MANISH SHARMA

manish.tinkering@gmail.com GitHub LinkedIn Twitter Kaggle
+91-9342533525 Bangalore, 560068

EDUCATION

B.Tech — Electronics and Communication Engineering *2018 – 2022*
Nitte Meenakshi Institute of Technology **GPA: 8.98**

Research Assistant — Speech and Vision *2021 – 2022*
Indian Institute of Science, Bangalore — SpireLabs

WORK EXPERIENCE

Lead AI Engineer — GenAI *Aug 2025 – Present*
UsefulBI, Bangalore *Website* | *LinkedIn*

- **Multi-Agent Platform Engineering:** Designed and implemented the **Gen AI Studio / CSR multi-agent orchestration** layer using **AWS** and **Strands**, enabling modular agent workflows, task routing, and reusable orchestration patterns for enterprise GenAI use cases.
- **No-Code Agent Workflow Builder:** Built a generic drag-and-drop workflow canvas for composing agent pipelines, connecting configurable agent nodes, and enabling reusable pipelines for **CSR, PLPS, DSUR**, and related operational flows.
- **Agent Evaluation Framework:** Built a structured multi-agent evaluation workflow to benchmark orchestration quality, **tool call, compliance score and accuracy**, response fidelity, and task completion across agentic pipelines using evaluation frameworks, experiment tracking, and comparative testing.
- **IQ Quality Intelligence Platform:** Built an AI-powered **quality-events intelligence system** that uses historical quality events as user-query context with multimodal evidence, stores event representations in a **vector database**, and retrieves similar past events for future quality checks. Owned the end-to-end **User, FAM, and QE workflow**, backend architecture, and **AWS/EKS deployment**.
- **AI Platform Development:** Architected and built the **SUNY course discovery platform** enabling students to explore and receive recommendations across all SUNY campuses and departments.
- **Data Infrastructure:** Designed the **GQMD unified data repository** by integrating **Smartsheet Data Vault** and **GPLM architecture**; built **Databricks pipelines** for centralized storage and processing of qualified materials.
- **Technical Leadership:** Led a team of **4 engineers** on client projects, ensuring timely delivery and high quality of deliverables.
- **AI Hiring:** Conducted **10+ technical interviews** for mid and senior AI roles across **ML, LLMs, and system design**.
- **Hackathon Leadership:** Conceptualized and conducted a **company-wide hackathon**, authored the medical science and pharma domain problem statement, defined judging criteria, coordinated execution, and managed the cash prize structure end-to-end.
- **Knowledge Sharing:** Curated and shared **50+ AI/ML learning resources** across the organization to support continuous learning.

Machine Learning Engineer – 2 *Jan 2024 – July 2025*
Parspec.io, Bangalore *Website* | *LinkedIn*

- **Latency Reduction:** Built a Datasheet Recommendation System for model number-to-datasheet mapping across 5M documents with **80% latency reduction** using **AWS DynamoDB** caching and Gemini 2.0 Flash LLM via OpenRouter.
- **Model Fine-Tuning:** Fine-tuned **Llama 3.3 70B Instruct** on a custom Alpaca-format dataset for attribute extraction on **Modal Labs using H100**.
- **Family Name Extraction:** Helped develop a family name extraction algorithm with significant improvements by integrating the Gemini 2.5 Flash model via the Gemini SDK, boosting **recall from 89% to 96%**.
- **Kubernetes Migration:** Migrated the entire AI-dev Kubernetes workload to **AWS EKS** with guidance from the Snapsoft team. Mapped the load balancer to API Gateway and created isolated partitions for staging and production environments.
- **Order Detection Algorithm:** Designed and deployed an order information detection algorithm for lighting datasheets using T5-base, **achieving 95% accuracy** and automating data extraction.
- **Algorithm Enhancement:** Enhanced header and column detection algorithm, increasing capacity from 4 to **8 columns with 97% accuracy** and reducing manual processing by 30%.
- **LLM-as-Judge Pipeline:** Designed and implemented an LLM-as-a-Judge pipeline to assist human annotation workflows. Leveraged **GPT-4o** and **Gemini 2.5 Flash** in parallel execution to evaluate datapoints for manual annotation, **decreasing human annotation to ∼74%**.
- **Multimodal RAG Pipeline:** Built a RAG pipeline supporting the LLM-as-Judge framework to evaluate retrieved knowledge base chunks. The KB included both images and text, using BGE for text embeddings, CLIP for image embeddings, and **FAISS** as the vector store. **Achieved MRR of 0.94 in retrieval and 0.96 accuracy in generation.**

Machine Learning Scientist *Dec 2022 – Dec 2023*
Docsumo AI, Bangalore *Website* | *LinkedIn*

- Designed wireframes and implemented advanced **Document KV + Table Extractor** using **LayoutLM, BROS, and YOLO** architectures for both fixed and unstructured document categories. Deployed to production.
- Successfully integrated ML & DL architectures into **10+ custom APIs** for clients with MRR in the range of **$80K–$100K**.
- Built and integrated **Chat-AI**, a powerful LLM integration using **LangChain** and **Pinecone DB** for QA and support tasks within the product.
- **Reduced annotation time** from 1 full day to **∼2 hours** using a GPT-KV LLM Extractor powered by GPT-4, significantly reducing human effort.

Machine Learning Engineer *Aug 2021 – Nov 2022*
Zealth-AI (YC W21)

- Built and deployed **OCR extraction** and **medical mapping** flow for medical documents, reports, and prescriptions using **Amazon Textract**, **RxNORM**, **MedXN**, and **SciSpacy**.
- Built a customized Symptom Checker Algorithm using **Neo4j Aura** with **Cypher** to construct an entire **Knowledge Graph Database** with **11k+ relations** and **3k+ nodes** and relationship mappings.
- Leveraged **Rasa NLU** to build a **Medical AI Chatbot** capable of querying problems related to symptoms and providing necessary care management.

PROJECTS / SIDE BUILDS

Rag — Video-Based RAG System GitHub | Loom
Qdrant, RAG, Video-Querying

- Developed a system allowing users to query video content via YouTube URL or uploaded videos.
- Implemented video chunking and indexing with **Qdrant** for efficient vector search.
- Integrated a QA pipeline to retrieve **relevant frames, timestamps, and precise answers**.
- Experimented with 3 vector databases to optimize performance.
- Deployed the solution with **Streamlit** for an intuitive user interface.

AutoCommit Generator GitHub | Twitter
Mistral, Ollama, GitHub, LLM

- Built an AutoCommit Generator using **Ollama** and **Mistral** to automatically generate commit messages for projects locally and quickly.
- Fully **local** solution with no privacy concerns, ensuring secure usage.
- Developed as a **bash script** for quick installation and seamless terminal integration.
- Designed with under **100 lines of code** for simplicity and powerful functionality.

Company Scraper — AI-Powered Agent Agent UI | Twitter
Relevance AI, LLMs, Markdown

- Built a lightweight agent that generates clean, structured summaries from company URLs.
- Auto-extracts: Overview, Products, Features, Audience, Integrations, and more.
- Powered by **Relevance AI** + custom LLM prompts; outputs are markdown-formatted for easy reading.
- Ideal for due diligence, competitor analysis, interviews, and startup research.

TECHNICAL SKILLS

- **Languages & Frameworks:** Python, FastAPI, Flask, GitHub, Cursor, Ollama, OpenRouter, Langchain, Langgraph, LlamaIndex, Strands Agent, Linux, Lovable, HuggingFace, Grok etc
- **AI & Machine Learning:** ML, DL, NLP, PyTorch, TensorFlow, HuggingFace Transformers, Multi-Agents, AWS Bedrock, Azure Foundary, LLMs (GPT series), Fine-tuned Models, RAG, Multi-Agent RAG, VectorDB, GenAI Solutions
- **Natural Language Processing:** NLTK, SpaCy, SciSpaCy, MedXN, Librosa, PyTesseract, OCR, Embeddings
- **Databases & Analytics:** MySQL, Neo4j, Tableau, Amplitude Analytics, Hasura DB, Amazon DynamoDB, S3, GCS
- **Cloud & DevOps:** AWS, GCP, Google Colab, Kubernetes, Docker, EKS, ECS
- **Data Structures & Algorithms:** Strong foundation in problem-solving and optimization