SHRUTI KARMARKAR

AI Engineer | Machine Learning | LLM Systems

spk9869@nyu.edu • LinkedIn • GitHub • Portfolio

EDUCATION

Master of Computer Engineering: Data Science & AI Specialization *May 2026*
New York University, New York, NY | GPA: 3.7
Coursework: Machine Learning, Deep Learning, NLP, MLOps, Large Language Models & Applications, Reinforcement Learning, Bayesian Methods, Big Data Analytics, Cloud Computing, Statistical Inference

B.Tech in Computer Science *June 2023*
Rajiv Gandhi College of Engineering, India | GPA: 3.9
Coursework: Artificial Intelligence, Neural Networks, Computer Vision, Data Structures & Algorithms, Software Engineering, Database Management, Probability & Statistics, Linear Algebra, Compiler Design

WORK EXPERIENCE

AI / Data Science Intern | Xometry | New York, NY *June 2025 – August 2025*

- Saved **$180K** by designing and executing **A/B experiments** across **~50K CNC and sheet metal jobs** to validate ML-driven pricing changes, confirming improved acceptance rates with zero revenue loss findings presented directly to and adopted by leadership.
- Identified partner capacity constraints as the primary driver of acceptance rate decline by building a **multi-variable root cause analysis pipeline** on **50K+ records** using **Python and Snowflake SQL**, directly reshaping partner engagement strategy firm wide.
- Trained a **Random Forest classifier** and **regression model** on **~24K job records** with geographic feature engineering, uncovering **net shipping losses** across partner locations findings triggered an immediate strategic review of **localized incentive structures** preventing further investment loss.
- Prototyped an **LLM-assisted reporting tool** using the **Claude API** to translate **Snowflake SQL outputs** into plain-English executive briefs, cutting reporting time from hours to under **10 minutes** with full data privacy compliance.

Research Data Scientist | Gilbert Research Center | Tamil Nadu, India *April 2023 – Feb 2024*

- Deployed a **brain tumor detection model** as a **REST API** using **FastAPI**, serving CNN and VGG16 predictions on MRI data with **94% test accuracy** and **0.93 F1 score** optimized for edge inference via **TensorFlow Lite** with model quantization reducing model size for deployment.
- Automated the end-to-end **MRI data preprocessing and inference pipeline** using **OpenCV and Python**, processing **10K+ MRI scans** and reducing per-scan prediction time from minutes to under **5 seconds** by eliminating manual notebook execution.

AI/ ML PROJECTS

SaaS Churn Copilot: Churn Intelligence Platform | Python · XGBoost · LangChain · LangGraph · FAISS · PostgreSQL · Docker

- Improved **30-day churn prediction** accuracy on **300K+ account records** by building a temporally correct **XGBoost pipeline with Platt calibration**, resolving right-censoring and data leakage that inflated baseline accuracy and producing honest probabilities validated against true population churn rate.
- Extended the platform with a **hybrid RAG and SQL router** using **LangChain and LangGraph**, embedding **daily activity records** into a **FAISS vector index** with sentence-transformers and implementing **LangSmith tracing** to monitor retrieval quality and response latency across all query types.
- Hardened the production pipeline by exposing churn scores via a **FastAPI endpoint** with **input validation, rate limiting, and Pydantic response schemas**, replacing the unsafe replace DB load with an **atomic staging swap** protected by a **10% row count validation threshold.**

ClaimIQ; Multi-Agent Insurance Adjudication System | Next.js 14 · TypeScript · Anthropic Claude API · LangSmith · Vercel

- Reduced claim processing time to near-zero perceived latency by engineering a **4-agent sequential LLM pipeline** with **real-time token streaming**, retry logic with exponential backoff, and parallelized Agent 4 execution cutting final document generation time by **~50%**.
- Improved verdict reliability across **4 automated fraud checks** by enforcing **structured JSON output** between agents and routing PDF ingestion through **claude-haiku-4-5** as a lightweight preprocessor before the main **claude-sonnet-4-6** adjudication pipeline, reducing preprocessing cost by **60%**.
- Eliminated blind-spot regressions by implementing **LangSmith tracing** on every agent call capturing latency and token usage and building a **golden eval dataset** with an automated **verdict regression suite** that exits non-zero on any verdict mismatch.

TECHNICAL SKILLS

- **Languages:** Python (advanced), SQL, JavaScript, Spark SQL, R (basic)
- **AI and LLMs:** Anthropic Claude API, LangChain, LangGraph, LangSmith, RAG Pipelines, Prompt Engineering, Prompt Versioning, Structured Outputs, Pydantic, Tool Calling and Function Calling, Agentic Workflows, Multi-Agent Systems, Embeddings, FAISS, Chroma, pgvector, Semantic Search, Retrieval Evaluation, Hugging Face Transformers, Fine-tuning
- **ML and DL:** PyTorch, TensorFlow, scikit-learn, XGBoost, CNNs, RNNs, LSTMs, SHAP Explainability, A/B Testing, Feature Engineering, Causal Inference, NLP, Sentiment Analysis, Model Calibration, MLflow
- **Data Engineering:** ETL Pipelines, Apache Spark, PySpark, Snowflake, PostgreSQL, Docker, Airflow, dbt, CI/CD, GitHub Actions, Data Quality Validation, Query Optimization
- **Backend and Tools:** FastAPI, REST APIs, Streaming APIs, Pytest, Git, Azure, GCP, Linux/Unix, Jupyter, Tableau