

PRIYANSHU RAWAT

+1 (585)-230-2026 | prawat3@ur.rochester.edu | linkedin.com/in/prwt | github.com/PRawat00

WORK EXPERIENCE

Center for Integrated Research Computing, UofR

New York

Data Scientist

Feb 2025–present

- [Transformers, NLP] Fine-tuned multi-task BERT classifier in PyTorch for automated ticket categorization across 4 teams and 4 priority levels on 200K+ tickets, achieving 85% precision and 83% recall, deployed as authenticated REST API in Docker.
- [Large Language Models, Data Privacy] Deployed API for automated ticket summaries and resolution suggestions using LLMs, with retrieval from compliance-ready knowledge base built through sensitive data anonymization of historical tickets.
- [Vector Search, Agentic AI] Built multi-source RAG system (SQL databases, documentation, real-time cluster metrics) using LangGraph orchestration, FastAPI, and Docker, serving 1000+ users for HPC job scheduling, resource optimization, and troubleshooting.
- [MLOps, DevOps] Deployed automated CI/CD pipelines using GitLab CI/CD with Prometheus and Grafana monitoring for ML classification models and RAG applications, enabling continuous deployment with automated testing and real-time performance tracking.

FLX AI

New York

Data Science Intern

Jun 2025–Jul 2025

- [Workflow Automation, Financial AI] Automated fund manager due diligence with LangGraph agent on 112+ SEC filing questions (10-K, 10-Q), achieving 70-75% precision via multi-step validation and cutting analysis time from days to minutes.
- [OCR, Document Processing] Engineered an OCR pipeline (90% accuracy on financial tables) using Docling that implemented section-based chunking while preserving and linking citation metadata to enable precise source attribution.
- [RAG Systems, Vector Search] Engineered a RAG pipeline (FastAPI/pgvector) with a ColBERT v2 re-ranker, cutting latency by 75% through dynamic batching for concurrent question processing.

K-Labs: Continual Learning Lab, UofR

New York

ML Research Assistant

Sep 2024–Jan 2025

- [Computer Vision, Domain Generalization] Conducted a comparative study on domain generalization by engineering and modifying ViT and CNN models for object detection across real-world photos, sketches, and illustrations.
- [Model Optimization, Distributed Training] Reduced model training time by approximately 50% through Distributed Data Parallel (DDP) and Automatic Mixed Precision (AMP) implementation for multi-GPU training in PyTorch.
- [Deep Learning Research, MLOps] Tracked and visualized model activations using Weights & Biases, comparing regularization effects on shape-focused vs texture-focused learning behaviors and presenting comparative results to the research team.

EDUCATION

MS, Data Science

Aug 2024–Dec 2025

University of Rochester, New York

GPA: 3.83/4.00

B.S. Computer Science

Jul 2020–Jul 2024

Graphic Era Hill University

GPA: 3.4/4

TECHNICAL SKILLS

- GenAI & LLMs:** RAG Systems, Agentic AI, LangGraph, LangChain, Prompt Engineering, Multi-modal AI
- ML Frameworks:** PyTorch, TensorFlow, Transformers (Hugging Face), XGBoost, MLflow
- Model Optimization & Fine-tuning:** LoRA/QLoRA, PEFT, Quantization, vLLM, DDP, AMP, Computer Vision (ViT, CNN)
- Vector & Retrieval:** Vector Search, ChromaDB, Pinecone, pgvector, Elasticsearch, Semantic Search
- Databases & SQL:** PostgreSQL, MongoDB, SQL (Complex Queries, Window Functions, CTEs)
- Big Data & Streaming:** Apache Spark, Spark Streaming, Kafka, Airflow, Databricks, Delta Lake, ETL Pipelines
- Cloud & DevOps:** Docker, Kubernetes, AWS, FastAPI, Streamlit, CI/CD, MLOps, REST APIs
- Languages & Tools:** Python, R, SQL, Git, OCR (Docling), spaCy, LLM APIs (OpenAI, Anthropic)

PROJECTS

CyberIntel Summarizer: Real-Time Threat Intelligence System

Jan 2025–Feb 2025

- [LLM Engineering, Security Intelligence] Engineered cybersecurity intelligence pipeline ingesting NVD, CISA, and MITRE ATT&CK feeds, generating LLM-powered summaries and severity classifications for 100+ daily CVE updates with automated monitoring.
- [LLM Fine-tuning, Inference Optimization] Deployed LoRA-fine-tuned model with 4-bit quantization using vLLM, achieving 3x throughput improvement and 60% memory reduction vs. FP16 baseline through dynamic batching and paged attention.
- [Full-Stack Development, Deployment] Built production-ready API (FastAPI, PostgreSQL) with Streamlit dashboard displaying real-time threat analytics by severity, attack vector, and vendor with interactive performance benchmarking.

Capstone: Gluten Sensitivity Prediction System (in collaboration with *Wegmans Food Market*)

Aug 2025–present

- [Machine Learning, Feature Engineering] Built **XGBoost classification models** on 5.6M transaction records to predict gluten sensitivity with feature engineering (purchase cadence, brand preferences, substitution patterns) and handling **severe class imbalance** via undersampling and cost-weighted learning.
- [Model Optimization, Evaluation Metrics] Designed a **threshold optimization framework** across 4 strategies (F1-optimal, Youden's J, cost-weighted FN:FP, PR-AUC), improving precision by **49%** while maintaining sensitivity for imbalanced classification.
- [Business Analytics, Cost Analysis] Conducted **cost-sensitive analysis** to quantify business trade-offs between false positives and false negatives, demonstrating reduced coupon waste and higher marketing ROI through optimized targeting.