Huzaifa Arif

☎ (518) 961-8482 | ✉ huzaifaarif20@gmail.com | ✉ arifh@rpi.edu | 🌐 Website | 🎓 Google Scholar

PROFILE SUMMARY

Ph.D. candidate in Electrical and Computer Systems Engineering at Rensselaer Polytechnic Institute specializing in trustworthy AI, with six first-author publications in privacy-preserving machine learning and LLM safety. My research focuses on developing methods for **federated learning** and **LLM** that enhance **privacy**, promote group **fairness**, and improve **robustness** to **adversarial attacks**, with applied experience at IBM Research and Lawrence Livermore National Laboratory.

EDUCATION

Ph.D. in Electrical and Computer Systems Engineering **2021 - Ongoing**
Rensselaer Polytechnic Institute | Advisor: Alex Gittens
B.S. in Electrical Engineering 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**

- Designed *policy patching*, an enhanced design of prefix tuning that enterprises can ship as a low-cost hotfix to already-deployed LLMs between major releases.
- Achieved safety performance comparable to full detoxification fine-tuning while using ≈ 800× fewer trainable parameters and ≈ 56× less GPU time than SOTA methods for fixing safety policy.
- Demonstrated up to ≈ 70% toxicity reduction with only 1–2k training examples per patch across multiple open-source LLM families (e.g., Llama-2/3, Aya-23, Mistral, Gemma) with minimal (≈ 2.5%) inference overhead. (Work under Review)

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

- **Project 1)** Developed a prompt-based PII leakage auditing tool that stress tests association leakage vulnerabilities in LLMs utilizing internal attention patterns, revealing how jailbroken models expose sensitive data
- This method achieves 3× increase in private data recall using only 0.003% trainable parameters (Work under review)
- **Project 2)** Designed FedProxGrad, a novel kernel-based federated learning algorithm with theoretical guarentees that reduces computational overhead while achieving *group fairness* across distributed classification and regression tasks (published **in TMLR 2025**)

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

- Discovered critical vulnerabilities in FourCastNet weather prediction model by developing novel evasion-based adversarial attacks, revealing security gaps prior to production deployment.
- Demonstrated systematic susceptibility to localized perturbations that compromise prediction accuracy, establishing new threat models for ML-based weather forecasting systems. (**Accepted at AAAI 2025**)

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

- Developed PEEL - a method for backward feature inversion in residual networks, achieving 10x MSE improvement over generative methods
- Published findings at **IEEE SATML 2025 (29% Acceptance Rate)**; filed patent currently under review

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

- Created Reprogrammable-FL framework adapting model reprogramming to differentially private federated learning
- Achieved 60% accuracy improvement under same privacy budget against other finetuning methods in federated learning (**Published at IEEE SATML 2023** 26% Acceptance Rate)
- Patent filed (US20240256894A1); authored book chapter in "Federated Learning for Medical Imaging"

SELECTED PUBLICATIONS – ALL FIRST AUTHOR

- **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) (26% Acceptance Rate)*** , 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) (29% Acceptance Rate)*** , 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. "Patching LLM Like Software: A Lightweight Method for Improving Safety Policy in Large Language Models." (Under Review)
- **H. Arif**, A. Gittens, K. Murugesan, P. Das, & P.-Y. Chen. "SPARK: Amplifying Association Leakage in LLMs via Soft Prompting and Attention Head Steering." (Under Review)

PATENTS

- US20240256894A1: "Differentially Private Federated Learning Using Model Reprogramming"
- Patent Application: "Method for Quantifying Private Leakage in Pretrained Neural Networks" (Filed)
- Patent Application: "Patching: A Lightweight Method for Mitigating Safety Risks in Large Language Models" (Filed)

BOOK CHAPTER

Contributed a chapter (first author) called "Utility Privacy Tradeoff" in Federated Learning for Medical Imaging

HONORS & AWARDS (Graduate School)

- Founders of Excellence 2025 (Top 1% of Students for Excellence in Research and Overall Leadership)
- 2nd Place Runner-Up, 3MT (Three Minute Thesis) 2025, RPI
- 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, JAX(basic); TensorFlow; NumPy/Pandas/SciPy; Matplotlib
Privacy/Alignment Tools: Differential Privacy; **Opacus**, **TF-Privacy**; DP-SGD; privacy auditing/evaluation; prompt/prefix/soft-prompt steering; jailbreaking/red-teaming; safety classifiers.
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 2025/6,ICASSP 2023/2025, AISTATS 2023, IEEE MLSP 2023
Teaching Assistant (4.5/5 avg rating): Machine Learning, Embedded Control, Signals & Systems — 8 semesters total (2021–2025)