

Toluwani Samuel Aremu

AI Safety | Trustworthy AI | Responsible AI

[Certifications](#) • [Email](#) • [GitHub](#) • [Google Scholar](#) • [LinkedIn](#) • [Website](#)

KEY COMPETENCIES

- **Skills:** Research, Mathematics, Statistics, Machine Learning, Deep Learning, Data Science, Project Management, Programming, Writing.
- **Tools:** Python, VB.Net, PyTorch, Lightning, TensorFlow, Keras, Jax, Scikit-learn, NumPy, Matplotlib, Visual Studio, Visual Studio Code, PyCharm, Jupyter, LaTeX, Office 365, Google [Docs, Sheets, Slides].
- **Research Interests:** Generative AI, Watermarking, Alignment (Self/Agentic), Scalable Oversight, Debate.

EDUCATION

Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE <ul style="list-style-type: none">• Doctor of Philosophy (PhD) in Machine Learning.• Research Areas: AI Safety	AUG 2023 – PRESENT
Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE <ul style="list-style-type: none">• Master of Science (MSc) in Machine Learning.• Research Area: Privacy-Preserving ML.	JAN 2021 – DEC 2022
University of Ibadan (UI), Nigeria <ul style="list-style-type: none">• Master of Science (MSc) in Computer Science.• Research Area: Cryptography.	MAY 2018 – MAY 2020
Adeleke University, Nigeria <ul style="list-style-type: none">• Bachelor of Science (BSc) in Computer Science.• Minor: Philosophy.	OCT 2012 – JUL 2016

RESEARCH EXPERIENCE

Doctoral Researcher, Trustworthy-ML Lab, MBZUAI, UAE <ul style="list-style-type: none">• Research Area: Safe and Trustworthy Generative AI.• Selected Publications (* denotes equal contribution):<ul style="list-style-type: none">◦ S. Fares, K. Ziu, T. Aremu*,..., "MirrorCheck: Efficient Adversarial Defense for Vision-Language Models," arXiv, 2024. (under review at TMLR)◦ A. Diaa, T. Aremu, & N. Lukas, "Optimizing Adaptive Attacks against Content Watermarks for Language Models," arXiv, 2024. (under review at ICML '25, accepted at ICLR-WMARK '25)◦ N. Tasthan, S. Fares, T. Aremu, S. Horvath, and K. Nandakumar, "Redefining Contributions: Shapley-Driven Federated Learning," 33rd International Joint Conference on Artificial Intelligence (IJCAI), Jeju, Korea, 2024.	AUG 2023 – PRESENT
Research Assistant, MCR-Lab, MBZUAI, UAE <ul style="list-style-type: none">• Research Area: AI Applications in Smart Cities.• Selected Publications:<ul style="list-style-type: none">◦ T. Aremu et al., "On the reliability of Large Language Models to misinformed and demographically informed prompts," (AAAI) AI Magazine, vol. 46, no. 1, 2025.◦ T. Aremu, L. Zhiyuan, R. Alameeri, M. Khan, and A.E. Saddik, "SSIVD-Net: A Novel Salient Super Image Classification and Detection Technique for Weaponized Violence," Lecture Notes in Networks and Systems, pp. 16–35, Jan. 2024.◦ W. Y. Kang, T. Aremu, Y. Balah, M. Nadeem, I. G. Navarette, and A. E. Saddik, "ScholarFace: Scanning Faces, Discovering Minds," 2024 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–5, Jan. 2024. (US Patent App. 18/672,708)	FEB 2023 – AUG 2023
Applied Science Intern, M42 HealthCare, UAE <ul style="list-style-type: none">• Achievements:<ul style="list-style-type: none">◦ Developed an end-to-end pipeline for efficiently downloading and preprocessing the NHANES dataset, ensuring streamlined data preparation.◦ Integrated AutoML capabilities to automate analysis, training, and evaluation of statistical models while allowing flexibility for custom configurations.◦ Engineered a feature which generates a comprehensive TRIPOD report post-evaluation, providing structured insights for model assessment and transparency.	FEB 2023 – APR 2023
Student Researcher, SPriNT-AI Lab, MBZUAI, UAE	JAN 2021 – DEC 2022

- **Research Area:** Privacy-Preserving ML.
- **Publication:**
 - **T. Aremu** and K. Nandakumar, "PolyKervNets: Activation-free Neural Networks For Efficient Private Inference," 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), Raleigh, NC, USA, 2023.

OTHER EXPERIENCE

Projects

- Accented Speech Recognition AUG 2021 – DEC 2021
 - Implemented VQ-VAE to disentangle style and content features in the latent space of ASR systems.
 - Accuracy on accented speech improved by 3.8%.
- Racial Bias Mitigation in Self-Supervised Face Recognition Architecture JAN 2021 – MAY 2021
 - Implemented various preprocessing and model-centric methods such as downsampling, GANs, weighted sampling, etc, to reduce racial biases in detecting faces.
 - Improved face recognition of people of color in SIM-CLR by up to 20%. However, face recognition for well represented races in the dataset reduced by up to 5%.
- Gender Bias Mitigation in Word Embeddings JAN 2021 – MAY 2021
 - Investigated several data-centric methods proposed to mitigate gender bias in GloVe.
 - Concluded that most of these methods often lead to new forms of biases in NLP systems they are used in.

Teaching

- Teaching Assistant, Mathematical Foundations of AI (MTH701), **MBZUAI**, UAE AUG 2024 – DEC 2024
- Teaching Assistant, Object Oriented Programming (OOP), **University of Ibadan**, Nigeria JAN 2019 – MAY 2019

Leadership

- Graduate School Mentorship for Africans JAN 2021 – PRESENT
- Associate Editor, MBZUAI Research Blog JAN 2024 – PRESENT

Reviewing

- **Conferences:** JAN 2024 – PRESENT
 - [NeurIPS](#) | [ICLR](#) | [ICML](#) | [AISTATS](#) | [AAAI](#) | [DLI](#)
- **Workshops:** AUG 2024 – PRESENT
 - [HRAIM@NeurIPS](#) | [SafeGenAI@NeurIPS](#) | [WMark@ICLR](#)
- **Journals:** AUG 2023 – PRESENT
 - [IEEE Access](#) | [QSI](#) | [CHBAH](#) | [AI Magazine](#)

Talks/Presentations

- "Literature Review and Research Methodologies", LyngualLabs MAY 2025
- "Optimizing Adaptive Attacks against Content Watermarks for Language Models", ICLR-WMARK APR 2025
- "Ethical Perspectives of AI", AI Summer, Department of Material Sciences, University of Denver JUL 2022

HONORS & AWARDS

- MBZUAI Conference Travel Scholarship APR 2025
- MBZUAI PhD Fully Funded Fellowship AUG 2023
- MBZUAI MSc Fully Funded Fellowship JAN 2021
- UAE Golden Visa for Talented Persons/Specialists in Science OCT 2022
- MBZUAI Award of Appreciation for Iconic Representation and Student Hospitality AUG 2022
- ProjectSet Innovation Challenge for Entrepreneurship (ICE-22) MAY 2022
- Top 100, DeepLearning.AI Data-Centric AI Competition AUG 2021
- NYSC-FRSC Award for the Most Creative Corp Member OCT 2017
- 2015 AUE-NACOSS Award for the Best Programmer JUL 2015

REFERENCES

- [Dr. Nils Lukas](#) - Assistant Professor, MBZUAI - nils.lukas@mbzuai.ac.ae.
- [Dr. Karthik Nandakumar](#) - Associate Professor, MBZUAI - karthik.nandakumar@mbzuai.ac.ae.
- [Prof. Kun Zhang](#) - Professor, MBZUAI & Carnegie Mellon University - kun.zhang@mbzuai.ac.ae.
- [Prof. Abdulmotaleb El Saddik](#) - Distinguished Professor, UOttawa - elsaddik@ottawa.ca.