

# Diksha Menghmalani

[dmenghmalani@gmail.com](mailto:dmenghmalani@gmail.com) | [linkedin-diksha-menghmalani](https://www.linkedin.com/in/diksha-menghmalani) | [github-dikshamenghmalani](https://github.com/dikshamenghmalani) | [portfolio-diksha](https://portfolio-diksha.github.io)

## EDUCATION

### University of Wisconsin–Madison

Sep 2024 – May 2026

*MS in Data Engineering (Computer Science); GPA: 3.82/4*

*Madison, WI*

- **Courses:** Distributed Systems, Big Data Systems, Machine Learning, Foundation Models, Deep Learning & Generative AI, Topics in Database Management Systems, Data Exploration

### Cummins College of Engineering for Women

Aug 2017 – Aug 2021

*B.Tech in Electronics and Telecommunication; GPA: 8.6/10*

*Pune, India*

## SKILLS

**Programming:** Python, Go, SQL, C++, Bash, TypeScript, JavaScript

**Software & Infrastructure:** Git, Docker, Kubernetes, CI/CD, Linux, AWS, GCP

**Machine Learning:** PyTorch, TensorFlow, LangChain, LangGraph, scikit-learn, Weights & Biases, MLflow, Jax

**Data Systems:** Spark, Kafka, Hadoop, Redshift, Azure Data Explorer

**Frontend:** React, WebSockets

## EXPERIENCE

### AI Software Engineer Intern

May 2025 – Aug 2025

*Zscaler*

*San Jose, CA*

- Reduced P95 connection latency by **10x** and eliminated **1,000+ support tickets** by resolving large-scale misrouting across **1M+ users**.
- Built a 0-to-1 graph-structured routing system that models device-IP interaction patterns and latency minima to infer optimal data center routing, replacing external geolocation data at production scale.
- Developed a supervised routing model and feedback pipeline to automate **40K+** daily IP corrections, stabilizing routing using real-world latency signals.

### Machine Learning Engineer

Jul 2021 – Aug 2024

*ZS Associates*

*Pune, India*

- Built customer-facing ML pipelines generating call plans for **1M+** physicians across **300** field reps, driving **7%** YoY NBRx growth for 2 pharma brands.
- Developed Random Forest and XGBoost ranking models using rolling-window training to prevent look-ahead leakage; optimized via AUPRC.
- Engineered 200+ features and improved models via 3-month territory A/B tests and field-rep feedback loops over a 2-year lifecycle.
- Optimized pipelines with PySpark, reducing runtime by **71% (24 to 7 hours)** and automating delivery to downstream territory optimization systems.
- Partnered directly with non-technical stakeholders to scope requirements, translate clinical domain insights into engineering features, and validate model outputs.

## PROJECTS

### Distributed F1 Stream Processing Engine ([Github](#))

Jul 2026 – Present

- Building a real-time Kafka pipeline ingesting subsecond OpenF1 telemetry, partitioning topics by driver ID for ordered stream processing.
- Designing anomaly detection using Redis geospatial indexing to flag off-track events, streaming live alerts to a React UI via WebSockets.

### RAG-Based Research Paper Explainer Chatbot ([Github](#))

Apr 2026 – Jul 2026

- Built an end-to-end RAG pipeline over ML arXiv papers using Qdrant and section-aware chunking, improving retrieval relevance by **25%** over fixed-size chunking baselines.
- Developed a grounded QA system using hybrid retrieval (BM25 + dense embeddings) and cross-encoder reranking, achieving **85%** accuracy on a 50-question evaluation set while significantly reducing hallucinated responses.

### LLM Pretraining Data Deduplication & Optimization ([Github](#))

Feb 2025 – May 2025

- Engineered a data deduplication for **10M+** documents to reduce memory overhead and compute costs, cutting dataset size by **14%** and fine-tuning runtime by **15%** using MinHash and Bloom filters.
- Built a distributed multi-GPU training pipeline using Hugging Face Accelerate, enabling scalable parallel training and automated experiment tracking with Weights & Biases.