MANISH SHARMA

manish.tinkering@gmail.com Portfolio LinkedIn

+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.
- **AIQE Quality Intelligence Platform:** Built an AI-powered **quality-events intelligence system** that retrieves similar historical events with multimodal evidence as context for new quality checks. Owned the end-to-end **User, FAM, and QE workflow**, backend architecture, and **AWS/EKS deployment**.
- **Agent Evaluation & Regression Gating:** Built the eval harness for the AIQE **Decision Tree Agent** (**Strands + AWS Bedrock**): golden datasets, **tool-call trajectory** graders, **span-level tracing** of tokens, cost and latency, a human-calibrated **LLM-as-a-Judge**, **pass@k / pass**^{k} consistency checks, and a **regression gate** that blocks releases that underperform the baseline.
- **Online Evals:** Converted reviewer decisions into production labels by comparing **AI-suggested vs human-final answers** per decision node, giving a continuous **AI-human agreement** metric with zero extra instrumentation.
- **Agent Memory & Context Engineering:** Designed the AIQE memory layer: role-scoped **short-term memory** in **DynamoDB**, **long-term episodic memory** of past quality events in a **vector database** for grounding, and **audit trails** for every override, keeping agent decisions explainable.
- **AI Platform Development:** Architected the **SUNY course discovery platform** for recommendations across all SUNY campuses.
- **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, owning delivery and quality.
- **AI Hiring:** Conducted **10+ technical interviews** for mid and senior AI roles across **ML, LLMs, and system design**.
- **Hackathon Leadership:** Ran a **company-wide pharma-domain hackathon** end-to-end: problem statement, judging and prizes.

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.
- **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.

PROJECTS / SIDE BUILDS

Manish Code — Terminal Coding Agent Harness GitHub
Python, Agent Loop, Tool Calling, OS Sandboxing, Context Engineering, OpenRouter

- Built a pip-installable **Claude Code-style coding agent** from scratch: one readable agent loop with file, shell, edit, planning and skill tools over any **OpenAI-compatible** model endpoint.
- Enforced safety in two layers: a **permission engine** (read-only commands auto-run, writes ask, destructive commands always blocked, bypass attempts like redirection and command substitution caught) and a kernel-level **OS sandbox** (seatbelt / bubblewrap) with no network, project-only writes and kernel-protected .env secrets.
- Implemented **context management**: capped tool output, turn trimming, automatic **LLM-summarised compaction** at 85% of the context window, and cache-friendly **late injection** of git state, todos and changed files.
- Added **read-only subagents** with isolated context windows, a **todo planner**, on-demand **SKILL.md** skills, and persistent JSONL sessions with resume and rewind.

OpenMemoryUI — Transparent Agentic Memory Glassbox GitHub | Demo
JavaScript, Agentic Memory, LLMs, MCP, LocalStorage

- Built a transparent agentic memory interface that visualizes **session, episodic, semantic, and contextual memory** in real time.
- Implemented a traceable **six-stage pipeline**—Receive, Retrieve, Act, Generate, Extract, and Write with retrieval scores, prompt inspection, provenance, and memory edit history.
- Integrated live **OpenAI, Anthropic, and OpenRouter** models, 20+ browser-callable tools, and runtime connections to Streamable HTTP **MCP servers**.
- Designed the application as a framework-free static web app with browser-local persistence, memory import/export, consolidation, and zero-setup demo mode.

OpenMCPUI — Visual MCP Playground GitHub | Demo
TypeScript, Next.js, MCP, LLMs, LocalStorage

- Built a visual MCP playground where users ask in plain English and watch the client choose tools, prepare arguments, and execute server calls in real time.
- Implemented transparent protocol tracing across connect, **tools/list**, intent routing, and **tools/call**, exposing handshake, tool selection, arguments, and results.
- Added a no-code server workshop that generates MCP TypeScript SDK servers, plus live connections to **OpenAI, Anthropic, OpenRouter, and Gemini**.
- Shipped built-in utility, web, and productivity servers with browser-local persistence for custom servers and a zero-setup demo mode.

TECHNICAL SKILLS

- **Languages & Frameworks:** Python, FastAPI, Flask, GitHub, Cursor, Ollama, OpenRouter, Langchain, Langgraph, LlamaIndex, Strands Agent, Linux, Lovable, HuggingFace, Grok, JavaScript, HTML, CSS, Bash, REST APIs, Streamlit
- **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, Agentic AI, Agentic Memory, LLM Evaluation, LLM-as-a-Judge, Multimodal RAG, Prompt Engineering, Model Fine-Tuning, Retrieval Evaluation
- **Natural Language Processing:** NLTK, SpaCy, SciSpaCy, MedXN, Librosa, PyTesseract, OCR, Embeddings, Semantic Search, Information Extraction, BGE, CLIP, T5, LayoutLM, BROS
- **Databases & Analytics:** MySQL, Neo4j, Tableau, Amplitude Analytics, Hasura DB, Amazon DynamoDB, S3, GCS, Qdrant, FAISS, Pinecone, Vector Databases, Databricks, LocalStorage
- **Cloud & DevOps:** AWS, GCP, Google Colab, Kubernetes, Docker, EKS, ECS, AWS API Gateway, Netlify, Modal Labs, CI/CD
- **LLM APIs & Agent Tooling:** OpenAI API, Anthropic API, Gemini SDK, Model Context Protocol (MCP), Tool Calling, Agent Orchestration, OpenRouter API, Agent Sandboxing, Context Compaction, Agent Tracing & Evals
- **Data Structures & Algorithms:** Strong foundation in problem-solving and optimization