Sifat Muhammad Abdullah

+1 (540)-449-2710 | sifat.abdullah577@gmail.com | https://sifatmd.github.io | Google Scholar

EDUCATION

PhD in Computer Science, Virginia Tech, advisor: Dr. Bimal Viswanath   **1/2021 - expected 12/2025**

BS in Computer Science and Engineering, BUET (CGPA: 3.91/4.0)   **2/2015 - 4/2019**

RESEARCH INTERESTS

Broad interest in the security of Generative AI and Machine Learning. Specifically, I study adversarial robustness of Multimodal LLMs & deepfake defenses using Foundation models & test-time reasoning. Also studied toxicity mitigation in LLMs, and Multi-LLM reasoning optimization with Debate.

PUBLICATIONS

- Antonio Guillen-Perez, Ashwin Ramesh Babu, Sahand Ghorbanpour, Avisek Naug, Vineet Gundecha, **Sifat Muhammad Abdullah**, Ricardo L. Gutierrez, Soumyendu Sarkar. "Sustainable Control of Geo-Distributed Datacenters by Distilling Numerical Experts into Adaptive LLM Agents" *NeurIPS MLForSys Workshop*, 2025.

- Aravind Cheruvu, Shravya Kanchi, **Sifat Muhammad Abdullah**, Nicholas Kong, Daphne Yao, Murtuza Jadliwala, Bimal Viswanath. "TuneShield: Mitigating Toxicity in Conversational AI while Fine-tuning on Untrusted Data" *ArXiv Preprint*, 2025.

- Shravya Kanchi, Neal Mangaokar, Aravind Cheruvu, **Sifat Muhammad Abdullah**, Shirin Nilizadeh, Atul Prakash, Bimal Viswanath. "Taming Data Challenges in ML-based Security Tasks: Lessons from Integrating Generative AI" *ArXiv Preprint*, 2025.

- **Sifat Muhammad Abdullah**, Aravind Cheruvu, Shravya Kanchi, Taejoong Chung, Peng Gao, Murtuza Jadliwala, Bimal Viswanath. "An Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape" *IEEE Symposium on Security and Privacy (S&P)*, 2024.

- Aravind Cheruvu(co-lead), Connor Weeks(co-lead), **Sifat Muhammad Abdullah**, Shravya Kanchi, Danfeng Yao, Bimal Viswanath. "A First Look at Toxicity Injection Attacks on Open-domain Chatbots" *Annual Computer Security Applications Conference (ACSAC)*, 2023.

- Jiameng Pu(co-lead), Zain Sarwar(co-lead), **Sifat Muhammad Abdullah**, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, Bimal Viswanath. "Deepfake Text Detection: Limitations and Opportunities" *IEEE Symposium on Security and Privacy (S&P)*, 2023.

- Md Ashiqur Rahman (co-lead), Abdullah Aman Tutul(co-lead), **Sifat Muhammad Abdullah**(co-lead), Md Shamsuzzoha Bayzid. "CHAPAO: Likelihood and hierarchical reference-based representation of biomolecular sequences and applications to compressing multiple sequence alignments" *PLOS ONE Journal*, 2022.

- Shadman Saqib Eusuf, Kazi Ashik Islam, Mohammed Eunus Ali, **Sifat Muhammad Abdullah**, Abdus Salam Azad. "A Web-Based System for Efficient Contact Tracing Query in a Large Spatio-Temporal Database" *Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*, 2020.

WORK EXPERIENCE

ML Research Associate Intern | Hewlett Packard Enterprise Labs   **5/2025 - 8/2025**

- RL policy distillation into LLMs by parameter efficient fine-tuning (PEFT) for spatio-temporal optimization of Geo-distributed Data Center (DC) cooling performance, outperforming rule-based controllers by up-to 24.3% in energy consumption, evaluated by customizing LLaMA 3.2 and Qwen 3 models.

- Studied application of Multi-LLM reasoning systems with multi-turn Debate for DC cooling optimization by deploying QwQ-32B, achieving 43.7% lower energy usage and improving scalability and explainability over single-LLM controllers.

Graduate Research Assistant | Virginia Tech, SecML Lab   1/2022 - 4/2025 | 8/2025 - Present

- Defending Multimodal LLMs, e.g., LLaMA, LLaVA, MiniGPT-4 against a suite of adversarial attacks using multimodal inference-time reasoning with OpenAI o3, QVQ-72B-Preview & Kimi-VL-A3B-Thinking, and generative AI strategies including Diffusion models (e.g., FLUX, Stable Diffusion) and Autoregressive models (e.g., OpenAI GPT-4o), achieving more than 98% CLIPScore gain in image captioning in one representative scenario, among several cases evaluated.

- Evaluated robustness of 8 defensive protection schemes, e.g., watermarking, preventing deepfake manipulation & style mimicry, achieved up-to 100% attack success while preserving image utility, using generative AI (GenAI) based image translation approach.

- Developed defense framework *TuneShield* for toxicity mitigation during LLM fine-tuning on untrusted data, utilizing safety alignment with Direct Preference Optimization (DPO), outperforming industry APIs by upto 28.4% by evaluating 7 LLMs from 4 model families, including LLaMA, FLAN-T5, OPT-IML and Vicuna.

- Analyzed robustness of 8 state-of-the-art deepfake image detectors by developing practical & low-cost adversarial attacks, achieving more than 70% performance (recall score) degradation, using Stable Diffusion and StyleGAN-based text-to-image (T2I) generators with LoRA fine-tuning and OpenCLIP multimodal foundation model. — *Resources requested by **40 research groups**.*

- Performed toxicity injection attacks on BART and BlenderBot chatbots after deployment in a Dialog-based learning setup, eliciting up-to 60% response toxicity rate by building adversarial attacks using GPT-J model with prompt engineering.

- Developed Nimai, a generative AI pipeline enabling highly controlled data synthesis, which improves security classifier accuracy by 32.6% (even in data constrained settings), leveraging discrete latent space of VAE-based architecture.

- Evaluated state-of-the-art deepfake text detectors, e.g., BERT & GPT-2 based defenses, on our collected real-world datasets, and achieved up-to 91.3% evasion rate by crafting high-probability token replacement using public LLMs without any query to surrogate or victim defenses. — *Resources requested by **158 research groups**.*

Graduate Research Assistant | BUET, DataLab   1/2020 - 12/2020

- Developed highly efficient web-based contact tracing query system to locate COVID-19 patients utilizing QzR-tree with PostgreSQL database.

Software Engineer | REVE Systems, Dhaka, Bangladesh   5/2019 - 12/2019

- Built a chatbot system for company website using BART with PyTorch and Django framework.

TEACHING EXPERIENCE

Graduate Teaching Assistant | Intro to Python & Java   1/2021 - 12/2021

- Conducted office hours, programming labs, and graded assignments for undergraduate courses.

PROFESSIONAL SERVICE

Technical Program Committees

- Deepfake, Deception, and Disinformation Security Workshop (3D-Sec), 2025

- 4th Workshop on the Security Implications of Deepfakes and Cheapfakes (WDC), 2025

Reviewer for Journals

- IEEE Transactions on Information Forensics and Security (IEEE TIFS), 2025

- Pervasive and Mobile Computing (PMC) Journal, 2025

ACHIEVEMENTS

- Pratt Fellowship, CS@VT   **2025**
- CCI SWVA Cyber Innovation Scholarship   **2024 - 2025**
- CCI Research Showcase   **6/2024**
- Invited Talk: VT Skillshop Series: Leveraging Creative Technologies   **10/2023**
- CCI Student Spotlight   **2023**
- BUET Dean's List Award   **2015 - 2019**

MEDIA COVERAGE

- *The Dark Side of AI* - VPM News Focal Point   **10/2023**
- *The Rise of the Chatbots* - Communications of the ACM   **7/2023**
- *The strengths and limitations of approaches to detect deepfake text* - TechXplore   **11/2022**

SKILLS

- **GenAI Technologies**:   MLLMs/VLMs, LLMs, T2I models, LoRA, PEFT, SFT, DPO
- **Languages**:   Python, C/C++, Bash, Java, JavaScript, Assembly
- **Frameworks**:   PyTorch, TensorFlow, Keras, Django
- **Libraries**:   Scikit-learn, NumPy, pandas, vLLM, transformers, peft, trl
- **Developer Tools**:   Git, Vim, Jupyter Notebook, VS Code, Cursor, Markdown, LaTeX, Linux, Docker

REFERENCES

- **Bimal Viswanath**, Associate Professor, Department of Computer Science, Virginia Tech.
- **Peng Gao**, Assistant Professor, Department of Computer Science, Virginia Tech.
- **Murtuza Jadliwala**, Associate Professor, Department of Computer Science, UT San Antonio.