Shashank Mankala

New York, NY | Open to Relocate | shashankmankala.5@gmail.com| +1 (716) 547-1045 | GitHub | LinkedIn | Portfolio

Summary

AI Engineer / Applied Machine Learning Engineer with experience building production-grade LLM systems, inference engines, and end-to-end ML pipelines. Strong background in GenAI architectures, decoding strategies, RAG systems, and scalable ML infrastructure across cloud environments

Experience

Data Scientist Intern | New Era Cap | Buffalo, United States | September 2025 – December 2025

- Designed and deployed an end-to-end demand forecasting system with automated retraining pipelines, integrating Snowflake and Azure for scalable batch inference and data orchestration
- Executed model-driven allocation logic and capacity planning workflows, productionizing forecasting outputs for downstream systems across 40k+ SKUs
- Partnered with engineering teams to define data schemas, build ETL pipelines, validate model outputs, and ensure reproducibility, monitoring, and auditability of ML workflows

Lead Data Analyst | Shadowfax | Bangalore, India | September 2023 - July 2024

- Engineered data pipelines and analytical workflows to optimize workforce allocation and logistics routing across high-volume delivery networks
- Improved performance monitoring dashboards and analytical frameworks that informed operational decision-making at scale
- Collaborated with cross-functional stakeholders to translate operational requirements into data-backed system improvements

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

- Built automation frameworks and backend test systems across four large-scale projects, improving execution efficiency by 30 percent through reusable, modular design
- Contributed to the development of a Central Bank Digital Currency platform in collaboration with NPCI, supporting production release through system validation and reliability improvements
- Applied data-driven analysis to optimize testing pipelines, reducing execution cycles and improving system-level observability across distributed components

Projects

Production-Grade Video Summarization and Q&A Platform using Whisper, and Distributed Workers | LINK

- Designed and deployed a distributed, asynchronous video processing pipeline using FastAPI, Redis, and RQ workers to handle long-form YouTube videos with background ingestion, job tracking, and fault recovery
- Deployed the system on cloud infrastructure using Docker Compose, addressing real-world challenges such as platform bot detection, CPU-only ML constraints, and long-running background workloads

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

- Implemented 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
- Exposed the engine with OpenAI-compatible Chat Completions API with streaming, Jinja2-based prompt templates, and inference metrics (TTFT, tokens/sec), benchmarking Hugging Face and vLLM backends to study latency and memory tradeoffs

Education

State University of New York at Buffalo	August 2024 - December 2025
Master’s in Data Science (GPA: 3.9/4.0)	Buffalo, United States
Lovely Professional University	August 2019 - May 2023
Bachelor’s in Computer Science and Engineering	Punjab, India

Skills

- Languages: Python, SQL, C++, JavaScript
- Machine Learning and AI: LLMs, Transformer Architectures, Decoding Strategies, RAG Systems, Natural Language Processing, Embeddings, Vector Databases, Model Evaluation, PyTorch, TensorFlow, LangChain
- MLOps & Data Engineering: Docker, Kubernetes, MLflow, CI/CD, Airflow, Spark, Distributed Systems, Model Deployment, Inference Optimization, Version Control, Git, APIs, Scalable Automation
- Cloud Platforms: AWS, Google Cloud (GCP), Azure, Snowflake, Databricks
- Data Visualization & Storytelling: Tableau, Power BI, Looker Studio
- Database Systems: PostgreSQL, MySQL, MongoDB, BigQuery, FAISS, Redis