Research Statement

Redwanul Karim

My research focuses on AI safety, model integrity, and behavioral reliability in adaptive machine learning systems, with particular emphasis on agentic and language model based systems operating under interaction and distribution shift. I study mechanisms for verification, stress testing, and behavioral constraint that go beyond aggregate performance metrics. Methodologically, my work lies at the intersection of machine learning security, efficient machine learning, and interpretable machine learning.

At Fraunhofer IIS, I worked on watermarking and ownership verification methods for federated learning under heterogeneous client participation. I implemented several black-box and white-box watermarking approaches, including zero-knowledge-proof watermark verification, WAFFLE, and class-hidden client-side watermarking, for GNSS jammer detection and localization models. These studies covered both black-box and white-box verification scenarios and exposed practical challenges in ownership verification under distributed training. This line of work strengthened my interest in model provenance, attribution, and authenticity guarantees for continuously updated systems.

At the FAU Pattern Recognition Lab, I worked on physics informed graph neural networks for AC power flow prediction, where outputs are required to satisfy physical consistency constraints. I implemented attention based GNN architectures and applied LoRA based parameter efficient fine tuning to reduce trainable parameters while preserving most of the performance gains. This work demonstrated that architectural structure and domain constraints can improve reliability and reduce unstable behavior. It also motivated my interest in constraint driven approaches for agent and language model systems.

Taken together, these projects showed that strong benchmark performance often does not translate into reliable behavior under distribution shift, adversarial conditions, or extended interaction. The gap is more pronounced in adaptive systems whose parameters or usage context evolve over time. This observation motivates my focus on safety oriented evaluation and robustness testing for adaptive and agent based systems.

My current research questions include:

- How can we design behavioral, cryptographic, and provenance signatures that remain detectable for continuously adapted, reused, or tool using agents?
- How can we build evaluation protocols for long horizon agent interaction that reveal hidden objective drift, situational deception, and loss of controllability over time?
- Can interpretability and attribution signals be converted from post hoc explanations into online safety monitors or constraint signals?
- How do parameter efficient adaptation methods such as LoRA, quantization, and distillation affect safety, controllability, ownership verification, and red-team robustness after deployment?
- How should we stress test contextual and memory enabled agents whose behavior changes over time and interaction history?

My goal is to develop mechanisms and evaluation frameworks that allow the safety and reliability of advanced AI systems to be tested, audited, and falsified under realistic interaction settings. I am particularly interested in research that connects theoretical foundations with practical system implementations for safety in high-stakes, safety-critical environments.