Rishi Guptha Mankala

Texas (Open to Relocate) • 934-263-3087 • rishiguptha.mankala@gmail.com • linkedin.com/in/rishi-guptha • github.com/rishiguptha

Software Engineer & Data Engineer with 2+ years building production AI systems. Specialized in LLM applications and scalable data-driven pipelines.

Technical Skills

Programming Languages: Python (Expert), SQL, Java, JavaScript/TypeScript, R, Git, Linux, Shell Scripting
Machine Learning & AI: PyTorch, scikit-learn, LangChain, RAG, OpenAI API, FAISS, Prompt Engineering, NLP, Transformers
Data Engineering & Processing: Apache Spark, Airflow, dbt, Snowflake, PostgreSQL, BigQuery, ETL/ELT, Pandas, NumPy
Cloud & Infrastructure: AWS (Lambda, S3, EC2, Glue, RDS, CodePipeline), GCP (BigQuery, Cloud Functions), Docker, CI/CD
Web Development & APIs: FastAPI, Next.js, Flask, Supabase, REST APIs, Node.js, React, SQLite

Experience

Nevara AI — New York, NY
Software Engineer (Intern) — *Jun 2025 - Aug 2025*

- Reduced manual sales review time 82% by architecting an AI-powered auditing MVP using Next.js, Supabase, and AWS Lambda, integrating OpenAI and Deepgram APIs to generate automated coaching reports for 100+ active users.
- Achieved 40% latency reduction by designing scalable FastAPI-based ML pipelines with PostgreSQL, optimizing serverless concurrency, async I/O, and batching strategies to reliably handle transcript ingestion and LLM analysis during traffic spikes.
- Improved LLM output reliability by 30% by implementing advanced prompt engineering frameworks (multi-call context weighting, schema validation, contradiction checks), significantly reducing downstream production errors.
- Secured data integrity and access control by designing schema-driven APIs using Supabase Auth (JWT), PostgreSQL, and Prisma ORM, implementing role-based authorization models to support secure access for 100+ users.
- Enabled continuous model improvement by building CI/CD pipelines with automated benchmarking, systematically comparing AI-generated outputs against human feedback to track regressions and drive measurable accuracy gains.

Kimberly-Clark — Bengaluru, India
Data Engineer (Intern) — *Jan 2023 - Jul 2023*

- Led migration of 10,000+ data assets into an enterprise data catalog, authoring Python and SQL validation scripts to replace legacy Excel trackers and ensure GDPR-compliant data governance and audit readiness.
- Saved 10+ hours per week by building cron-orchestrated metadata extraction pipelines in Python and Java, ingesting schemas from Data Lakes, Data Marts, and SAP HANA to streamline dataset discovery for analytics teams.
- Accelerated stakeholder reporting cycles by 35% by deploying Python and SQL lineage validation workflows, improving internal audit traceability and near real-time visibility into data dependencies across global teams.

Matchday AI — Hyderabad, India
Software Engineer (Contract) — *Jan 2022 - Dec 2022*

- Pioneered real-time tactical analysis for 6+ ISL teams by building computer-vision pipelines using OpenCV and homography algorithms to convert live broadcast footage into 2D tactical maps for coaching staff.
- Boosted training data labeling speed by 50% by developing a full- stack annotation platform using Flask, JavaScript, and SQLite, exposing RESTful APIs and integrating directly with Star Sports broadcast feeds.
- Reduced release latency by 38% by architecting CI/CD pipelines with AWS CodePipeline, Docker, and Git webhooks, standardizing automated testing and deployments for the analytics team.
- Cut post-match reporting time by 27% by implementing an active learning feedback loop to generate high-quality ground-truth datasets, enabling continuous model refinement and faster tactical insights.

Projects

Text-to-SQL Autonomous Agent | ***Python, LangChain, LoRA, RAG, Vector DB***

- Developed a self-correcting agentic system using RAG and LoRA to map natural language to complex SQL schemas, utilizing iterative reasoning loops to improve query accuracy by 40% vs. baseline LLMs.

Distributed Spatial-Temporal Clustering Engine | ***Python, Pandas, Scikit-Learn, DBSCAN, mpi4py***

- Processed 3.9M rows of NYC building permit data (1991-2025) using distributed DBSCAN and mpi4py parallel processing to analyze spatial-temporal patterns and identify construction hotspots at scale.

NutriSmart: Semantic Search Engine | ***Python, FastAPI, AWS EC2, Docker, FAISS***

- Launched a RAG service using Sentence Transformers and FAISS to index 1M+ tokens of medical papers, enabling sub-second semantic search queries via FastAPI REST APIs.

Snowflake Data Warehouse Modernization | ***dbt, Snowflake, S3, SQL***

- Modernized data infrastructure with an S3-to-Snowflake ELT pipeline on 20M+ records, using dbt incremental models and SCD2 logic to track historical changes while minimizing query costs.

NYC Taxi Insights | ***Python, GCP, Airflow, BigQuery***

- Orchestrated an Airflow workflow on GCP to process 10M+ records, designing star-schema data marts and integrating BERT feature embeddings for predictive modeling.

Education

Stony Brook University — Stony Brook, NY
M.S. Data Science — *Jan 2024 - Dec 2025*

Vellore Institute of Technology — India
B.Tech Computer Science — *Jun 2019 - May 2023*