Rishi More

USA | rmore2@jhu.edu | +1 (650) 220-9450 | linkedin.com/in/rishimore102 | github.com/rishi-more-2003 | rishi-more-2003.github.io

Summary

ML Engineer with research and production experience in **LLM inference optimization, RAG systems, and preference learning.** Co-first author of an **ICML 2026** paper on risk-controlled test-time compute for reasoning LLMs. Seeking roles at the intersection of applied ML research and large-scale systems.

Education

Johns Hopkins University **Expected May 2026**
M.S.E. in Computer Science *GPA: 4.0/4.0*
- **Coursework:** Deep Learning, Human-in-the-Loop ML, Theory of Replicable ML, Non-Stationary Environments

University of Mumbai **July 2024**
B.Tech. in Computer Science, Honors in AI/ML *CGPA: 9.85/10*
- **Coursework:** Natural Language Processing, Big Data Analytics, Data Warehousing, Quantitative Analysis

Professional Experience

JHU Data Science and AI Institute (DSAI) **May 2025 – Present**
Graduate Research Assistant (Advisors: Prof. Daniel Khashabi, Prof. Eric Nalisnick) *Baltimore, MD*
- **Co-first author** on *Compute When Worth It: Risk Control for Reasoning on a Compute Budget*, accepted at **ICML 2026**; earlier version accepted at the **NeurIPS 2025 Workshop on Efficient Reasoning**.
- Ran large-scale inference experiments across **4 models**, **4 benchmarks**, and **2.7K+** samples; improved accuracy by **8%** while operating under **50–70%** compute budgets.
- Implemented threshold search, signal ensembling, and validation/test split automation to make calibration results reproducible across models and benchmarks.

JHU Center for Language and Speech Processing (CLSP) **January 2025 – December 2025**
Graduate Researcher (Advisor: Prof. Jason Eisner) *Baltimore, MD*
- Built a temporal-aware RAG pipeline over **10K+** non-stationary documents using time-decay retrieval weighting and metadata-aware indexing.
- Improved retrieval precision by **23%** over dense-retrieval baselines by prioritizing temporally relevant evidence in high-volatility knowledge domains.
- Developed a reference-free consistency scoring framework for hallucination detection, reducing false retrievals by **31%**.

Mastek **November 2022 – July 2023**
Machine Learning Engineer Intern *Mumbai, India*
- Engineered a low-latency voice authentication service **(<100ms p99 latency, 96.6% accuracy)** deployed in production; placed 2nd of 2,400 teams company-wide.
- Engineered a TensorFlow audio-processing pipeline for **1M+** voice samples, including spectral-gating denoising and reliability evaluation under noisy conditions.
- Improved authentication reliability by **20%** in high-noise environments through preprocessing, feature extraction, and model-evaluation updates.

Projects

Asymmetric Relation Dynamics in Language Agents | *Python, LLM Agents, LoRA, AsyncIO, BM25* | GitHub
- Built an async multi-agent training system with Qwen3 caregiver and LoRA-adapted child agents, salience-gated BM25 memory retrieval, and **4** controlled ablation settings.
- Reduced dialogue turns by ∼**25%** versus Solo/Peer baselines, with ablations isolating scaffolding dependency as the transfer bottleneck.
- Implemented as a reproducible multi-stage pipeline with centralized config, JSONL metrics logging, and transcript storage; reproducible across 12 ablations via a single CLI command

Label-Preserving Domain Adaptation via KL-Regularized Optimal Transport | *Python, Sinkhorn OT, t-SNE* | GitHub
- Extended Sinkhorn OT with pairwise KL-divergence and intra-cluster consistency penalties to preserve label structure during unsupervised domain adaptation without target labels.
- Outperformed standard Sinkhorn by **12.6%** (Two-Moons, **99.8%**) and **5.3%** (MNIST→USPS, **82.7%**) under 1-NN evaluation across unsupervised and semi-supervised settings.

Moderator-Aligned Toxicity Detection via Direct Preference Optimization | *Python, TensorFlow, DPO* | GitHub
- Fine-tuned a BiLSTM toxicity classifier on ∼**15K** Reddit comments using DPO with pairwise preference signals derived from community upvote-ratio gaps.
- Achieved **99% accuracy/F1** on **974** moderator-curated labels and improved generalization over behavior cloning on context-dependent toxicity cases.

Technical Skills

Languages: Python, C++, Java, C
ML & NLP: PyTorch, TensorFlow, Scikit-learn, LangChain, Hugging Face, vLLM, OpenCV, NLTK
Cloud & MLOps: AWS (SageMaker, EC2, S3), Docker, Apache Spark, Kubernetes, Apache Airflow, Jenkins
Data Engineering: SQL, PostgreSQL, MongoDB, Cassandra, Spark, Tableau, Pandas, NumPy