Sifat Muhammad Abdullah

+1 (540)-449-2710 | sifat@vt.edu | https://sifatmd.github.io | Google Scholar

EDUCATION

Virginia Tech, Ph.D. in Computer Science, advisor: Dr. Bimal Viswanath — Jan 2021 - expected Apr 2026

BUET, B.S. in Computer Science and Engineering (GPA: 3.91/4.0) — 2015 - 2019

RESEARCH INTERESTS

Security and Adversarial Robustness of Large Multimodal Models, LLMs & Generative AI Defenses, Improving Defenses with better Content Semantics understanding using Multimodal Foundation models, toxicity mitigation in Large Language Models.

SELECTED PUBLICATIONS

[**IEEE S&P'24**] **1st author**. "*An Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape*".

[**ACSAC'23**] **2nd author**. "*A First Look at Toxicity Injection Attacks on Open-domain Chatbots*".

[**IEEE S&P'23**] **2nd author**. "*Deepfake Text Detection: Limitations and Opportunities*". Dataset requested by **143 research groups**.

SELECTED PROJECTS

Deepfake Image Detection | Published in **IEEE S&P'24**
- Studied 8 state-of-the-art deepfake image detectors using Diffusion and GAN-based text-to-image generators
- Developed adversarial attacks using LoRA and Vision Foundation models without adding adversarial noise
- Used metrics for measuring attack success, along with underlying semantic meaning and quality of images
- Achieved more than 70% recall score degradation against most of the deepfake image detectors

Toxicity Injection Attacks | Published in **ACSAC'23**
- Studied toxicity injection attacks on chatbots after deployment in a Dialog-based Learning setup
- Evaluated BART and BlenderBot LLM chatbots
- Proposed fully automated indiscriminate and backdoor attacks using public LLMs eliciting up-to 60% response toxicity rate

Deepfake Text Detection | Published in **IEEE S&P'23**
- Collected and released real-world deepfake text dataset, including T5 and GPT-3 powered bots' data
- Evaluated state-of-the-art deepfake text detectors, e.g., BERT and GPT-2 based defenses
- Developed fully black-box and low-cost adversarial attack without access to defender or surrogate model
- Our adversarial attack achieves up-to 91.3% evasion rate while maintaining linguistic quality of text

EXPERIENCE

Virginia Tech SecML Lab – Graduate Research Assistant — Jan 2022 - Present

Virginia Tech – Graduate Teaching Assistant — Jan 2021 - Dec 2021

BUET DataLab – Graduate Research Assistant — Jan 2020 - Dec 2020

REVE Systems – Software Engineer — May 2019 - Dec 2019

ACHIEVEMENTS

- **Invited Talk**: VT Skillshop Series: Leveraging Creative Technologies (Oct 2023)
- CCI SWVA Cyber Innovation Scholarship: 2024-2025
- CCI Research Showcase: 2024
- *The Dark Side of AI* - VPM News Focal Point: 2023
- *The Rise of the Chatbots* - Communications of the ACM: 2023
- *The strengths and limitations of approaches to detect deepfake text* - TechXplore: 2022

TECHNICAL SKILLS

- **GenAI Technologies**: LMMs/VLMs, LLMs, T2I models, LoRA, Model Fine-tuning
- **Languages**: Python, C/C++, Bash, Java, JavaScript, Assembly
- **Frameworks**: PyTorch, TensorFlow, Keras, Django
- **Developer Tools**: Git, Vim, Jupyter Notebook, VS Code, Markdown, LaTeX, Linux, Docker