Rishika Mamidibathula

(332) 251-6686 · rm4318@columbia.edu · linkedin.com/in/rishika-mamidibathula · github.com/rishika1099 · rishika-m.netlify.app

Education

Columbia University New York, NY

M.S. Data Science, GPA: 3.873/4.0 *Aug 2025 – Expected Dec 2026*

- Coursework: Deep Learning, Generative AI Systems, Agentic AI, High Performance Computing, Causal Inference, Statistical Inference
- Teaching Assistant for Artificial Intelligence for Public Policy; Data Science Institute Student Council Member

Vellore Institute of Technology Vellore, IN

B.Tech. Computer Science and Engineering (Data Science specialisation), GPA: 4.0/4.0 *Jul 2019 – May 2023*

- Ranked 7th out of 200 (Top 4%); Merit Scholarship recipient (2019–2023); Program Representative (2019–2023)
- Coursework: Machine Learning, Artificial Intelligence, Natural Language Processing, Image Processing, Social Information Networks

Skills

Programming & Frameworks: Python, SQL, R, C++, PyTorch, TensorFlow, scikit-learn, FastAPI, NumPy, Pandas

LLM & AI Systems: RAG, LangChain, Hugging Face, ChromaDB, FAISS, Multimodal AI, Prompt Engineering, Agent Workflows

Data Engineering & Infrastructure: Databricks, PySpark, MongoDB, MySQL, BigQuery, Redis, Docker, Git, REST APIs

Cloud & MLOps: AWS, Azure DevOps, Weights & Biases, CI/CD, Experiment Tracking, Feature Engineering, Model Monitoring

Work Experience

Data Science Intern Jun 2026 – Aug 2026

NYC Administration for Children's Services (ACS) *New York, NY*

- Developing predictive risk models on child welfare administrative data with explainable machine learning, fairness auditing, and causal adjustment to support transparent decision-making in high-stakes public-sector settings.

Software Engineer Aug 2023 – Jul 2025

Shell *Bengaluru, IN*

- Developed machine learning forecasting solutions in Databricks (PySpark, SQL) across 12 business units, reducing forecast error by 23% and enabling $100K+ annual cost optimization decisions.
- Built production data pipelines, feature engineering workflows, and 5 Blue Prism RPA bots with logging and retry logic, cutting manual reporting effort by 85% (120+ hours/quarter) and improving SLA compliance from 92% to 99%.

Technical Analyst Intern Jan 2023 – Jul 2023

Novartis *Hyderabad, IN*

- Engineered a NLP-driven clinical trial analysis workflow for drug-level sentiment mining, automated summarization, and outcome extraction from unstructured research documents, reducing manual literature review and evidence synthesis effort by 40% per quarter.
- Built predictive and time-series forecasting pipelines on environmental operations data, integrating web-scraped external signals, feature engineering, and temporal trend analysis to support sustainability initiatives targeting a 19% annual reduction in carbon emissions.

Research Experience

Research Assistant – Clinical LLM & Phenotyping Jan 2026 – Present

Columbia University, Irving Medical Center *New York, NY*

- Built an information extraction system featuring hybrid regex/LLM de-identification, longitudinal EHR reconstruction, token-aware chunking, and hallucination-aware validation to transform clinical notes into 56 structured cardiac sarcoidosis phenotype variables.
- Implemented phenotype harmonization and temporal feature synthesis across multi-specialty EHR data, generating analysis-ready cohorts for clinical outcome modeling and translational research.

Research Assistant – Human Rights LLM Evaluation Jan 2026 – Present

Columbia University, Graduate School of Arts and Sciences *New York, NY*

- Developed a retrieval-augmented LLM evaluation framework integrating automated web search, evidence synthesis, and chain-of-thought reasoning to generate explainable human-rights due diligence scores across 27 defense manufacturers, reducing manual review effort by 80%.
- Quantitatively evaluated LLM-human agreement using weighted kappa, Krippendorff's alpha, rank correlation, MAE, and confusion-matrix analysis, establishing reliability benchmarks for automated policy-risk assessment.

Projects

Folio: Clinical Multimodal RAG https://github.com/rishika1099/Folio-Clinical-Multimodal-RAG | https://folio-health.vercel.app

- Built a multimodal clinical intelligence platform unifying retrieval-augmented generation, document understanding, speech transcription, vector retrieval, and longitudinal patient record management across text, PDF, image, and audio inputs.
- Engineered a consensus extraction pipeline using embedding-cluster voting across multiple LLMs, achieving 85.1% extraction micro-F1, 100% RAG recall@1, 100% grounded-chat correctness, 100% PII recall, and sub-2s median inference latency.

KV Cache Optimization for LLM Inference https://github.com/rishika1099/KV-Cache-Implementation

- Developed and benchmarked KV-cache optimization techniques for Llama-2-7B, including KIVI quantization, TopK sparse selection, SnapKV eviction, and MLA latent compression, using Triton kernels, distributed experimentation, and LongBench evaluation.
- Achieved 4× KV-cache compression with KIVI 4-bit while maintaining LongBench quality parity (0.292 vs. 0.291 baseline), improving decode throughput by 1.93× and increasing maximum batch capacity from 32 to 128, resulting in a 3.1× peak throughput improvement.

Colon Cancer Trial Causal Analysis https://github.com/rishika1099/Colon-Cancer-Trial-Causal-Analysis

- Applied causal inference methods including ATE estimation, heterogeneous treatment-effect modeling, mediation analysis, and transportability assessment to a randomized colorectal cancer trial involving 929 patients.
- Estimated a treatment hazard ratio of 0.69 and quantified the effects of post-treatment conditioning, demonstrating collider bias that reversed the direction of the estimated treatment effect (HR 0.69 → 1.10) and underscored the importance of proper causal adjustment strategies in clinical studies.