

Shikhar Srivastava

+91 9335031075 | shikharsrivastava3004@gmail.com | LinkedIn | GitHub

SUMMARY

Final-year CS (Data Science) student specializing in **Generative AI and Agentic Systems**, with 9+ months of production experience building LLM pipelines, multi-agent orchestration platforms, RAG systems, and real-time AI tools at Zypp Electric(EV logistics company). Passionate about applying AI to financial services, customer experience, and operational intelligence.

EDUCATION

Bennett University

B.Tech in Computer Science (Data Science) – CGPA: 8.46/10

Delhi Public School

PCM with Computer Science (XII) – 87.2%

Delhi Public School

Secondary Education (X) – 92.2%

Greater Noida, Uttar Pradesh

Sept 2022 – May 2026

Lucknow, Uttar Pradesh

May 2021

Lucknow, Uttar Pradesh

May 2019

EXPERIENCE

Zypp Electric (Rs.2800 Cr EV Logistics, 20,000+ Riders)

AI/ML Engineer Intern

Gurgaon, Haryana

June 2025 – Current

- Built **Zia**, a production multilingual voice + text agentic chatbot serving **20,000+ riders** in 6 languages – LLaMA 4 Scout 17B with dynamic tool calling, RAG (Qdrant), Whisper ASR, GPT-4o TTS, and a **zero-code Tool Manager** (JSON schema discovery) enabling new tools via config drop with no codebase changes; reduced support query load by **40%**.
- Engineered an **automated Call Sentiment Analysis pipeline** processing **1,000+ recordings/day** – speaker diarization (Pyannote AI), multilingual transcription (Whisper Large V3), LLM-based analysis (sentiment, urgency 1–10, resolution status); saved ops team **~15 hrs/week** of manual call review with a Streamlit/Plotly leadership dashboard.
- Developed an **MCP-based Lead Generation Chatbot** (AnythingLLM + FastMCP) capturing and routing sales leads across 6 business programs into Google Sheets with real-time duplicate detection and per-program validation rules across 9 cities – eliminating **100%** of manual sales data entry.
- Optimized **10+ critical SQL dashboards** – rewrote slow queries using indexed CTEs, reducing average query runtime by **60%**; implemented data masking for PII compliance across operational datasets.
- Currently building **spare parts demand forecasting** and an **aesthetic quality control** system using computer vision for EV fleet maintenance.

PROJECTS

AI Factory (Idea 0.1) | Python, FastAPI, PostgreSQL, Redis, Docker, LLMs | [GitHub](#)

- Developed a **production-grade multi-agent orchestration platform** with a DAG-based execution engine supporting complex, branching AI workflows – designed for reliability in high-stakes environments.
- Designed a **3-tier AI governance system** (CPO Council → Managers → Workers) to enforce agent alignment, resolve conflicts, and audit decisions – directly analogous to risk governance in financial AI.
- Implemented **deterministic execution and replay**, runtime safeguards (timeouts, retries, cancellation, rate limiting), and full audit trails for reproducible, compliant AI workflows.

AI Lead Generation Chatbot | Python, FastMCP, AnythingLLM, Google Sheets API | [Live](#)

- Built an **MCP-based agentic chatbot** integrated with AnythingLLM that captures and routes sales leads across **6 business programs** – each program exposed as a discrete MCP tool for context-aware LLM routing.
- Implemented **real-time duplicate detection** (mobile + email normalization) and per-program validation rules across 9 cities; built an **auto-updating analytics dashboard** tracking chatbot-sourced leads per program via marker-based attribution.

Payout Automation Chatbot | Python, LLMs, SQL, FastAPI

- Developed an **AI chatbot for real-time payout query resolution** – integrated live DB lookups for wallet balances, transaction history, and deduction breakdowns; reduced payout ticket handling by **35%**.
- Automated repetitive payout escalations using **intent-detection logic** – if a rider repeats the same query 3 times, the system auto-raises a support ticket.

TECHNICAL SKILLS

AI / GenAI: LLMs, Agentic AI, RAG Pipelines, Tool Calling, MCP (Model Context Protocol), Prompt Engineering, NLP, Computer Vision, Speaker Diarization

Frameworks: LangChain, LangGraph, FastAPI, PyTorch, TensorFlow, HuggingFace, OpenCV, Scikit-learn, Streamlit

Languages: Python, SQL, JavaScript (Node.js), TypeScript, C++

Databases & Vector DBs: PostgreSQL, MySQL, MongoDB, Qdrant, Pinecone, Redis

APIs & Models: OpenAI API (GPT-4o, Whisper, TTS), OpenRouter, LLaMA 4, Gemma, MiniLM, Pyannote AI

Tools & Platforms: Docker, Git, AWS, Azure, GCP, FastMCP, AnythingLLM, APScheduler