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, PyTorch, Lightning, TensorFlow, Keras, Jax, Scikit-learn, NumPy, Matplotlib, Visual Studio Code, PyCharm, Jupyter, etc.
- **Research Interests**: AI Safety, AI Security, Trustworthy AI, Responsible AI.

EDUCATION

- Doctor of Philosophy (PhD) in Machine Learning, **Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE.** *(AUG 2023 – PRESENT)*
- Master of Science (MSc) in Machine Learning, **Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE.** *(JAN 2021 – DEC 2022)*
- Master of Science (MSc) in Computer Science, **University of Ibadan (UI), Nigeria.** *(MAY 2018 – MAY 2020)*
- Bachelor of Science (BSc) in Computer Science, **Adeleke University, Nigeria.** *(OCT 2012 – JUL 2016)*

RESEARCH EXPERIENCE

- **Doctoral Researcher, Trustworthy-ML Lab, MBZUAI.** *(AUG 2023 – PRESENT)*
    - ***Research Areas****: Generative AI Security, Watermarking, Alignment, Adversarial Robustness, Policy.*
- **Research Assistant, MCR-Lab, MBZUAI.** *(FEB 2023 – AUG 2023)*
    - ***Research Areas****: Generative AI, AI in Smart Cities, AI Applications.*
- **Graduate Student Researcher, SPriNT-AI Lab, MBZUAI.** *(FEB 2023 – AUG 2023)*
    - ***Research Areas****: Homomorphic Encryption, Secure Multi-Party Computation, Private Inference, Privacy-Preserving ML.*

SELECTED PUBLICATIONS **(* denotes equal contribution)**

- **Aremu, T.**, Hussein, N., Nwadike, M., Poppi, S., Zhang, J., Nandakumar, K., ... & Lukas, N. (2025). Mitigating Watermark Stealing Attacks in Generative Models via Multi-Key Watermarking. arXiv preprint arXiv:2507.07871.
- Diaa, A., **Aremu, T.**, & Lukas, N. (2025). Optimizing adaptive attacks against content watermarks for language models. *The Forty-second International Conference on Machine Learning (ICML)**[Spotlight]**.*
- **Aremu, T.**, Akinwehinmi, O., Nwagu, C., Ahmed, S., Orji, R., Amo, P., & Saddik, A. (2025). On the reliability of Large Language Models to misinformed and demographically informed prompts. *AI Magazine (AAAI), 46(1), e12208.*
- Nwadike, M., Iklassov, Z., **Aremu, T.**, Hiraoka, T., Bojkovic, V., Heinzerling, B., Alqaubeh, H., Takač, M., & Inui, K. (2025). RECALL: Library-Like Behavior In Language Models is Enhanced by Self-Referencing Causal Cycles. *The 63rd Annual Meeting of the Association for Computational Linguistics (ACL).*
- Fares, S.*, Ziu, K.*, **Aremu, T.***, Durasov, N., Takač, M., Fua, P., Nandakumar, K., & Laptev, I. (2024). Mirrorcheck: Efficient adversarial defense for vision-language models. *arXiv preprint arXiv:2406.09250.*
- Tastan, N., Fares, S., **Aremu, T.**, Horvath, S., & Nandakumar, K. (2024). Redefining contributions: shapley-driven federated learning. *In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (pp. 5009-5017).*
- **Aremu, T.**, & Nandakumar, K. (2023). Polykervnets: Activation-free neural networks for efficient private inference. *In 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) (pp. 593-604). IEEE.*

WORK EXPERIENCE

- **Applied Scientist (Intern), G42 (M42 HealthCare), UAE.** *(FEB 2023 – APR 2023)*
    - 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.
- **Data Analyst, State Government of Osun, Nigeria.** *(JAN 2018 – DEC 2018)*
    - Carried out exploratory data analysis (EDA) on survey data gotten from the State's workers.
    - Used insights from the analysis helped in the Government's decision-making processes, improving workers' satisfaction and increasing the half year internal revenue by 27.7%.
- **IT Support Engineer, Government of Ekiti State, Nigeria.** *(NOV 2016 – NOV 2017)*
    - Installed, configured, and maintained softwares, computer systems and workstations for the State's Ministry of Finance.

OTHER EXPERIENCE

- **Projects**
    - **Accent Speech Recognition, MBZUAI.** *(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, MBZUAI.** *(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%.
    - **Gender Bias Mitigation in Word Embeddings, MBZUAI.** *(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.** *(AUG 2024 – DEC 2024)*
  - Teaching Assistant, Object Oriented Programming (OOP), **University of Ibadan.** *(JAN 2019 – MAY 2019)*
- **Talks/Presentations**
  - "Literature Review and Research Methodologies", presented at **LyngualLabs, Nigeria.** *(MAY 2025)*
  - "Optimizing Adaptive Attacks against Content Watermarks for Language Models", presented at **ICLR-WMARK, Singapore.** *(APR 2025)*
  - "SSIVD-Net: A Novel Salient Super Image Classification and Detection Technique", presented at **SAI, UK.** *(JUN 2024)*
  - "PolyKervNets: Activation-free Neural Networks For Efficient Private Inference", presented at **IEEE SaTML, USA.** *(FEB 2023)*
  - "Ethical Perspectives of AI", presented at the **Department of Material Sciences, University of Denver, USA.** *(JUL 2022)*
- **Reviewing**
  - **Conferences**: NeurIPS | ICLR | ICML | AISTATS | AAAI | DLI
  - **Workshops**: HRAIM@NeurIPS | SafeGenAI@NeurIPS | WMark@ICLR
  - **Journals**: IEEE Access | OSJ | CHBAH | AI Magazine
- **Leadership/Volunteering**
  - Volunteering Member, "**EVOLVE-MBZUAI".** *(AUG 2024 – PRESENT)*
  - Co-founding Member, "**Spruce Up Your Space".** *(AUG 2024 – PRESENT)*
  - Associate Editor, **ML Department Research Blog.** *(JAN 2024 – PRESENT)*
  - Student Editor, **MBZUAI Research Blog.** *(JAN 2024 – PRESENT)*
  - Mentor, **AI Research/Graduate School Mentorship for Africans.** *(JAN 2021 – PRESENT)*

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, University of Michigan | 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*