Elias Hossain

CONTACT INFORMATION

12800 Pegasus Dr
Orlando, FL 32816

mdelias.hossain@ucf.edu
Google Scholar | Personal Site | GitHub | LinkedIn

PhD researcher in machine learning working on trustworthy and safe large language models: robust alignment under corrupted preference supervision, calibrated uncertainty for foundation models, runtime safety for tool-using agents, and inference-time control of frozen policies. Lead author across NeurIPS / ICLR / ACL venues and Q1 journals, with biomedical discovery and clinical decision support as primary application domains.

RESEARCH PROFILE

I build methods that keep large language models safe, calibrated, and auditable when deployed in high-stakes settings where humans cannot manually verify every output. My work spans the model lifecycle, from alignment and uncertainty estimation to agent runtime safety and post-deployment control, and is validated on two demanding application domains: antimicrobial-resistance (AMR) discovery and clinical decision support. I drive research end to end: problem formulation, theory, large-scale empirical evaluation, open-source release, and first-author publication at top-tier ML venues and flagship Q1 journals.

RESEARCH IMPACT

- **915 total citations**, **h-index 12**, **i10-index 12** (Google Scholar, June 2026).
- **Accepted as a Regular Paper in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*** (Q1, SCIE-indexed, JIF 8.7, 5-Year IF 9.1): a lead-author comprehensive review of safe and scalable collaboration in multi-agent LLM systems, accepted after ~10 months of peer review.
- **491 citations** on a single lead-author systematic review in *Computers in Biology and Medicine* (2023), an Elsevier Q1 flagship journal in biomedical informatics with a ~12% acceptance rate.
- **63 citations** on the BarkPlug V.2 university-deployed chatbot paper (IEEE FIE 2024).
- **62 citations** on the multimodal brain-tumor segmentation paper (*CMC*, 2022).
- **Peer-reviewed and accepted: 12+ first-author publications** in Q1 venues (*IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *Computers in Biology and Medicine*, *Briefings in Bioinformatics*, *Scientific Reports*, *Frontiers in Bioinformatics*), an accepted **ICLR 2026 Trustworthy AI Workshop** paper, and IEEE conference proceedings (FIE, ICICT).
- **In the pipeline:** first-author manuscripts under review at **NeurIPS 2026** (4) and **ACL ARR** (2), plus active first-author drafts targeting ICLR and AAAI.
- **Sustained peer-review service for 7 international journals** across Nature Portfolio, Elsevier, PLOS, SAGE, Taylor & Francis, and Springer, plus workshop reviewing at an ICLR-affiliated venue.
- **4 publicly released benchmarks** on Hugging Face Hub spanning **74,590+ records** for biomedical NLP, knowledge-conflict evaluation, and cancer-document classification, alongside **14+ open-source research repositories**.

RESEARCH OVERVIEW

My research spans four methodological directions in trustworthy AI, anchored by published and accepted work, and grounded in biomedical and clinical applications.

Trustworthy alignment under corrupted supervision. RLHF/DPO pipelines assume clean preference data and independent judges, both of which break in practice. I introduced **multi-agent verification of preference supervision** (*Fault-Tolerant Preference Alignment via Multi-Agent Verification*, ICLR 2026 Trustworthy AI Workshop; *SAVe V.1*, Amazon Trusted AI Symposium, 2026) to make alignment robust to corrupted feedback.

Calibrated uncertainty for foundation models and clinical AI. Reliable deployment requires models that know when they are wrong. I developed **UAT-LITE**, an inference-time uncertainty-aware attention mechanism for pretrained transformers (under ACL ARR commitment review), and

MedBayes-Lite, a Bayesian uncertainty-quantification framework for safe clinical decision support, building on my lead-author systematic review of NLP in electronic health records (*Computers in Biology and Medicine*, 2023).

Runtime safety for tool-using agents. Tool-using LLM agents take consequential actions before any human can intervene. I am the first author of **NEXUS**, a structured runtime safety monitor that combines deterministic rules, argument-level inspection, and a calibrated risk scorer to route plans to one of four graded interventions (allow / block / confirm / revise), reporting a 27.3 percentage-point gain in intervention accuracy over rule-only baselines.

Inference-time control of frozen policies. Deployed policies often cannot be retrained, yet still need correction. My work *When Policies Cannot Be Retrained* (under review, NeurIPS 2026) gives a unified closed-form view of post-training steering in offline RL, and **MIRAGE-BIO** applies inference-time screening to biomolecular candidate selection under unreliable surrogate evidence.

Biomedical applications and open evaluation. I ground these methods in antimicrobial-resistance research as first author of a comprehensive Q1 review of computational paradigms for AMR (*Briefings in Bioinformatics*, 2026, synthesizing 156 studies from 2016–2025) and of **BIOGEN**, an evidence-grounded multi-agent reasoning framework for transcriptomic interpretation in AMR (*Frontiers in Bioinformatics*, 2026), and I support the community through open biomedical benchmarks released on Hugging Face Hub.

Education

University of Central Florida (UCF), College of Engineering and Computer Science, Orlando, FL, USA

Ph.D. in Industrial Engineering (Concentration in Computer Science) — Aug 2025 – Present

- Advancing to Ph.D. candidacy (qualifying examination, Fall 2026); research focus on trustworthy AI for large language models, with antimicrobial resistance and clinical reasoning as primary application domains.
- Advisor: Dr. Niloofar Yousefi, Assistant Professor, Department of Industrial Engineering & Management Systems; affiliated with the Complex Adaptive Systems Laboratory.

Mississippi State University (MSU), Mississippi State, MS, USA

M.S. in Computer Science (Research), GPA 3.88/4.00 — Jan 2024 – July 2025

- Thesis: *A Multi-Stage AI Framework for Topic Discovery in Scientific Abstracts*.
- Advisors: Dr. Andy D. Perkins and Dr. Shahram Rahimi; degree conferred August 2025.

Daffodil International University, Dhaka, Bangladesh

B.S. in Software Engineering — Jan 2016 – Sep 2020

Research Experience

UCF Complex Adaptive Systems Laboratory, Orlando, FL, USA

Graduate Research Assistant — Aug 2025 – Present

- Lead an independent research program on trustworthy and safe large language models: robust alignment under corrupted preference supervision, calibrated uncertainty for foundation models, runtime safety for tool-using agents, and inference-time control of frozen policies.
- First-author results submitted to NeurIPS, ICLR, and ACL ARR, with peer-reviewed work accepted at the ICLR 2026 Trustworthy AI Workshop and presented at the Amazon Trusted AI Symposium.

- Apply these methods to antimicrobial-resistance discovery and clinical decision support, including **BIOGEN** (multi-agent transcriptomic reasoning) and **MedBayes-Lite** (Bayesian clinical uncertainty quantification).

Department of Computer Science & Engineering, Mississippi State University, MS, USA

Graduate Research Assistant May 2025 – July 2025

- Built a contrastive learning framework for anomaly detection in large-scale cybersecurity event logs.
- Co-developed **BarkPlug V.2**, a RAG-based chatbot deployed at MSU; engineered FAISS-based retrieval pipelines evaluated via RAGAS metrics and the System Usability Scale.

Selected Peer-Reviewed Publications

Highest-impact peer-reviewed work (published or accepted), lead-author first. Full publication list and live citation counts on Google Scholar.

7A. Q1 / Flagship Journal Articles

Elias Hossain, Md. Mehedi Hasan Bhuiyan Nipu, Mohammad Sakib Mahmood, Md. Jakir Hossen, M. F. Mridha. "Safe and Scalable Collaboration in Multi-Agent LLM Systems: A Comprehensive Review." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (IEEE), 2026. *Accepted as a Regular Paper (in production).* **[Q1; SCIE-indexed flagship IEEE Transactions journal; JIF 8.7, 5-Year IF 9.1.]**

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R. Pisani, Kathryn Turner. "Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review." *Computers in Biology and Medicine* (Elsevier), 2023. [Available Online] **[Q1 in Medicine / Computer Science Applications; Elsevier flagship for biomedical informatics; ~12% acceptance rate; 491 Google Scholar citations.]**

Elias Hossain and Niloofar Yousefi. "Computational paradigms for antimicrobial resistance prediction: integrating multi-omics, structural modeling, and foundation artificial intelligence systems." *Briefings in Bioinformatics* (Oxford University Press), vol. 27, no. 3, 2026. [Available Online] **[Q1; IF 7.7 (2024 JCR); OUP flagship review journal in computational biology and bioinformatics.]**

Elias Hossain, Mehrdad Shoeibi, Ivan Garibay, Niloofar Yousefi. "BIOGEN: evidence-grounded multi-agent reasoning framework for transcriptomic interpretation in antimicrobial resistance." *Frontiers in Bioinformatics* (Frontiers Media), vol. 6, 2026 (Section: Drug Discovery in Bioinformatics). [Available Online] **[IF 3.9; CiteScore 4.9 (2024); major open-access bioinformatics venue.]**

Elias Hossain, Tasfia Nuzhat, Shamsul Masum, Shahram Rahimi, Noorbakhsh Amiri Golilarz. "R-GAT: Graph Neural Networks for Cancer Document Classification." *Scientific Reports* (Nature Portfolio), 2026. [Available Online] **[Nature Portfolio flagship multidisciplinary open-access journal; IF 3.9 (2024 JCR).]**

Noorbakhsh Amiri Golilarz, **Elias Hossain**, Shahram Rahimi, Hossein Karimi. "A hybrid approach for pattern recognition and interpretation in age-related false memory." *Frontiers in Psychology* (Frontiers Media), 2025. [Available Online] **[IF 2.9; CiteScore 6.3 (2024); peer-reviewed open-access journal in cognitive science.]**

7B. Peer-Reviewed Workshop & Symposium Papers (AI Safety / Trustworthy AI)

Elias Hossain, Maryam Rahimimovassagh, Subash Neupane, Mohammad Jahid Ibna Basher, Ivan Garibay, Niloofar Yousefi. "Fault-Tolerant Preference Alignment via Multi-Agent Verification." *Accepted at the ICLR 2026 Trustworthy AI Workshop* (workshop track, peer-reviewed). **[ICLR is a top-3 venue in machine learning; h5-index 304.]**

Elias Hossain, Maryam Rahimimovassagh, Niloofar Yousefi. "SAVe V.1: Multi-Agent Safe Alignment Verification for Preference-Based LLM Optimization." *Trusted AI Symposium, Amazon JFK27, Manhattan, NY, USA*, 2026. **[Poster; industry–academic symposium on safe and responsible AI deployment.]**

7C. Peer-Reviewed Conference Proceedings

Subash Neupane, **Elias Hossain**, Jason Keith, Himanshu Tripathi, Farbod Ghiasi, Noorbakhsh Amiri Golilarz, Amin Amirlatifi, Sudip Mittal, Shahram Rahimi. "From questions to insightful answers: Building an informed chatbot for university resources." *Proceedings of the 2024 IEEE Frontiers in Education Conference (FIE)*, 2024. [Available Online] **[IEEE flagship conference in engineering and computing education; h5-index 31; 63 citations.]**

Elias Hossain, Umesh Biswas, Charan Gudla, Sai Phani Parsa. "Learning Robust Representations for Malicious Content Detection via Contrastive Sampling and Uncertainty Estimation." *Accepted at the 9th International Conference on Information and Computer Technologies (ICICT)*, USA, 2026. [arXiv preprint] **[IEEE-affiliated international conference on information and computing technologies, peer-reviewed.]**

7D. Peer-Reviewed Book Chapters (Springer Nature)

Seven peer-reviewed book chapters in Springer Nature series (LNCS, LNNS, CCIS, LNEE, LNICST), all Scopus-indexed and DBLP-listed: edge-based learning under adversarial noise (CCIS vol. 2720, 2026); machine learning for COVID-19 risk-factor identification (LNNS vol. 436, 2022); deep learning for COVID-19 chest X-ray diagnosis (Advances in Sustainability Science and Technology, 2022); next-generation telemedicine and health-advice system (LNNS vol. 236, 2022); LLM-based English-to-Bangla translation (LNNS vol. 1014, 2024); mutation-triggering for genetic algorithms (LNEE vol. 678, 2020); and assistive mobile navigation for the visually impaired (LNICST vol. 325, 2020). Full citations on Google Scholar.

7E. Additional Refereed Journal Publications

Additional peer-reviewed first- and co-author journal papers, including: **multimodal brain tumor segmentation** (*Computers, Materials & Continua*, 2022; SCIE-indexed; **62 citations**); **Surformer v1** multimodal transformer for tactile-vision surface classification (*Information* (MDPI), vol. 16, no. 10, 2025); **chronic kidney disease diagnosis** via ensemble ML (*Machine Learning with Applications* (Elsevier), 2022); deep-learning systematic review for COVID-19 (*SN Computer Science*, 2022); machine-learning forecasting for mental stress (*CMC*, 2022); diabetes-diagnosis mobile health system (*CMC*, 2022); data-driven news retrieval (*Intelligent Automation & Soft Computing*, 2024); and two papers in *International Journal of Electrical and Computer Engineering* (2019, 2021). Full citations on Google Scholar.

Manuscripts Under Review

First-author manuscripts currently under peer review at top-tier international venues.

Elias Hossain, Mohammad Jahid Ibna Basher, Ivan Garibay, Ozlem Garibay, Niloofar Yousefi. "When Policies Cannot Be Retrained: A Unified Closed-Form View of Post-Training Steering in

Offline RL." *Under review at NeurIPS 2026 (Main Track).* **[NeurIPS Main Track: top-3 venue in machine learning; h5-index 337; ~26% acceptance rate.]**

Elias Hossain, et al. "MIRAGE-BIO: A Robust Inference-Time Screening System for Biomolecular Candidate Selection under Unreliable Surrogate Evidence." *Under review at NeurIPS 2026 (Main Track).* **[NeurIPS Main Track.]**

Elias Hossain, et al. "BioDivergence: A Leakage-Aware Benchmark for Contextual Biomedical Contradiction Evaluation." *Under review at NeurIPS 2026 (Datasets & Benchmarks Track).* **[Premier dedicated venue for community-adopted datasets and benchmarks in ML.]**

Elias Hossain, Subash Neupane, Ivan Garibay, Niloofar Yousefi. "Before the Gradient: Preference Alignment Requires Verified Supervision." *Under review at NeurIPS 2026 (Position Track).* **[NeurIPS Position Track: peer-reviewed venue for position papers shaping the ML research agenda.]**

Elias Hossain, Md Mehedi Hasan Nipu, Tasfia Nuzhat Ornee, Rajib Rana, Niloofar Yousefi. "NEXUS: Structured Runtime Safety for Tool-Using LLM Agents." *Under review at ACL ARR 2026 (May cycle).* [Code] **[ACL ARR is the shared peer-review system for the flagship NLP venues ACL, EMNLP, and NAACL.]**

Elias Hossain, Shubhashis Roy Dipta, Subash Neupane, Rajib Rana, Ravid Shwartz-Ziv, Ivan Garibay, Niloofar Yousefi. "UAT-LITE: Inference-Time Uncertainty-Aware Attention for Pretrained Transformers." *ACL ARR 2026 (March cycle); under commitment review.* [arXiv preprint] **[ACL ARR is the rolling peer-review system shared by the flagship NLP venues ACL, EMNLP, and NAACL.]**

Elias Hossain, Mehedi Hasan Nipu, Maleeha Sheikh, Rajib Rana, Subash Neupane, Niloofar Yousefi. "MedBayes-Lite: Bayesian Uncertainty Quantification for Safe Clinical Decision Support." *Under review at IEEE Access, 2026.* [arXiv preprint] **[IEEE flagship multidisciplinary open-access journal; IF 3.6 (2024 JCR); SCIE-indexed.]**

Current Research Focus

Active first-author research I am currently developing.

CorrFilter: Structured Co-Failure in LLM Judge Banks. Studying whether LLM-judge banks fail independently, the assumption behind k-of-n consensus filtering. I measure the inter-judge error-correlation structure on a stratified calibration set, show that consensus filtering becomes unreliable under correlation drift, and build a small-calibration adaptive filter that recovers near-oracle false-retention rates. In collaboration with **Prof. Ser-Nam Lim** (Associate Professor of Computer Science, UCF; previously Senior Research Scientist Manager at Meta AI / Facebook AI Research).

INFOSHIELD: Safety-Constrained Information Flow for Multi-Agent LLM Systems. Developing an information-theoretic safety framework that regulates unsafe information propagation across the planner, memory, tool-use, retrieval, and inter-agent channels of multi-agent systems, evaluated against prompt injection, reasoning-state poisoning, and consensus-collapse attacks.

CareBench: Measuring Unsafe Reasoning behind Correct Diagnoses. Building a process-level clinical-agent benchmark that operationalizes "lucky correctness" (a correct diagnosis reached through an unsafe reasoning trajectory) via pathway-structured metrics over a hidden patient state and an EHR-style toolset. In collaboration with a research scientist at **Amazon** and clinical faculty at **Meharry Medical College**.

BeliefNet: Equilibrium Reasoning for Inconsistency Detection in Frozen LLMs. Designing a low-capacity equilibrium reasoner over a typed belief graph on top of a frozen LLM; it matches a symbolic oracle (AUROC 1.000) on a synthetic inconsistency ladder where text-only baselines collapse to chance.

Open-Source Software & Public Benchmarks

Public datasets and code released to the international research community in support of trustworthy AI and biomedical discovery. All artifacts available at huggingface.co/EliasHossain and github.com/eliashossain001.

Public Benchmarks (Hugging Face Hub).

- **nanobubbleeval** [link]: schema-constrained extraction benchmark over nanocarrier literature, 51,566 records with 40 gold-annotated examples.
- **BioDivergence-Silver-v1.0** [link]: 11,900 biomedical claim pairs for natural-language inference and contradiction detection.
- **ptc-benchmark** [link]: 9,250-example temporal QA benchmark for knowledge-conflict handling in LLMs.
- **CancerAbstracts** [link]: 1,874 biomedical abstracts labeled by cancer type (used in R-GAT).

Selected Research Code Repositories.

- **NEXUS** [code]: runtime safety monitor for tool-using LLM agents.
- **MIRAGE-BIO** [code]: robust inference-time biomolecular candidate selection.
- **BioDivergence** [code]: silver benchmark and evaluation framework for biomedical contradiction detection.
- **PoE-Composition** [code]: inference-time steering of frozen offline policies via product-of-experts refinement.
- **Temporal Attractor Steering (TAS / PTC)** [code]: inference-time detection and steering of parametric temporal conflicts in language models.
- **qwen-scratch-0.6B** [code]: end-to-end Qwen-style 0.6B transformer built from scratch in PyTorch.
- **Continual-pre-training** [code]: domain-adaptation recipes for LLMs without catastrophic forgetting.
- **efficient-gemma3-finetuning** [code]: LoRA / QLoRA fine-tuning for Gemma3 on single- and multi-GPU setups.
- **MiniHealthLM** [code]: domain-adapted language model for healthcare and clinical tasks.
- **Bone-Fracture-Detection** [code]: deep-learning pipelines (ResNet / DenseNet / EfficientNet) for X-ray fracture detection.

Invited Talks & Conference Presentations

Invited Talk on artificial intelligence in rural American healthcare, Mississippi Health Innovation Conference, Millsaps College, USA. [Video] 2024

Conference Presentation, "CITE V.1 / BIOGEN: Interpretable RNA-Seq Clustering with an LLM-Based Agentic Evidence-Grounded Framework," 7th Molecular Machine Learning Conference (MoML), MIT Jameel Clinic, Cambridge, MA, USA Oct 2025

Poster, "SAVe V.1: Multi-Agent Safe Alignment Verification for Preference-Based LLM Optimization," Amazon Trusted AI Symposium, Amazon JFK27, Manhattan, NY, USA Jan 2026

Poster, "BIOGEN: Evidence-Grounded Multi-Agent Reasoning for Transcriptomic Interpretation in Antimicrobial Resistance," RECOMB Satellite Workshop on Regulatory & Systems Genomics (RECOMB-RSG), Greece 2026

International Conference on Sentiment Analysis & Deep Learning, Songkla University, Thailand 2024

Earlier IEEE-affiliated international conferences (ICONCS 2020, ICCIT 2020, SPICSCON 2019), Bangladesh. 2019–2020

Professional Service (Peer Review)

Journal Peer Reviewer, 2024–Present. Review activities focus on methodological rigor, calibration, uncertainty analysis, reproducibility, and ethical deployment of AI/ML systems in healthcare and biomedical data science.

- *Scientific Reports* (Nature Portfolio) – since March 2025 (invited via Dr. Ayman El-Baz)

- *Computers in Biology and Medicine* (Elsevier) – since April 2025
- *Computer Methods in Biomechanics and Biomedical Engineering* (Taylor & Francis) – since December 2024
- *PLOS ONE* – since August 2024
- *INQUIRY: Journal of Health Care Provision and Public Health* (SAGE) – since July 2024
- *Digital Health* (SAGE) – since September 2024
- *International Journal of Computational Intelligence Systems* (Springer) – since May 2025

Conference / Workshop Peer Reviewer.
- *ICLR Trustworthy AI Workshop*, 2026.

Selected Honors, Awards & Recognition

- **Graduate Presentation Fellowship**, University of Central Florida, 2025: selective university fellowship funding research presentation at the 7th Molecular Machine Learning Conference (MoML), MIT Jameel Clinic.
- **Finalist**, Dr. Pradeep P. Thevannoor Innovation Awards, India, 2019: selected through a competitive international process to the final round. [Campus News]
- **Recognized international collaborations:** sustained research collaboration with faculty at UCF (Dr. Ivan Garibay, Dr. Niloofar Yousefi), New York University (Dr. Ravid Shwartz-Ziv), Mississippi State University (Dr. Shahram Rahimi, Dr. Andy D. Perkins), the University of Southern Queensland (Dr. Rajib Rana), and the University of Portsmouth (Dr. Adrian Hopgood, Dr. Alice Good, Dr. Alexander Gegov).
- **Earlier student-era awards (2017–2023):** Fourth Industrial Revolution Skills Summit Award (2023); DIU Research Award (2019); Winner, IEEE SS12 Innovation Challenge & Maker Fair (2018), recognizing the Kidnap Prevention Mobile App published in IEEE Xplore; Divisional Champion & Global Nominee, NASA Space Apps Challenge (2017).

Industry Experience

REVE Systems, Dhaka, Bangladesh

Software Engineer — Dec 2022 – Nov 2023

- Engineered and deployed speech-recognition models (DeepSpeech, Wav2Vec 2.0), achieving **20% lower Word Error Rate** and **15% faster inference** in production.
- Delivered a voice-enabled extension for Bangladesh government digital initiatives, supporting large-scale e-governance services.

Time Research & Innovation, Portsmouth, United Kingdom

Senior AI/ML Researcher (Remote) — Oct 2020 – Nov 2022

- Designed deep-learning architectures for COVID-19 diagnosis from X-ray and CT imagery.
- Co-developed a telemedicine platform for the UK healthcare system during the pandemic.

Teaching

Graduate Teaching Assistant, Department of Industrial Engineering & Management Systems, University of Central Florida — Aug 2025 – Present

- STA 3032: Probability and Statistics for Engineers.

Graduate Teaching Assistant, Department of Computer Science & Engineering, Mississippi State University — Jan 2024 – May 2025

- Supported instruction in five advanced courses: Software Testing & Quality Assurance, Computer Forensics, Machine Learning & Soft Computing, Artificial Intelligence Fundamentals, and Advanced Machine Learning.

TECHNICAL SKILLS

- **Generative AI & Modeling:** LLM fine-tuning (RLHF, DPO, SFT), LLM reasoning, multimodal and agentic AI systems; probabilistic and Bayesian inference, uncertainty quantification, model calibration, graph neural networks.
- **Frameworks & Data:** PyTorch, TensorFlow, scikit-learn, Hugging Face Transformers; Apache Spark, Hive, Presto, FAISS, Chroma, RAG pipelines.
- **Programming & Infrastructure:** Python, R, JavaScript, Kotlin, C#, SQL, Bash; Docker, Git, Azure, Google Cloud, MySQL, PostgreSQL.