Zafaryab Haider

+1 (207) 719-0411 | zafaryab.haider@maine.edu | Orono, ME, USA | zafaryab |

SUMMARY

AI security and trustworthy-ML researcher: adversarial and bit-flip attacks on multimodal and speech models, RLHF training-signal integrity, harm evaluation for multi-agent LLM systems, and the evaluation tooling behind them.

EXPERIENCE

University of Maine | Graduate Research Assistant | Orono, ME, USA ***August 2022 – Present***

- Built **BLADE** and **SMASH**, differentiable bit-level fault-injection frameworks for quantized vision–language and speech-recognition models; BLADE reached **2.4x higher attack success rate** than a prior baseline at a 20-bit budget.
- Built **COBRA** and **ACT**: consensus-based reward for RLHF under malicious feedback (**˜73% vs. 48.78%**, **˜23%** over prior work at DF=10) and anchor-calibrated trusted/corrupted cohort recovery.
- Built distributed LLM inference across **heterogeneous NVIDIA Jetson and RTX GPUs**; models up to 15B.
- Built an internal **Streamlit** evaluation platform: structured scoring, diff review, CSV export.
- **AI/ML Consultant**, ARCSIM (2024–2026); guest lectures for **ECE 479/579** and the UMaine **AI Seminar**.

Aligarh Muslim University | Assistant Professor | Aligarh, India ***January 2015 – August 2022***

- Designed and taught information security, networking, data structures and C programming to **200+ students annually**.
- Led and mentored applied ML and systems projects across healthcare ML, IoT sensing and intelligent decision support.

SELECTED RESEARCH & IP

Peer-reviewed & accepted

- Nafi, **Haider, Z.** et al. *DASH: MetaAttack Framework for Synthesizing Effective & Stealthy Adversarial Example.* **CVPR**.
- **Haider, Z.** et al. *A Framework for Mitigating Malicious RLHF Feedback in LLM Training using Consensus Based Reward.* **Nature Scientific Reports**, 2025.
- **Haider, Z.** et al. *ACT: Anchor-Calibrated Trust Estimation When Corrupted Sources Form Majority.* **IEEE CARS'26**.
- Rahaman, Xu, **Haider, Z.** et al. *(A)iSpy: Parasitic Trojans for Machine Learning Infrastructure.* **ACM CCS**.

Under review

- **Haider, Z.** et al. *BLADE: Bit-Flip Attacks for Semantic Misdirection in Quantized Image Captioning.* **IEEE TIFS**.
- **Haider, Z.** et al. *SMASH: Probing Speech Recognition Robustness via Semantically Targeted Bit Flips.* **NeurIPS**.
- Rahman, **Haider, Z.** et al. *HARP: Measuring Harm Amplification in Multi-Agent LLM Systems.* **IEEE S&P**.

Patents & IP — four invention families

- **Automated Neural Architecture Generation.** PCT Application No. PCT/US2025/013847, 2025.
- **COBRA.** PCT application filed March 9, 2026; priority to U.S. Provisional Application No. 63/768,361.
- **BLADE.** U.S. Provisional Patent Application No. 63/930,131.
- **Explainable Agentic Cross-PDK Migration Difficulty.** U.S. Provisional Patent Application No. 64/075,062, 2026.

SELECTED PROJECTS & SYSTEMS

INTACT September 2026 – Present

Studying whether autoregressive-model logit signals can identify unreliable or incorrect generated output.

REWIRE: Grid Reconfiguration World Models August 2026 – Present

Topology-aware world model ranking grid line-switching candidates via learned rollouts, verified by exact AC power flow.

GRANTS & FELLOWSHIPS

SentriAlign: Training-Signal Integrity for LLMs — Co-Investigator, MIRTA 8.0 Accelerator ($25,000) 2026

Maulana Azad National Fellowship (MANF), Government of India 2014

Graduate Aptitude Test in Engineering (GATE) Scholarship, Government of India August 2011 – July 2013

TECHNICAL SKILLS

Languages & Tools: Python, C/C++, SQL, Shell, Git, Jupyter, LaTeX
Frameworks: PyTorch, scikit-learn, NumPy, CuPy, Pandas, Streamlit, REST APIs
AI/ML & Security: adversarial ML, bit-flip & fault injection, RLHF and reward modeling, LLM and agentic evaluation
Systems & Edge: Slurm, Apptainer, NVIDIA containers, multi-GPU/HPC, GCP, Jetson (Orin Nano, Xavier NX, AGX)

EDUCATION

Ph.D., Electrical and Computer Engineering — University of Maine, Orono, ME, USA August 2022 – Present

M.Tech., Computer Engineering — Aligarh Muslim University, India (GPA 9.33/10.0) August 2011 – July 2013