Isaac Kobby Anni

Ph.D. Candidate in Data Science | Machine Learning Researcher

Bowling Green, Ohio | isaacka@bgsu.edu | linkedin.com/in/isaac-kobby-anni | isaackobby.com

Research Profile

Ph.D. candidate studying when machine-learning actions generated by language models, such as new features, data-cleaning steps, and pipeline changes, can be trusted. My dissertation develops an evaluation framework that runs generated actions under controlled, leakage-safe conditions, measures whether they add predictive value that holds up across evaluators, data splits, and held-out data, and turns that evidence into clear decisions about which actions to keep, repair, or reject.

Education

Bowling Green State University *Dec. 2027*
Ph.D. in Data Science Bowling Green, Ohio
Dissertation: Trustworthy Evaluation of Language-Model-Driven AutoML Actions and Pipelines
Advisor: Dr. Md Main Uddin Rony

Bowling Green State University *May 2024*
M.S. in Data Science Bowling Green, Ohio

African Institute for Mathematical Sciences *July 2022*
M.S. in Mathematical Sciences (focus: Big Data) Dakar, Senegal
Thesis: Structured Approach to Credit Risk Modeling

University for Development Studies *June 2019*
B.Sc. in Financial Mathematics Ghana

Research Experience

Research Assistant *Aug. 2026–Present*
Bowling Green State University Bowling Green, Ohio

- Conduct doctoral research with Dr. Md Main Uddin Rony on trustworthy human-AI collaboration and the evaluation of language-model-driven data-science actions.
- Lead study design, experiments, and manuscript writing for 2 ongoing projects; first author on one manuscript under review.
- Prepare manuscripts, presentations, and research materials for technical and interdisciplinary audiences.

Graduate Assistant and Data Scientist *Aug. 2022–Apr. 2026*
Student Success Analytics and Technologies, Bowling Green State University Bowling Green, Ohio

- Built Python, SQL, and Power BI pipelines for predictive modeling and reporting used by academic colleges, advising teams and student success leaders to support data-informed decisions across the university serving approximately 19,700 students.
- Developed and evaluated student-success prediction models on complex educational data, using cross-validation and focusing on minority-class performance.
- Automated 37 recurring reports, cutting preparation time by 65%.
- Maintained privacy-protecting, reproducible analysis workflows for sensitive student records.

Research Fellow in Natural Language Processing *Mar. 2022–Aug. 2022*
Computational Language Understanding (CLU) Lab, University of Arizona, and ICRISAT Remote

- Co-authored a peer-reviewed paper at WASSA 2023, an ACL workshop.
- Contributed to text annotation, information extraction, semantic similarity, and experimental evaluation in an international research collaboration.

Selected Research

Trustworthy Evaluation of Language-Model-Driven AutoML *2026–Present*
Dissertation Research

- Developed an evaluation framework that scores generated actions on six criteria: executability, incremental predictive value, stability, evaluator generality, robustness, and decision cost.
- Ran leakage-safe matched baseline-candidate experiments across 19 datasets using cross-validation, paired bootstrap analysis, and held-out evaluation; none of the 14 features that passed cross-validation gates were confirmed on held-out data, and 6 (43%) caused measurable harm.
- Built evidence-tracking and output-validation workflows that make each keep/repair/reject decision reproducible. Code available on request.

Language-Model-Driven Automated Feature Engineering *2024–Present*
Ph.D. Research

- Designed a workflow in which language models propose executable features that must pass controlled-execution, target-exclusion, leakage, and predictive-value checks.
- Evaluated the approach on 14 public classification datasets, with specialist components for numerical, temporal, and categorical features; improved TabPFN on 11 of 14 datasets (mean ROC-AUC gain +0.0038) and Random Forests on 12 of 14 (mean gain +0.0111).
- Presented as a poster at the 2026 Tapia Celebration; full manuscript under review.

Automated Data Cleaning and Leakage Detection *2026*

Research Project

- Built and tested a constrained proposal-and-validation pipeline for model-suggested cleaning actions, covering downstream evaluation, leakage prevention, and failure handling; detected train/eval duplicate leakage in 2 of 5 benchmark datasets (up to 20% eval-row overlap).

Publications and Presentations

Peer-Reviewed Publications

- **Anni, I. K.**, and Dasigi, V. G. (2025). Analysis of document representation for text classification. In K. Arai (Ed.), *Advances in Information and Communication* (FICC 2025), Lecture Notes in Networks and Systems, vol. 1285, pp. 261–273. Springer. 10.1007/978-3-031-84460-7_17
- Alexeeva, M., Hyland, C., Alcock, K., Beal Cohen, A. A., Kanyamahanga, H., **Anni, I. K.**, and Surdeanu, M. (2023). Annotating and training for population subjective views. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA)*, pp. 416–430. Association for Computational Linguistics. 10.18653/v1/2023.wassa-1.36

Presentations

- **Anni, I. K.** (2026). SLM-first automated feature engineering. Poster, CMD-IT/ACM Richard Tapia Celebration of Diversity in Computing, Atlanta, Georgia.

Awards and Honors

- Tapia 2026 Conference Scholarship (2026)
- Bowling Green State University Tuition Scholarship (2022-present)
- AIMS scholarship/fellowship (2020-2022)

Service

- Peer reviewer for AAAI, IEEE ICDM, and Africa AI at NeurIPS.

Technical and Research Methods

Languages: Python, SQL, R

ML and NLP: scikit-learn, PyTorch, TensorFlow, XGBoost, Hugging Face; language models, tabular machine learning, automated feature engineering, model interpretation

Methods: cross-validation, paired bootstrap evaluation, experimental design, distribution-shift evaluation, leakage detection, reproducible experimentation

Infrastructure: Git, Linux, SLURM, high-performance computing, Docker, pytest, Jupyter

References

Available on request.