TASSALLAH AMINA ABDULLAHI, PH.D.

she/her/hers ◇ Providence, RI
tassallah_abdullahi@brown.edu ◇ tassabdul.github.io

RESEARCH PROFILE

Researcher specializing in responsible AI, model behavior, and evaluation methodology. My work combines evaluation research, ML engineering, large-scale expert human studies, and adversarial testing to characterize safety, reliability, bias, and behavioral variation in generative models. I develop evaluation methods and benchmarks across multilingual, multimodal, and tool-using systems, connecting model behavior to the people and contexts affected by AI deployment.

Core Interests: Responsible AI, Sociotechnical Evaluation, Model Behavior, AI Safety, Evaluation Methodology, Human–AI Judgment, Adversarial Robustness.

RESEARCH AND INDUSTRY EXPERIENCE

Humana Remote, USA
Senior Responsible AI Data Scientist, R&D Jun 2026 – Present

- Evaluate a heterogeneous portfolio of internal and third-party AI systems prior to production release, spanning conversational agents, IVR voice systems, text-to-SQL, and tool-using agentic workflows.
- Assess risk and define governance criteria, building structured evaluation pipelines that translate ambiguous policy requirements and qualitative failure modes into measurable safety metrics.
- Work cross-functionally across engineering, product, policy, and legal to translate evaluation findings into mitigation strategies and deployment decisions.

Brown University Providence, RI
Ph.D. Researcher, Health NLP Lab & Singh Lab Sept 2021 – May 2026

- Designed and conducted multimodal LLM safety evaluations using 20,000+ expert ratings of 7,000+ point-of-care questions from 283 community health workers in Nigeria, measuring model behavior across five languages and text and speech inputs.
- Published research on limitations of LLM-based evaluation, arguing that model-generated judgments require independent calibration and stronger benchmarking methodology (NeurIPS 2025 position paper).
- Developed UbuntuGuard, a policy-based safety benchmark for low-resource African languages, investigating how safety behavior transfers across linguistic and cultural contexts.
- Investigated persona conditioning as a behavioral prior in clinical LLMs, finding that effects varied across clinical tasks and that automated judges could diverge from expert clinical reasoning.

Optum AI, UnitedHealth Group Remote, USA
Graduate Research Intern – AI/ML Jun 2025 – Oct 2025

- Designed adversarial evaluations targeting hallucination and prompt-injection failure modes in production clinical retrieval systems.
- Built retrieval-augmented generation systems and measured changes in factual grounding against pre-RAG baselines.
- Combined automated metrics, adversarial testing, and clinical expert review into a unified evaluation framework; fine-tuned embedding models with contrastive learning for semantic retrieval.

Council for Scientific and Industrial Research (CSIR) Cape Town, ZA
Research Assistant, Climate and Health Modelling Feb 2019 – Apr 2021

- Extended a statistical early-warning framework for malaria into a machine-learning system forecasting climate-driven diarrhoeal disease outbreaks across southern African provinces; first-author publication in *PLOS ONE*.
- Evaluated ML forecasts against incumbent statistical baselines across provinces, characterizing where prediction performance degraded.

SELECTED SYSTEMS

LLM-as-a-Judge Evaluation Library Humana, 2026
Coauthor and maintainer; internal Python package

- Developed an installable Python library for structured LLM-as-a-Judge evaluation, calibration against human expert labels, and metric aggregation across internal evaluation pipelines.
- Implemented defaults to mitigate known judge failure modes, including position bias, verbosity bias, and self-preference.

Guardrail Recommendation Engine Humana, 2026
Designed and piloting

- Designed an automated workflow that infers a deployment's AI capabilities, derives associated risks, maps candidate controls, and recommends deployment guardrails.
- Structured around a reusable capability-to-risk-to-control taxonomy to support a repeatable, auditable process for guardrail recommendations.

SELECTED PUBLICATIONS

- **Benchmarking is Broken—Don't Let AI Be Its Own Judge.** NeurIPS 2025 (Position Paper).
- **Evaluating Clinical Safety of Multimodal LLMs Across Text and Speech in Low-Resource Languages.** NeurIPS 2026 Workshop on Africa in AI (under review).
- **Benchmarking Multimodal LLMs for Medicine Across African Languages and Context.** ML4H 2026 Symposium, Findings (under review).
- **UbuntuGuard: A Culturally-Grounded Policy Benchmark for Equitable AI Safety in African Languages.** ACL 2026.
- **The Persona Paradox: Medical Personas as Behavioral Priors in Clinical Language Models.** ACL ARR 2026.
- **AfriVox: Probing Multilingual and Accent Robustness of Speech & Multimodal LLMs.** EACL 2026.
- **AfriMed-QA: A Pan-African Multi-Specialty Medical QA Dataset.** ACL 2025. Best Social Impact Paper Award.
- **Machine Learning for Climate-Informed Early Warning of Diarrhoeal Disease Outbreaks.** PLOS ONE.

EDUCATION

Brown University Providence, RI
Ph.D. in Computer Science May 2026
Thesis: *Towards Trustworthy Clinical AI: Knowledge Grounding, Inference Reliability, & Behavioral Control*
Advisors: Carsten Eickhoff and Ritambhara Singh; Sc.M. Computer Science (GPA 4.00/4.00), May 2023

University of Cape Town Cape Town, South Africa
M.Sc. Computer Science (Distinction) May 2021

Bayero University Kano Kano, Nigeria
B.Sc. (Honors) Computer Science May 2014

TECHNICAL SKILLS

Programming	Python, Java, R, SQL
Evaluation	LLM-as-a-Judge design & calibration, benchmark construction, expert annotation, adversarial and red-team evaluation, human-in-the-loop review, failure analysis
Safety	Threat modeling, guardrail design, risk taxonomy, policy-to-metric translation
Statistics	Experimental design, hypothesis testing, inter-rater reliability, forecasting
ML	PyTorch, Hugging Face Transformers, scikit-learn, FAISS, LangChain, Azure, GCP
Languages	English, Hausa, Yoruba, Pidgin English

SELECTED TALKS AND LEADERSHIP

- Invited talk, Google Health AI 2026: *Engineering Trustworthy Clinical AI.*
- Presented *UbuntuGuard*, NVIDIA GTC 2026 Safety Track.
- ACM SIGHPC Computational and Data Science Fellowship (2022–2025); Brown University Diversity Fellowship; National Research Foundation Scholarship, South Africa.