

Shashank Saxena

APPLIED SCIENTIST — GENERATIVE AI, MACHINE LEARNING & SOFTWARE ENGINEERING

+91 90117 52453 • shashank.saxena1010@gmail.com • linkedin.com/in/shashanksaxena1010 • Bengaluru, India

PROFESSIONAL SUMMARY

Full-stack AI scientist with 7+ years and an M.Sc. in Mathematics, spanning the complete product surface — **Android, web and backend software engineering** through to **research-grade AI taken to production**. Across banking, healthcare and enterprise I have shipped nearly every class of AI problem: **GenAI** (LLM fine-tuning with LoRA, agentic prompt optimisation, Graph & Agentic RAG, multimodal decisioning, synthetic data via multi-model consensus), **computer vision** (deepfake, forgery and multi-face detection), **NLP, speech, document AI, time series forecasting, recommendation and reinforcement learning**. Owns problems end-to-end — framing, research, architecture, deployment and org-wide adoption — delivering **95–99% production metrics** and **₹30 Cr/year (~\$3.6M)** in measurable impact.

SKILLS

- Generative AI & LLMs (GPT, Claude, Llama, Mistral)
- Agentic AI, Multi-Agent Systems & MCP
- LangChain / LangGraph / LlamaIndex
- Machine Learning & Deep Learning (PyTorch)
- Document AI & Information Extraction
- MLOps & Deployment (Docker, Kubernetes, CI/CD)
- Big Data & Databases (Spark, Hadoop, SQL/NoSQL)
- LLM Fine-Tuning (LoRA, QLoRA, RLHF) & Prompting
- RAG, Graph RAG & Agentic RAG
- Vector DBs & Semantic Search (Milvus, FAISS)
- NLP, Computer Vision & Multimodal Transformers
- Model Evaluation, Quantization & Distillation
- Cloud AI (AWS Bedrock/SageMaker, Azure, Vertex AI)
- Python, Java & System Design

EDUCATION

M.Sc. (Hons.) Mathematics — BITS Pilani 2014–2019
B.E. (Hons.) Electrical & Electronics — BITS Pilani 2014–2019

PROFESSIONAL EXPERIENCE

IDFC First Bank Senior Data Scientist Sep 2024 – Jun 2026
vKYC AI — Research to Production

- Researched and productionized a suite of novel biometric & anti-fraud models for RBI-compliant video-KYC — **third-party speaker validation** (95% recall), **deepfake detection** (85% recall), **multi-face detection** (100% recall) and **PAN/PVC document verification** (99% accuracy) — each taken from problem framing and experimentation to live deployment.
- Integrated Deepgram/Greylabs speech-to-text and built **transcript-based LLM validation** of mandatory financial declarations, replacing manual call review.
- Impact:** Carried research-stage speaker-verification and deepfake models to production serving **3,500–5,000 calls/day**, cutting approval SLA **15 min → 4 min** while strengthening fraud defense.

Document Authenticity & Forgery Detection — Research Lead

- Researched generative approaches (**GANs, VAE, AE**) for detecting forgery in bank and credit-card statements, and designed a **hybrid detection method combining unsupervised clustering with rule-based heuristics** for robustness against unseen tampering; collaborated closely with forensics & risk-containment teams.
- Impact:** Production authenticity layer flagging document fraud earlier in the pipeline, cutting downstream risk exposure.

Document AI Extraction Platform (GenAI)

- Architected a GenAI extraction platform for multi-value entities across diverse banking documents: designed a **three-tier agentic prompt-optimisation system** — a **generator** drafting candidate prompts from annotations, an **evaluator** scoring them against ground truth, and a **reflector** critiquing failures and steering the next iteration — converging to **98%+ production accuracy** without manual prompt tuning.
- Built **synthetic ground truth via multi-model consensus** (agreement across independent LLMs replacing costly human annotation), **LoRA fine-tuned Mistral, Qwen, GPT and Llama** for POI/POA extraction, and shipped auto-prompt-training + structured/tabular extraction modules adopted **team-wide for all LLM tasks**.
- Established the **LLM evaluation & benchmarking harness** (accuracy, hallucination and regression checks) governing model selection and safe rollout across the bank's GenAI use cases.
- Impact:** Estimated **₹30 Cr/year (~\$3.6M/year)** savings through automation and higher straight-through processing.

StatusNeo Consulting Senior AI Scientist May 2023 – Jul 2024
Legal Intelligence — Chat with Contracts (AT&T)

- Built a GenAI contract-understanding platform with chatbot Q&A and semantic search over Milvus + MongoDB; engineered **Graph RAG** — extracting entities/relationships into a knowledge graph for multi-hop, relationship-aware retrieval — and **Agentic RAG** — agents decomposing queries, planning retrieval and self-correcting answers — with dynamic chunking, semantic mapping and embedding pipelines.
- Impact:** Automated contract amendment tracking that previously required manual legal review.

Claims Audit AI (Carelon)

- Led end-to-end delivery of a claims-audit platform — document-AI pipelines, annotation frameworks, a **96% F1** attachment classifier; benchmarked and deployed **LLM + multimodal decisioning (GPT-4, Falcon, Vicuna)** for entity extraction and reimbursement adjudication with production QA gates.
- Impact:** Automated reimbursement audit at scale for a major healthcare payer.

JPMorgan Chase & Co. Application Developer Mar 2021 – May 2023
Financial Forecasting & Expense Intelligence

- Deployed multi-class expense categorization (ensemble, **F1 88%**) and liquidity/cash-flow/Nostro forecasting (LSTM, Prophet, ARIMA) via FastAPI; migrated the Sales Dashboard to JPMC's BRIE data lake with a custom ETL pipeline. **Impact:** treasury cash-position visibility and **50% lower dashboard latency**.

Talentica Software Software Developer Mar 2020 – Mar 2021
User Lifecycle & Payments Platform: led activation/subscription systems and global payment API integrations on a serverless AWS stack. **Impact:** revenue +15% (~\$1,100/day).

Reflexis Systems Software Developer Jul 2019 – Mar 2020
RL Workforce Optimization: built RL models for open-shift generation and a collaborative-filtering recommendation engine (cosine similarity, Top-N) for retail scheduling.