

Shreyansh Verma

[Portfolio](#) | +91-7881110801 | [Gmail](#)

Summary: AI Engineer building production multi-agent systems, agentic pipelines, and LLM evaluation frameworks

WORK EXPERIENCE

Axtria

AI Engineer

Noida

Sep 2025 - Present

- Designed a production **multi-agent** DevOps platform (**LangGraph**, **Azure OpenAI**, **Flask**) that autonomously resolves **70%** of routine infrastructure requests at **~40%** faster turnaround, now used daily by the internal R&D team.
- Built the orchestration layer as a master agent handling LLM-based **intent classification**, parameter extraction, and multi-turn clarification, routing requests through a **state-machine graph** to five specialized domain worker agents.
- Standardized previously separate AWS and Azure Jenkins jobs into a single unified deployment **CI/CD pipeline**, adopted as the default across all environments, with **Terraform** and **Docker** driving cloud infrastructure provisioning.
- Reduced cluster infrastructure costs by **30%** while enabling demand-based dynamic autoscaling and improving workload scheduling efficiency, by migrating **8** Amazon **EKS** managed node groups to **Karpenter**-based NodePools.

NoBrokerHood

Product Analyst Intern

Bangalore

Aug 2024 – Nov 2024

- Cut data processing time from **50 minutes to ~7 minutes (~85% reduction)**, by building a **Python ETL pipeline** with optimized query design and parallelization that ingested **50+** column raw inputs from multiple product sales verticals.
- Designed the pipeline to clean and consolidate fragmented Google Form survey data into a **BigQuery** warehouse, keyed by unique **Kibana** identifiers, building a reliable single source of truth across **22,000+** tracked societies in India.
- Powered key business metrics on product adoption, occupancy, and customer lifecycle that enabled teams to actively drive a **10% churn reduction** and a measurable **4% lift** in app downloads across active flats over a 3-month period.

HERE Technologies

AI/ML Engineer Intern

Mumbai

Jun 2023 – Dec 2023

- Built an **NLP** preprocessing pipeline (**Regex**, **NLTK**) for feature selection and imputation across **75K+** daily map-task records, feeding a downstream binary classification and task-routing workflow that prioritized actionable map updates.
- Trained a binary text classifier (**GloVe** embeddings, chosen over **Word2Vec** after evaluation) to triage incoming map-correction tasks as actionable or escalate, reaching **0.84 accuracy** on noisy real-world HERE WeGo map data.

PROJECTS

InsightAgent | Python, LangGraph, Groq/Llama 3.3, LangSmith

[Link](#)

- Designed a multi-agent **LangGraph** pipeline answering natural-language questions over CSV/Excel/JSON, with parallel Stats and Viz agents via the **Send API** and executing LLM-generated pandas code in isolated **E2B** sandboxes.
- Built an **LLM-as-judge** eval framework pairing deterministic numeric checks with semantic scoring, proving **0.90 accuracy** on an unseen dataset (UCI Wine, 1,599 rows) and wiring it into a **CI** gate blocking any merge below **0.80**.
- Hardened for production with **Pydantic** structured outputs across **4** LLM nodes, a **LiteLLM** provider-fallback chain, and graceful partial-failure handling, deployed via **Docker** on Railway with detailed **LangSmith** node-level tracing.

Blood Report Analyser | Python, LangGraph, Groq/Llama 3.3, Qdrant

[Link](#)

- Built a **RAG**-powered agent answering natural-language questions over blood-report PDFs, using deterministic **pdfplumber** and regex extraction (no LLM in the pipeline) for fast, cheap, reliable parsing into a **Supabase** database.
- Engineered a **LangGraph ReAct** agent (Groq **Llama 3.3 70B**, **8B** fallback) routing across three tools: report search, chronological marker comparison, and cited retrieval from a curated **MedlinePlus** knowledge base within **Qdrant**.
- Embedded panel summaries and clinical docs into two **Qdrant** collections via sentence-transformers, with per-session memory through LangGraph **MemorySaver**, deployed on Streamlit Cloud with full **LangSmith** node-level tracing.

EDUCATION

Indian Institute of Technology (BHU), Varanasi

Integrated Dual Degree (B.Tech + M.Tech)

2020-2025

(8.26 CGPA)

SKILLS

- Languages & AI:** Python, SQL, LangChain, LangGraph, RAG, LLMs, HuggingFace, OpenAI, Gemini, LangSmith
- ML & Data:** PyTorch, TensorFlow, Scikit-learn, NumPy, Pandas, BigQuery
- Infrastructure & Tools:** Docker, Kubernetes, Terraform, AWS, Jenkins, CI/CD, FastAPI, Git, Cursor, Claude, Qdrant