **real time** multi user **scheduling for flights/hotels**, and integrating **smart budget/trend aware recommendation engines**.

SELECTED PROJECTS

- **EDIX - AI Video Engine, (Jun – Aug 2025)**
Timeline-less video engine compressing 4 hour workflows into minutes; 94% footage accuracy, 90% faster exports via distributed GPU orchestration (AWS, Redis), multimodal pipelines with Whisper, OpenCV, FFmpeg, WebSockets.
- **AEROX - Browser, (Jan – Mar 2025)**
Privacy-first agentic browser (Windows/macOS/iOS/Android); custom Electron Chromium IPC bridge, autonomous form-filling & voice navigation 1,000+ users, 4.3/5 rating.
- **Trax - Music Streaming App, (Mar – May 2025)**
Global music infra indexing billions of tracks; sub millisecond WebSocket sync for 1,000+ live Jam rooms, Vibe AI for voice navigated accent-personalized discovery.
- **Jurisight, Lexmont, FermWise, Insauré, YTV, BlindChef, AdMorph (Mar – Dec 2024)**
Production RAG & Agentic AI across LegalTech, InsurTech, AgriTech, MediaTech & AdTech, 80 to 91% latency reduction, 96% precision, 93% user satisfaction.

EDUCATION

Bachelor of Science (B.Sc.) – Physics with Computer Science

S.S. Jain Subodh P.G. College | University of Rajasthan | **Graduated: June 2026**

- **Coursework: Quantum Mechanics, ML & AI - Grade: O (Outstanding)**
- **ML Healthcare Analytics:** Engineered a 92% accurate predictive model **analyzing 5,000+ records** with Flask API
- **Research:** Reframing Bohr's Complementarity - six original quantum experimental architectures for wave particle exclusivity, published on Academia, OSF & Zenodo (Aug 2025), under editorial consideration at **Foundations of Physics (Springer Nature), 2026.**

INTERNSHIPS & CERTIFICATIONS

AI & GenAI – Anthropic (Claude & Prompt Engineering), OpenAI, MIT (GenAI), **Cloud** – AWS, GCP, Azure, NVIDIA,

Security – Cybersecurity, Tata Consultancy Services