

Toluwani Samuel Aremu

AI Safety Researcher | Applied Scientist | ML Engineer

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

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:** AI Safety, Trustworthy AI, Responsible AI.

EDUCATION

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

RESEARCH EXPERIENCE

- | | |
|---|---------------------|
| PhD Student, MBZUAI, UAE | AUG 2023 – PRESENT |
| <ul style="list-style-type: none">• Research Areas: Safe and Trustworthy Generative AI.• 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.○ A. Diaa, T. Aremu, & N. Lukas, "Optimizing Adaptive Attacks against Content Watermarks for Language Models," arXiv, 2024.○ 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.○ N. Tastan, 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.○ M. Nwadike, Z. Iklassov, T. Aremu,..., "RECALL: Library-Like Behavior In Language Models is Enhanced by Self-Referencing Causal Cycles," arXiv, 2025. | |
| Research Assistant, MBZUAI, UAE | FEB 2023 – AUG 2023 |
| <ul style="list-style-type: none">• Research Areas: AI Applications.• Publications:<ul style="list-style-type: none">○ 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.○ T. Aremu, "Unlocking Pandora's Box: Unveiling the Elusive Realm of AI Text Detection," Social Science Research Network, Jun. 06, 2023.○ 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. | |
| Applied Science Intern, M42 HealthCare, UAE | FEB 2023 – APR 2023 |
| <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.

MSc Student/Graduate Research Assistant, MBZUAI, UAE

JAN 2021 – DEC 2022

- **Research Areas:** Privacy-Preserving ML.
- **Publications:**
 - **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](#) | [OSI](#) | [CHBAH](#) | [AI Magazine](#)

Invited Talks

- "Ethical Perspectives of AI", AI Summer, Department of Material Sciences, University of Denver JUL 2022

HONORS & AWARDS

- MBZUAI MSc & PhD Fully Funded Fellowship JAN 2021 – MAY 2027
- 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.