UTKARSH SINGH

usingh7@jh.edu | +1 (410) 805-4462 | LinkedIn | ORCID | GitHub

EDUCATION

Johns Hopkins University, MSE Applied Mathematics — August 2025 – Expected December 2026
Honors: Dean's Master's Fellowship | GPA: 3.30/4.00
Relevant Coursework: Computing for Applied Mathematics, Data Science, Econometrics, Time Series Analysis

Maharashtra Institute of Technology, BTech in Electronics and Communication Engineering | GPA: 3.99/4.00 — July 2019 – July 2023
Relevant Coursework: Machine Learning, Natural Language Processing, Optimization Techniques, Artificial Neural Networks, Pattern Recognition

TECHNICAL SKILLS

- **Programming and Data:** Python, SQL, C++, R, MATLAB, Pandas, NumPy, PostgreSQL, MySQL
- **Machine Learning Libraries:** Scikit-learn, PyTorch, TensorFlow, Keras, LightGBM, XGBoost
- **Deep Learning and LLM frameworks:** Transformers, Hugging Face, LangChain, LangGraph, PydanticAI, Ollama, Faiss, ChromaDB
- **ML Systems and MLOps:** Git, Docker, Kubernetes, OpenShift AI, MLflow, Grafana, Prometheus, Jupyter
- **Platforms and Big Data:** Google Cloud, IBM Cloud, IBM WatsonX (AI/Data), Apache Spark, Apache Hive
- **Data Processing and Visualization:** PyPDF, Docling, Streamlit, Matplotlib, Seaborn
- **Certifications:** IBM Machine Learning Specialist, IBM Developer Profession, Redbooks Gold Author

EXPERIENCE

Research Assistant | Johns Hopkins Bloomberg School of Public Health — August 2025 – December 2025
- Designed class-imbalance mitigation strategy (cost-sensitive learning) for **99:1**–skewed EHR data; enhanced rare-event recall while controlling false positives.
- Fitted tree-based classifiers for longitudinal prediction tasks, optimizing predictive performance and calibration under data sparsity and label imbalance.
- Iteratively refined model stability and generalization by incorporating additional data sources and targeted feature and hyperparameter refinements.

Machine Learning Engineer | IBM Systems Development Lab — July 2023 – July 2025
- Deployed autoencoder-based anomaly detection models (TransformerAE, LSTM) on multivariate time-series data across **32** KPI groups and **150+** features.
- Built end-to-end data pipelines for telemetry ingestion, feature engineering, and anomaly scoring, persisting scored windows for downstream analysis.
- Calibrated detection thresholds using sequence-aware statistical tuning methods to balance recall and false positives under production constraints.
- Architected a real-time semantic triage engine using **BERT** and **Faiss indexing** to correlate live system anomalies with time-aligned logs and past tickets.
- Applied clustering and re-ranking techniques to compress high-dimensional incident windows and surface top resolution candidates within **sub-60s** latency.
- Created an instruction-tuned **LLM assistant** for Storage Insights using Granite/Llama with **QLoRA**-based alignment and light **RLHF**-style preference shaping.
- Implemented structured query translation from natural-language inputs to monitoring and observability APIs, improving reliability of automated diagnostics.
- Engineered an **open-source MCP** service exposing observability signals to agent-based workflows, reducing end-to-end issue identification latency by **15%**.

Software Engineer Intern | IBM Systems Development Lab — January 2023 – July 2023
- Optimized large-scale ingestion pipelines for multi-tenant storage telemetry, processing **2+ TB/day** of Protobuf-based configuration data on IBM Cloud.
- Reduced small-file overhead by **40%** through **dynamic batching** and **size-aware flush logic**, improving overall pipeline efficiency and space utilization.
- Launched scalable event-driven consumers on Kubernetes to stream, filter, and normalize millions of system metrics per hour for downstream ML workloads.
- Accelerated analytical queries by **2–3×** by optimizing schema layouts and ingestion paths for distributed SQL and Spark-based engines.
- Orchestrated Prometheus metric exporters across **10** ML pipelines and automated Grafana dashboard provisioning to enable monitoring of performance.
- Established standardized observability and performance baselines to detect bottlenecks and reliability issues across production data pipelines.

Data Scientist Intern | Sisai Technologies — November 2021 – February 2022
- Developed temporal convolution–based sequence models to process noisy, sparse IoT time-series data, improving feature extraction for forecasting tasks.
- Devised simulation-based stress tests spanning **100+** thermal-failure scenarios to evaluate model behavior under edge conditions and distribution shifts.

PROJECTS

Expert Specialization in Multilingual MoE Transformers | Johns Hopkins — October 2025 – December 2025
- Constructed a sparse **Mixture-of-Experts** (MoE) Transformer for multilingual language modeling across 4 languages, implementing top-k routing and load-balancing loss in PyTorch to scale model capacity with **3x** higher parameter count at constant inference FLOPs.
- Analyzed token-level expert routing behavior over **1M+** multilingual samples, quantifying expert specialization using routing entropy, utilization imbalance, and KL divergence, and identified language and script-specific expert preferences that reduced routing entropy by **18–25%** compared to random assignment.

LLM Sensitivity to Politeness and Emphasis | Johns Hopkins — September 2025 – November 2025
- Evaluated Llama 3, GPT-5, and Claude Sonnet 4.5 on the **GSM8K** test set across **multiple linguistic variants** (baseline, polite, emphatic, bold) to measure how phrasing affects math-reasoning reliability; auto-graded outputs with GSM8K gold labels and logged full token usage.
- Quantified the accuracy–compute tradeoff, observing under **1.5%** variance in accuracy but a **12–18%** increase in token usage for polite/emphatic prompts; validated results using paired statistical tests (McNemar, t-test, effect sizes), showing negligible accuracy change yet significant compute overhead.

Workload Placement Advisor | IBM — March 2025 – May 2025
- Formulated and trained a time-series forecasting engine for block-level I/O demand, benchmarking classical statistical models against **PatchTST** and **TTM** across **100+ FlashSystem arrays**, increasing out-of-sample prediction accuracy by **25%**.
- Operationalized the decision optimization layer that translated probabilistic forecasts into placement and migration actions, explicitly modeling capacity, latency, and risk tradeoffs through hybrid compatibility scores (model predictions + configuration rules) to prevent nearly **10,000** potential **SLA violations**.

Ransomware Threat Detection for FlashSystem | IBM — November 2024 – March 2025
- Trained **SnapML-based ensemble classifiers** on **500k+** FlashCore Module traces, detecting ransomware with **< 1% false-positive rate** on live workloads.
- Automated pipelines for class re-balancing, labeling, and **cross-family validation** to ensure the system generalized well to unseen ransomware variants.
- Integrated the inference layer into Storage Virtualize stack to trigger early alerts, immutable snapshots, and forensic trace dashboards for support teams.

Ticket Deflection | IBM — April 2024 – September 2024
- Developed a Gen-AI deflection workflow that sanitized **GDPR**-sensitive fields and clustered **26,000+** historical tickets into representative knowledge groups.
- Generated auto-validated FAQs from these groups, reducing Level 2 support workload by nearly **30%** and cutting MTTR by **20%** by deflecting repetitive issues.