Mohammed Pathariya

linkedin.com/in/mjpathariya | mjpathariya.com | github.com/MohammedPathariya
mjpathariya7@gmail.com | +1 (930) 333-7212 | Indiana, USA

Education

Master of Science in Data Science, Indiana University — Bloomington, IN, USA
GPA: 3.7/4.0 | Coursework: LLMs, Deep Learning, MLOps, Cloud Computing — *Aug. 2024 – Exp. May 2026*

BE in Artificial Intelligence & Data Science, Pune University — Pune, MH, India
GPA: 3.8/4.0 | ML, Data Structures & Algorithms, OS, Statistics — *Aug. 2020 – May 2024*

Experience

Graduate Research Assistant, AI in Sports Analytics — Aug. 2025 – Present
Indiana University, Kelley School of Business — *Bloomington, IN*

- Constructed a proprietary multimodal corpus of **450+ NBA press conferences**, utilizing **WhisperX** on **Slurm-managed HPC nodes** to accelerate data processing by **30x**, enabling large-scale sentiment analysis.
- Designed a hybrid feature extraction strategy combining **Sentence-BERT** embeddings with **VADER** lexical rules, capturing **granular emotional shifts** often lost by high-dimensional vector smoothing.
- Implemented automated statistical checks to detect **distribution drift** in sentiment features, ensuring longitudinal consistency across **50+ weekly games** for downstream **causal inference modeling**.

Data Engineering Intern — Jan. 2024 – Jun. 2024
Sparkwood IT Solutions — *Pune, India*

- Conducted root-cause analysis on a **5% KPI discrepancy** between Marketing and Finance reports, standardizing the "Net Sales" metric definition via a validated **SQL View** to ensure organizational alignment.
- Optimized legacy **PostgreSQL** reporting queries by analyzing execution plans and adding indices, reducing dashboard latency by **40%** (5m → 3m) to accelerate time-to-insight.
- Maintained **Airflow** ETL pipelines for the central data warehouse, implementing automated data quality checks and retries to ensure reliable daily reporting for business stakeholders.

Projects

AudioGroove | *Deep Learning, Research* | Code — Oct. 2023 – Apr. 2024

- Published **Tunes by Technology** in **IEEE ICC Robins**, statistically validating that **Bi-Directional LSTMs** outperform DCGANs by **85%** in structural coherence.
- Designed a distributed feature extraction pipeline using **Dask** to process **175,000+ MIDI files**, reducing data preparation latency by **90%** (20h → 2h) to enable rapid hypothesis testing.
- Conducted a rigorous hyperparameter study tracking **50+ experiments** via **MLflow**, identifying optimal sequence length (64) and hidden dimensions (256) to minimize validation loss to **0.78**.

LearnLoop | *GenAI Evaluation, Experimental Design* | Code — Jan. 2025 – Apr. 2025

- Implemented a **Session-Based RAG** workflow using ephemeral **FAISS** indexes, validating that strict data isolation reduces retrieval latency to <**50ms** compared to centralized baselines.
- Designed a **synthetic load simulation** with **500+ concurrent agents** to stress-test system throughput, utilizing statistical performance metrics to identify and resolve concurrency bottlenecks.
- Engineered a **self-correcting validation loop** using **Pydantic** that quantified and reduced LLM schema hallucination rates by **90%** (10% → <1%) for structured data extraction tasks.

The Digital Forge | *Multi-Agent Systems, Evaluation* | Code — May 2025 – Aug. 2025

- Designed a comparative benchmarking study for LLM code generation, demonstrating that multi-agent orchestration outperforms zero-shot baselines by **35%** (50% → 85%) on algorithmic tasks.
- Quantified the impact of autonomous "Reflection" loops, validating that iterative runtime error feedback resolves **70%** of syntax and logic defects compared to static generation.
- Established a **functional correctness benchmark** (Pass@1) by building a **Dockerized test harness**, filtering out **40%** of solutions that were syntactically valid but logically incorrect.

Technical Skills

Languages & Visualization: Python, SQL, Bash, Tableau, Streamlit, Matplotlib, Plotly, Pydantic
Generative AI & ML: PyTorch, Hugging Face, FAISS (RAG), XGBoost, Scikit-learn, Causal Inference, WhisperX
Big Data & Statistics: Dask, Pandas, NumPy, A/B Testing, Hypothesis Testing, Statsmodels, Experimental Design
Infrastructure & MLOps: Docker, MLflow, Apache Airflow, PostgreSQL, GitHub Actions, Locust, GCP (Vertex AI)
Strategy & Concepts: KPI Definition, Root Cause Analysis, Multi-Agent Systems, System Design, TDD