Elias Hossain

Research Scientist | Reliable AI, Multi-Agent Systems, and Adaptive Inference

Ph.D. Candidate | Expected Graduation: May 2028 | Open to Summer 2027 Research Internships and Full-Time Roles

mdelias.hossain@ucf.edu • Orlando, FL • Website • GitHub • LinkedIn • Scholar

Research Profile

Focus	Reliable AI at inference time: methods that adapt, calibrate, and safeguard LLMs, agents, and RL policies **without retraining**, plus rigorous evaluations of when they fail.
Publications	**8 first-author papers**. Accepted at **NeurIPS 2026** and **EACL 2027**, and in **Findings of AACL-IJCNLP 2026**.
Key results	Multi-agent prompt-attack success cut from **80% to 3%**; runtime agent safety at **under 0.1%** latency overhead; about **20% lower** calibration error with frozen weights.
Industry	**3 years** of production ML, including a speech recognition system deployed in a national e-governance service.

Education

University of Central Florida, Orlando, FL — Aug 2025 – May 2028 (Expected)
Ph.D. in Industrial Engineering, Computer Science concentration (Post-Master's Track), GPA 3.87/4.00
Research area: machine learning (LLM reliability, post-training, agent safety)
Advisor: Dr. Niloofar Yousefi | *Qualifying Exam: Passed*

Mississippi State University, Mississippi State, MS — Jan 2024 – Jul 2025
M.S. in Computer Science (Research), GPA 3.88/4.00
Thesis: "TopNet R1: A Multi-Stage AI Framework for Topic Discovery in Scientific Abstracts" | *Advisor: Dr. Andy Perkins*

Research and Technical Experience

Graduate Research Assistant — Aug 2025 – Present
Complex Adaptive Systems Laboratory, University of Central Florida — *Orlando, FL*

- Built multi-agent preference-data verification, reducing Qwen2-7B summarization refusals from **14.8% to 0.5%** without reducing win rate.
- Built a training-free inter-agent defense (ChannelGuard) that cut prompt-attack success from **80% to 3%** on an open model across 5,700+ runs.
- Released 5 public ML benchmarks (**74,590+ records**) with code.
- Trained and evaluated Qwen, Llama, and Mistral models from 1.5B to 8B parameters on UCF's Newton HPC cluster and a 2× NVIDIA TITAN RTX workstation.

Graduate Research Assistant — May 2025 – Jul 2025
Predictive Analytics and Technology Integration (PATENT) Lab, Mississippi State University — *Mississippi State, MS*

- Developed contrastive-learning and uncertainty-estimation methods for detecting rare malicious events in imbalanced security logs.
- Implemented FAISS retrieval and RAGAS evaluation for BarkPlug V.2, a retrieval-augmented assistant deployed to the campus community.

Software Engineer — Dec 2022 – Nov 2023
REVE Systems — *Dhaka, Bangladesh*

- Reduced speech-recognition word error rate by **20%** and inference time by **15%** using DeepSpeech and Wav2Vec 2.0.
- Deployed the models in a production e-governance voice interface.

AI/ML Researcher — Oct 2020 – Nov 2022
Time Research & Innovation (Timerni) — *Remote / UK*

- Developed deep-learning models for COVID-19 screening from chest X-ray and CT images.
- Contributed ML to a telemedicine and health-advice platform (web, iOS, Android) connecting patients, doctors, pharmacies, and hospitals; co-authored the system paper at *ICICT 2021* (Springer LNNS).

Selected Publications and Preprints (all first-author)

[1] **When Policies Cannot Be Retrained: A Unified Closed-Form View of Post-Training Steering in Offline Reinforcement Learning**
E. Hossain, M. J. I. Basher, I. Garibay, O. Garibay, N. Yousefi.
NeurIPS 2026, Main Conference Track Accepted arXiv Project Code
Adapts frozen offline-RL actors at deployment with no retraining; degrades gracefully under poor priors across 4 D4RL tasks.

[2] **UAT-LITE: Inference-Time Uncertainty-Aware Attention for Pretrained Transformers**
E. Hossain, S. R. Dipta, S. Neupane, R. Rana, R. Shwartz-Ziv, I. Garibay, N. Yousefi.
EACL 2027, Main Conference Track Accepted arXiv Project Code
Inference-time only, pretrained weights unchanged: about 20% lower calibration error vs. BERT-base, accuracy preserved.

[3] **NEXUS: Structured Runtime Safety for Tool-Using LLM Agents**
E. Hossain, M. M. H. Nipu, T. N. Ornee, R. Rana, N. Yousefi.
Findings of AACL-IJCNLP 2026 Accepted arXiv Project Code
0.205 ms median latency (under 0.1% overhead), with 0% indirect prompt-injection success at 99% benign allow.

[4] **ChannelGuard: From Borrowed Safety to Application-Controlled Defenses in Multi-Agent LLMs**
E. Hossain, M. M. H. Nipu, F. T. J. Faria, S. R. Dipta, M. Sheikh, T. N. Ornee, N. Yousefi.
Transactions on Machine Learning Research (TMLR) Under Review arXiv Project
Shows equal safety scores can rest on different protections (provider filter vs. model); training-free handoff checks stop all tested tool-output attacks on five models.

[5] **Right Knowledge, Wrong Answer: Characterizing Parametric Temporal Conflict in Open-Weight Language Models**
E. Hossain, S. Saha, T. N. Ornee, S. S. Jennifer, U. C. Biswas, S. R. Dipta, R. Rana, N. Yousefi.
AAAI 2027 Under Review, Phase 2 arXiv Project Code
Verified 8,746-transition Wikidata benchmark; activation patching flips 72–85% of cases across four open-weight models.

[6] **Agreement Overstates Evidence: Error Dependence in LLM Judge Consensus**
E. Hossain, N. Yousefi, S.-N. Lim. *ICLR 2027* Under Review arXiv Project Code
Ten LLM judges carry about as much information as 3.5 independent ones; shared errors flip up to 28% of significance calls.

Selected Technical Projects

- **MiniHealthLM** Code: Built a ∼600M-parameter Qwen-style Transformer from scratch in PyTorch (GQA, RoPE, RMSNorm) with medical-text preprocessing, PII filtering, custom BPE tokenization, 1024-token context, and checkpointing.
- **MLDriftGuard** Code: Installable Python package for numerical and categorical drift detection (KS and chi-squared tests) with visualizations and mitigation suggestions. *SciPy, pandas*

Technical Skills

RL & Post-training	RLHF, PPO, GRPO, DPO, SFT, preference learning, offline RL
LLM Systems	PyTorch, Hugging Face Transformers, vLLM, LoRA/QLoRA, multi-GPU training
Agents & Safety	Tool-using and multi-agent LLM pipelines, prompt-injection defense, runtime guardrails
Interpretability	Activation patching, activation steering, representation analysis
Evaluation	LLM-as-judge evaluation, statistical testing, uncertainty calibration, RAG (FAISS, RAGAS)
Tools	Python, SQL, R, Git, Docker, Azure, GCP

Service

- **Reviewer:** NeurIPS 2026 Workshop on Trustworthy AI for Good; ICML 2026 Workshop on Generative and Agentic AI for Biology