Abhinav Sharma

✉ abhinavs@umass.edu • abhinav-sharma • a1bhinav a1bhinav.netlify.app/

Education

University of Massachusetts, Amherst | M.S. Computer Science | **CGPA: 4.0 /4.0** **Sep 2024 - May 2026**
Coursework: Reinforcement Learning, Systems For Data Science, Industrial Research

Publications

- **[1] Abhinav Sharma**, Franck Dernoncourt, et al. *Test-Time Strategies for More Efficient and Accurate Agentic RAG*. Under review at ACL 2026.
- **[2]** Alexander Erlei, **Abhinav Sharma**, Ujwal Gadiraju. *Understanding Choice Independence and Error Types in Human-AI Collaboration*. CHI 2024.
- **[3]** Lakshmi Sathidevi, **Abhinav Sharma**, Nan Wu, Xun Jiao, Cong Hao. *PreAxC: Error Distribution Prediction for Approximate Computing Quality Control using Graph Neural Networks*. ISQED 2023.

Experience

Amazon | Applied Scientist Intern **Sep 2025 - Dec 2025**

- Architected a large-scale LLM simulation framework generating 5.4K+ synthetic sellers for multi-turn dialogue evaluation, employing a Mixture-of-Experts (MoE) model trained on 6.5M+ internal emails to capture latent persona distributions, behavioral heterogeneity, and conversational styles across seller archetypes.
- Engineered an end-to-end evaluation pipeline to quantify intent fidelity, scope adherence, and tool-call correctness for a production LLM agent, designing a 200-scenario adversarial benchmark that surfaced a 40% tool failure rate and systematic scope misclassifications under realistic multi-turn pressure.
- Translated evaluation findings into concrete mitigation recommendations adopted by the partner team, including targeted prompt restructuring, tool-schema hardening, and a regression harness that enables continuous monitoring of agent behavior across releases prior to production deployment.

IBM | Research Scientist Intern, Almaden Research **May 2025 - Aug 2025**

- Developed a custom Top-K token selection algorithm in C++ using IBM's proprietary parallel programming model on the NorthPole Inference Accelerator, computing top-K logits and routing them to custom downstream layers within the chip's tight on-chip memory and deterministic dataflow constraints.
- Benchmarked the NorthPole implementation against NVIDIA H100 GPU and multi-threaded CPU baselines on end-to-end inference workloads, achieving up to 16x higher energy efficiency and a 10x throughput speedup while maintaining numerical fidelity with reference implementations.
- Partnered with the hardware and compiler teams to profile kernel-level bottlenecks, iterate on memory layout and scheduling decisions, and document deployment patterns that informed NorthPole's evolving LLM inference stack and future accelerator generations.

Adobe | Graduate Student Extern Researcher **Feb 2025 - May 2025**

- Engineered an Agentic RAG framework for multi-hop QA, integrating iterative retrieval with a memory cache module that optimized context reuse and reduced redundant retrievals across long evidence chains.
- Designed a PPO/GRPO fine-tuning pipeline with an LLM-as-a-Judge reward, outperforming Search-R1 by 6.7% accuracy and reducing retrieval turns by 10.5% on HotpotQA, 2WikiMultiHopQA, and MuSiQue.
- Led first-author paper writing including experimental design and ablations over retrieval depth, memory policy, and reward shaping; submitted the resulting work to ACL 2026.

Projects

Optimization-Guided Adaptation of Pretrained Visual Representations for 3D Scene Understanding

- Designed a framework lifting dense 2D features from frozen foundation encoders (DINOv2, CLIP, MAE) into unified 3D representations via calibrated multi-view projection and visibility-weighted fusion.
- Formulated a constrained adaptation objective combining prototype alignment, k-NN graph smoothness, confidence-aware entropy minimization, and trust-region regularization; evaluated on ScanNet v2, S3DIS, and Matterport3D.

GPU-Accelerated Second-Order Rigid Body Dynamics

- Extended the GRiD library with CUDA kernels for second-order analytical derivatives of inverse/forward dynamics on branched-tree models (IIWA, HyQ, Atlas), exploiting limb-level parallelism and shared-memory tiling.
- Implemented and validated tensor-form IDSVA-SO expressions against a CPU reference, removing synchronization points and coalescing memory accesses to improve real-time robotic control precision.

Skills

- **Programming Languages:** Python, C++, CUDA, Java, MATLAB, SQL
- **ML Frameworks:** PyTorch, TensorFlow, PyTorch-Geometric (PyG), Deep Graph Library (DGL)
- **Cloud & Tools:** AWS (SageMaker, EC2), Redis, Git, Node.js, MERN Stack
- **Hardware & Systems:** NVIDIA H100, IBM NorthPole, Edge Devices (Oculus, Robotics)