Shashank Mankala

Buffalo, NY | Open to Relocation | shashankmankala.5@gmail.com | +1 (716) 547-1045 | github.com/Shashank-mankala1
linkedin.com/in/shashankmankala | shashank-mankala1.github.io/portfolio

Experience

Data Scientist Intern | New Era Cap | Buffalo, United States | Sep 2025 - Dec 2025

- Led a cross-functional team to design and deploy an end-to-end demand-forecasting pipeline in Snowflake with Azure integration, reducing manual planning time by 60% and improving forecast accuracy by 25%
- Orchestrated automation of vendor capacity and purchase planning by implementing allocation logic and retraining pipelines, eliminating manual Excel processes and enabling scalable, auditable forecasts for 40k+ SKUs
- Collaborated with supply-chain, sourcing, and engineering teams to design data schema, implement ETL, validate models, and deliver stakeholder-ready insights

Lead Data Analyst | Shadowfax | Bangalore, India | Sep 2023 - Jul 2024

- Achieved 98% accuracy in cleaning the employee database, a company-first in its 9-year history
- Optimized manpower allocation and delivery routes across East India using analysis and Looker dashboards, improving supply chain efficiency by 60%
- Drove 60% efficiency gains through actionable insights from in-depth hub performance analysis

Software Engineer Intern | Frugal Testing | Hyderabad, India | Jun 2022 - Jun 2023

- Created and implemented automation frameworks for four diverse projects, reducing manual testing and increasing efficiency by 30%
- Played a key role in developing the Central Bank Digital Currency (CBDC) product in collaboration with NPCI, contributing to its global release
- Pioneered test-methodology improvements through rigorous data analysis, informing cross-functional teams and reducing testing cycle times by 20% across four projects

Projects

High-Performance LLM Inference Engine with Grammar-Constrained Decoding | *LINK*

- Built a local LLM inference engine with full logits-to-decoding pipeline, supporting greedy and probabilistic decoding using temperature, top-k/top-p, and grammar-constrained text generation through token masking and incremental parsing
- Designed an OpenAI-compatible Chat Completions API with streaming, Jinja2-based prompt templates, and inference metrics (TTFT, tokens/sec), benchmarking Hugging Face and vLLM backends to evaluate latency and memory tradeoffs

TutorMind - GenAI-powered Personalized Tutor | *LINK*

- Developed a full-stack RAG-based GenAI application using LangChain, FAISS, and OpenAI to enable question answering from user-uploaded curriculum materials (PDFs, notes, videos)
- Integrated feedback-driven dual-model (factual + conceptual), and document-specific embedding logic, ensuring high answer accuracy with zero hallucination

Education

State University of New York at Buffalo — *Aug 2024 - Dec 2025*
Master's in Data Science — *Buffalo, United States*
Lovely Professional University — *Aug 2019 - May 2023*
Bachelor's in Computer Science and Engineering — *Punjab, India*

Skills

Languages: Python, SQL, R, C++, MATLAB, Javascript, HTML, CSS
Machine Learning and AI: Supervised Learning, Unsupervised Learning, Deep Learning, NLP, LLMs, RAG, Vector Databases, PyTorch, TensorFlow, LangChain
MLOps & Data Engineering: Docker, Kubernetes, Airflow, Spark, DagsHub, MLflow, CI/CD, Git, Data Pipelines, ETL/ELT, Kafka Streaming, Distributed Computing
Cloud Platforms: AWS, Google Cloud (GCP), Snowflake, Databricks
Data Visualization: Tableau, Power BI, Looker Studio
Database Systems: MySQL, PostgreSQL, MongoDB, BigQuery
Experimental Design: A/B testing, hypothesis testing, statistical inference