Huzaifa Arif

(518) 961-8482 | huzaifaarif20@gmail.com | Troy, NY | huzaifa-arif.github.io

PROFILE SUMMARY

A fifth-year Ph.D. candidate and researcher with multiple first-author publications in trustworthy AI. My work focuses on exposing and mitigating privacy and safety vulnerabilities in Large Language Models, demonstrated through research at IBM and LLNL. **I am seeking a full-time Research Scientist position starting December 2025** where I can apply my expertise in LLM alignment and privacy to build verifiably safe AI systems.

EDUCATION

Ph.D. in Electrical and Computer Systems Engineering **Expected Dec 2025**
Rensselaer Polytechnic Institute | Advisor: Alex Gittens
B.S. in Electrical Engineering **Graduated with High Merit**
Lahore University of Management Sciences (LUMS) | Dean's Honor List

RESEARCH EXPERIENCE

IBM T.J. Watson Research Center | *AI Research Extern - Trustworthy AI* **May 2025 – Aug 2025**

- Developed parameter-efficient steering method for LLM safety using lightweight prefix tuning (<0.01% parameters)
- Combined SFT and DPO training to achieve 80% reduction in toxic responses across Llama and Aya model families
- Extending framework to mitigate demographic bias and PII leakage (manuscript in preparation for ICLR 2026)

Rensselaer Polytechnic Institute | *Research Assistant* **Aug 2022 – Present**

- Discovered novel "association leakage" vulnerability in LLMs where jail-broken models reveal sensitive data
- Created SPARK attack method achieving 3x increase in private data recall using only 0.01% trainable parameters (manuscript in preparation for ICLR 2026)
- Published work on federated fair learning via kernel regularization in TMLR 2025

Lawrence Livermore National Laboratory | *Data Science Intern* **May 2024 – Aug 2024**

- Exposed critical vulnerabilities in FourCastNet weather prediction model through novel adversarial attacks
- Demonstrated susceptibility to localized evasion attacks leading to faulty predictions (Accepted at AAAI 2025)

IBM T.J. Watson Research Center | *AI Research Extern* **Jun 2023 – Aug 2023**

- Developed PEEL method for backward feature inversion in residual networks, achieving 10x MSE improvement
- Published findings at IEEE SATML 2025; filed patent currently under review

IBM T.J. Watson Research Center | *AI Research Extern* **Jun 2022 – Aug 2022**

- Created Reprogrammable-FL framework adapting model reprogramming to differentially private federated learning
- Achieved 60% accuracy improvement under same privacy budget (Published at IEEE SATML 2023)
- Patent filed (US20240256894A1); authored book chapter in "Federated Learning for Medical Imaging"

SELECTED PUBLICATIONS

- **H. Arif**, A. Gittens, & P.-Y. Chen. "Reprogrammable-FL: Improving Utility-Privacy Tradeoff in Federated Learning via Model Reprogramming." In *IEEE Conference on Secure and Trustworthy Machine Learning (SATML)*, 2023.
- **H. Arif**, A. Gittens, K. Murugesan, P. Das, & P.-Y. Chen. "Peel the Layers and Find Yourself: Revisiting Inference-time Data Leakage for Residual Neural Networks." In *IEEE Conference on Secure and Trustworthy Machine Learning (SATML)*, 2025.
- **H. Arif**, A. Gittens, J. Diffenderfer, P.-Y. Chen, & B. Kailkhura. "Forecasting Fails: Unveiling Evasion Attacks in Weather Prediction Models." In *AAAI Workshop on Artificial Intelligence for Science and Engineering (AI2SE)*, 2025.
- **H. Arif**, P.-Y. Chen, K. Murugesan, & A. Gittens. "Fair Federated Learning via Stochastic Kernel Regularization." *Transactions on Machine Learning Research (TMLR)*, 2025. (Journal)
- **H. Arif**, A. Gittens, K. Murugesan, I. Ko, P. Das, & P.-Y. Chen. "The Safety Patch: Lightweight Prefix Tuning with DPO for Controllable LLMs." (In Preparation: ICLR 2026)
- **H. Arif**, A. Gittens, K. Murugesan, P. Das, & P.-Y. Chen. "SPARK: Amplifying Association Leakage in LLMs via Soft Prompting and Attention Head Steering." (In Preparation: ICLR 2026)

PATENTS and INTELLECTUAL PROPERTY

- US20240256894A1: "Differentially Private Federated Learning Using Model Reprogramming" (Under Review)
- Patent Application: "Method for Quantifying Private Leakage in Pretrained Neural Networks" (Filed)
- **Book Chapter**:Utility Privacy Tradeoff in the book "*Federated Learning for Medical Imaging*"

HONORS & AWARDS

- Belsky Award for Computational Science and Engineering 2025 (Graduate Research Excellence)
- Top-5 Candidate in ECSE Research Qualifier Exam at RPI
- Travel Support Awards: IEEE SATML 2023 & 2025

Technical Skills

Programming & Languages: Python, C++, SQL, MATLAB
AI/ML Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn
Data Science Libraries: NumPy, Pandas, SciPy, Matplotlib, Seaborn
Research Areas: Trustworthy AI, Large Language Models, Federated Learning, Differential Privacy, Reinforcement Learning, Constrained Optimization, Stochastic Optimization, Transfer Learning, Spatio-temporal Data Analysis, NLP, Deep Learning, Alignment, RLHF, Prompt Tuning, Soft Prompt Engineering, Model Reprogramming
Tools & Other: Git/GitHub, Jupyter Notebooks, LaTeX, Tableau

SERVICE

Reviewer: ICLR 2025, ICASSP 2023/2025, AISTATS 2023, IEEE MLSP 2023
Teaching Assistant (4.5/5 avg rating): Machine Learning, Embedded Control, Signals & Systems (2021-2024)