Eric Nartey Onyame

website: eric-onyame.github.io ◇ reh6ed@virginia.edu ◇ ericnarteyonyame@gmail.com

My research advances trustworthy AI at the intersection of AI safety, interpretability, and multilingual AI. I develop methods for monitoring large language models, with a focus on making chain-of-thought monitoring more reliable for scalable oversight. My broader goal is to help build AI systems that are safe, trustworthy, and aligned with human values.

Education

University of Virginia — *Aug 2023 – Expected May 2028*
Ph.D. in Data Science — Charlottesville, VA
Advised by Dr. Chirag Agarwal — **GPA:** 3.95/4.0

The University of Tennessee at Chattanooga — *Aug 2021 – May 2023*
M.Sc. in Mathematics (Applied Mathematics) — Chattanooga, TN
Advised by Dr. Lakmali Weerasena — **GPA:** 3.80/4.0

University of Cape Coast — *Aug 2016 – Jul 2020*
B.Sc. in Mathematics with Economics — Cape Coast, Ghana
GPA: 3.76/4.0

Publications

* denotes equal contribution. Google Scholar.

1. **Eric Onyame***, Akash Ghosh*, Subhadip Baidya, Sriparna Saha, Xiuying Chen, Chirag Agarwal. *CURE-Med: Curriculum-Informed Reinforcement Learning for Multilingual Medical Reasoning.* In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, Main Conference, Oral, Top 50%. arXiv.

2. **Eric Onyame**, Miles Zhou, Kowshik Thopalli, Bhavya Kailkhura, Chirag Agarwal. *The Fragility of Chain-of-Thought Monitoring Across Typologically Diverse Languages.* arXiv.

3. **Eric Nartey Onyame**. *Covering Problem with Minimum Radius Enclosing Circle.* Master's thesis, Department of Mathematics, The University of Tennessee at Chattanooga, 2023. Thesis link. *712 downloads.*

Ongoing Projects

3. *White-Box Approaches for Chain-of-Thought Monitorability.*
Eric Onyame, Miles Zhou, Bhavya Kailkhura, Kowshik Thopalli, Chirag Agarwal. *In progress.*

4. *Counterfactual LLM Verifiers for Math and Logic Reasoning Tasks.*
Elita Lobo, **Eric Onyame**, Yair Zick, Chirag Agarwal. *In progress.*

Research Interests

1. **Chain-of-Thought Monitoring and Scalable Oversight:** Methods for monitoring model reasoning to detect unfaithful, deceptive, and scheming behavior in large language models.
2. **AI Safety and Interpretability:** White-box and interpretability-based approaches for understanding, evaluating, and improving the reliability of advanced AI systems.
3. **Multilingual AI under Distribution Shift:** Evaluation and training methods for studying how model behavior changes across languages, tasks, and contexts.

Awards and Honors

1. Quantitative Foundation Fellow, University of Virginia, 2023–2024.
2. Provost's Fellowship, University of Virginia, 2023–2024.
3. Best Mathematics Graduate Student Award, The University of Tennessee at Chattanooga, April 2023.
4. Inductee, Pi Mu Epsilon Mathematics Honor Society, The University of Tennessee at Chattanooga, April 2022.

Teaching

1. Deep Learning, Teaching Assistant
   University of Virginia Spring 2026
2. Research Methods in Data Science, Teaching Assistant
   University of Virginia Fall 2025
3. Decoding Large Language Models, Teaching Assistant
   University of Virginia Spring 2025
4. Data Engineering I: Data Management & Visualization, Teaching Assistant
   University of Virginia Fall 2024
5. Foundations of Data Science, Teaching Assistant
   University of Virginia Fall 2024
6. Elementary Statistical Analysis; College Algebra, Teaching Assistant
   The University of Tennessee at Chattanooga Aug 2021 – Jun 2022

Specialties

- **Trustworthy AI and Safety Evaluation:** Experience in AI reliability and safety evaluation, adversarial stress testing, chain-of-thought monitoring, and interpretable error analysis across multilingual settings.
- **RL for Reasoning and Alignment:** RL and preference optimization workflows for improving LLM reasoning quality, robustness, and safe behavior.
- **Multilingual AI:** Experience improving model reasoning, safety, and generalization across diverse languages, including low-resource settings.
- **LLM Training and Fine-Tuning:** PyTorch-based fine-tuning, data curation, scalable training, and systematic evaluation.

- **Statistical Modeling and Data Analysis:** Skilled in big data analysis, predictive modeling, and statistical inference using Python, with strong foundations in probability and linear algebra.
- **Research Engineering:** Reproducible experimentation, ablations, documentation, and large-scale runs on HPC.

Technical Skills

- **Languages:** Python, R, Julia, JavaScript, SQL.
- **ML and Data:** PyTorch, scikit-learn, NumPy, Pandas, Matplotlib, PySpark.
- **Engineering and Tooling:** Git, Docker, LaTeX, Cloud computing.

Selected Talks and Presentations

1. Applications of Large Language Models
   School of Data Science, University of Virginia 2025
2. Mathematical Model and Algorithm to Identify a Hub Location
   UTC Spring Research and Arts Conference, The University of Tennessee at Chattanooga 2023
3. Covering Problem with Minimum Radius Enclosing Circle
   Master's Thesis Defense, The University of Tennessee at Chattanooga 2023
4. Application of the Smallest Enclosing Circle Problem to Healthcare Systems
   American Mathematical Society Southeastern Sectional Meeting 2022
5. Carotenoid Levels and Self-reported Fruit and Vegetable Intake: Differences Across Age, Race, Sex, and BMI
   National and Regional Food Industries (Student Research Presentation) 2022

Service

1. Reviewer, NeurIPS
   Conference on Neural Information Processing Systems 2026
2. Mathematics department research Committee Member
   University of Tennessee at Chattanooga 2021-2023