

Closing the Sovereign Safety Gap: A Localized Adversarial Auditing Framework for African AI Governance

Author: Omonivie Cynthia Jatto

Regional Track / Sub-track: Africa, Policy & Governance

Abstract

Africa's AI governance challenge is not primarily a policy gap. Nigeria ranks 35th globally on Policy Capacity in the 2025 Oxford Government AI Readiness Index, yet 72nd overall, revealing a wide gap between policy design and the technical capacity to evaluate whether deployed AI systems actually behave as governed. This report maps that gap across Sub-Saharan Africa's top-ranked countries and proposes the Localized Adversarial Auditing Initiative (LAAI), a framework for building African-led technical audit capacity for AI systems deployed on the continent. Drawing on offensive security methodology, LAAI proposes redirecting Africa's existing penetration testing and red-teaming talent base toward AI system evaluation, addressing risks including data-colonial infrastructure dependency and algorithmic bias in healthcare. We present the audit framework's structure, a pilot case scoping for an African healthcare deployment, and a policy case for why technical evaluation capacity, not compute, is Africa's most realistic governance leverage point.

1. Introduction

AI systems are already being deployed across African healthcare, hiring, and financial services, often without any technical evaluation of how they behave in African contexts. The policy frameworks exist, or are emerging, in several African countries. What does not exist, almost anywhere on the continent, is the technical capacity to check whether deployed systems honor those frameworks.

This report calls that gap the Sovereign Safety Gap: the distance between a country's stated AI policy and its actual capacity to verify, technically, that AI systems operating within its borders meet that policy. It matters for AI safety in Africa specifically because the harms are not hypothetical or distant. They are concrete: deepfake-driven electoral interference, data-colonial dependency on foreign infrastructure, and algorithmic bias trained on non-local data surfacing in healthcare and hiring deployments across the continent.

This report addresses that gap from a specific angle. Africa already has the raw talent to build technical audit capacity. It has a large, underused base of offensive security practitioners, penetration testers, and red teamers whose skills transfer directly to adversarial AI evaluation. What is missing is a framework to redirect that talent toward AI system auditing, and a policy case for why governments and institutions should fund and mandate it.

The Localized Adversarial Auditing Initiative, LAAI, is the framework this report develops to fill that gap.

Our main contributions are:

- The LAAI audit framework: a four-stage methodology (scoping, adversarial testing, evidence documentation, institutional handoff) for conducting localized adversarial audits of AI systems deployed in African contexts, adapted from offensive security practice (Section 3).
- A regional readiness-gap analysis showing that the mismatch between policy capacity and technical verification capacity holds across Sub-Saharan Africa's top-ranked countries, not only in Nigeria (Section 4, Table 1).
- A scoped pilot case and reusable test matrix applying the LAAI framework to an African healthcare AI deployment context (Section 4, Table 2).

This is original work produced during the hackathon, drawing on the prior published research cited in Section 2 and on prior personal positioning work around the Sovereign Safety Gap concept.

2. Related Work

Existing work establishes both the urgency of the problem and the shape of a credible response, but stops short of operationalizing technical audit capacity at the level this report proposes.

Segun et al. (2025) map Africa's distinctive AI risk profile, from deepfake-fuelled electoral interference and data-colonial dependency to compute scarcity, labour disruption, and disproportionate exposure to climate-driven environmental costs, and note that while Computer Incident Response Teams exist across the continent, none currently hosts a dedicated AI Safety Institute or office. Their proposed five-point action plan includes establishing an African AI Safety Institute and developing an early-warning system with inclusive benchmark suites covering 25+ African languages. LAAI is best understood as one possible operational mechanism for delivering on that recommendation: a concrete methodology for what technical evaluation capacity could look like in practice, rather than a parallel proposal.

The Oxford Insights Government AI Readiness Index provides the clearest available evidence of the policy-capacity gap this report addresses. Ten African countries appear in the global top 100 of the 2025 index, led in Sub-Saharan Africa by Kenya, South Africa, Mauritius, and Nigeria. Section 4 presents a rank comparison across these four countries (Table 1). Policy capacity is not the same pillar as the technical and institutional capacity to verify deployed systems meet that policy, a gap consistent with the Sovereign Safety Gap thesis.

Neither of these sources proposes a concrete, replicable methodology for conducting an audit itself. LAAI's contribution is that operational layer: a structured process for translating the risk categories and capacity gaps they identify into an actual evaluation a team could run.

3. Methods

The core contribution developed during this hackathon is the LAAI audit framework: a structured methodology for conducting localized adversarial audits of AI systems deployed across African contexts, built on offensive security practice rather than purely academic evaluation design. The framework is designed to generalize across the continent. We anchor it in Nigeria as a worked case because of data availability and direct local context, not because the problem is Nigeria-specific. The regional readiness data situates Nigeria alongside Kenya, South Africa, and

Mauritius, the other Sub-Saharan African countries in the global top 100, to show the pattern holds beyond one country.

The framework has four components.

3.1 Scoping. Identify a deployed AI system operating in an African context and the specific population it affects. Scoping draws on the risk categories identified by Segun et al. (2025) and on Africa-specific harm categories absent from existing taxonomies.

3.2 Adversarial testing. Apply red-teaming methodology, the same structured approach used in penetration testing, to probe the system for failure modes specific to the local context: language-specific degradation, bias against locally underrepresented groups, and robustness gaps from training data that does not reflect the deployment population.

3.3 Evidence documentation. Produce findings in a format usable by both technical and policy audiences, following the documentation discipline standard in offensive security engagements (reproducible findings, severity rating, remediation guidance), rather than purely academic publication norms.

3.4 Institutional handoff. Route findings to the relevant regulatory or institutional body, with the audit functioning as an accountability mechanism rather than a one-off research exercise.

To ground the framework, we scoped a pilot case for an African healthcare context: an AI-based triage or diagnostic-support tool of the kind increasingly piloted in African health systems, evaluated against criteria including diagnostic accuracy across locally prevalent conditions, performance degradation for patient populations underrepresented in training data, and resilience against erroneous or adversarial input. Healthcare is the highest-stakes deployment category named directly in the hackathon's own risk framing, and a misdiagnosis-driving bias is harder to detect after the fact than a hiring rejection, which makes it a more demanding test of the audit methodology. This pilot is scoped, not executed, within the hackathon timeframe. We treat it as worked methodology rather than a completed evaluation.

4. Results

The readiness-gap analysis is the clearest finding, and it holds at a regional level, not just in Nigeria. Table 1 situates Nigeria against the other three Sub-Saharan African countries in the 2025 Oxford Insights Government AI Readiness Index global top 100.

Table 1. Government AI Readiness, Sub-Saharan Africa Top Four (2025 Oxford Insights Index)

Country	2025 Global Rank
Kenya	65th
South Africa	67th
Mauritius	71st
Nigeria	72nd

Nigeria offers the most detailed public breakdown of the four: it ranks 35th globally in Policy Capacity specifically, well ahead of its 72nd overall position, a gap consistent with our thesis that policy design is outpacing technical verification capacity. We treat Nigeria as an illustrative

case of a regional pattern rather than a uniquely Nigerian problem, since the same overall index structure applies to every country in it, and policy capacity is typically the easier dimension for any government to score well on, since it can be satisfied by publishing a strategy document, while technical verification capacity requires sustained institutional investment. We do not have the equivalent pillar-level breakdown publicly available for Kenya, South Africa, or Mauritius; this is flagged directly in the Limitations section below.

The pilot case scoping for the African healthcare-triage audit produced a working test matrix, organized by the same severity and reproducibility discipline used in penetration testing engagements. Table 2 maps each test category to its evaluation method and expected evidence format.

Table 2. LAAI Healthcare Audit Test Matrix (Pilot Scoping)

Test Category	Evaluation Method	Expected Evidence
Diagnostic accuracy across locally prevalent conditions	Comparative accuracy testing using regionally representative case sets	Per-condition accuracy delta against published baseline, documented with case IDs
Demographic performance gaps	Disaggregated performance testing across age, sex, and subgroups underrepresented in training data	Subgroup-level accuracy and false-negative or false-positive rates, severity-rated
Adversarial input resilience	Structured adversarial input testing using malformed, ambiguous, or manipulated patient data	Reproducible failure cases with severity rating and remediation guidance

This matrix is the concrete, reusable artifact from this hackathon. It is built to apply across African healthcare deployment contexts with minimal adaptation, and it is the first translation of offensive-security audit discipline into an AI-evaluation format aimed specifically at African deployment risk.

5. Discussion and Limitations

The central implication is a reframing of where Africa's leverage in AI governance actually lies, and it is a continental argument, not a national one. Compute and mineral supply chains are slow, capital-intensive, and already contested by better-resourced actors. Technical audit capacity is comparatively fast to build, because the underlying skill base, offensive security practitioners, already exists across multiple African countries and simply needs to be redirected toward AI-specific evaluation. This is a concrete answer to the question of what unique incentive Africa can deploy: not infrastructure ownership, but evaluation capacity that nobody outside the region can credibly replicate, because the harms in question (diagnostic bias in underrepresented populations, language-specific failure modes, context-specific misuse patterns) require local knowledge to detect.

Limitations and Dual-Use Considerations

This report presents a scoped methodology and a readiness-gap analysis, not an executed technical audit. The pilot test matrix has not been validated against a real deployed system, and its effectiveness is therefore unproven. The pillar-level depth of the readiness-gap analysis is

currently only available for Nigeria from public reporting. The Kenya, South Africa, and Mauritius comparisons in Table 1 are limited to overall rank, and a fully rigorous regional claim would require the same pillar-level breakdown for all four countries, which we did not have time to compile this weekend.

On dual use: the adversarial testing methodology described here, by design, surfaces ways AI systems can be manipulated or made to fail. In a healthcare context specifically, this means findings could reveal exploitable failure modes in diagnostic systems, information that, mishandled, could cause real patient harm before it is fixed. We mitigate this by recommending the framework operate under a responsible disclosure model, findings routed to the system owner or regulator before any public documentation, consistent with standard offensive-security practice and especially critical in a healthcare setting. A second risk is that an underfunded or politically captured audit body could use audit findings to selectively target specific AI deployments for reasons unrelated to actual safety. Institutional design, particularly independence and transparency of audit criteria, is the primary mitigation, and is flagged here as a question for future work rather than one resolved by this report.

Future Work

With more time, the next steps would be: (1) executing the pilot audit against a real healthcare AI tool rather than only scoping it, to validate the methodology; (2) replicating the readiness-gap analysis for Kenya and South Africa at the same level of detail available for Nigeria, to confirm the pattern holds with equal rigor across countries; (3) formalizing the test matrix into a reusable, standardized template that other auditors could apply directly; and (4) engaging a specific health regulatory body to test institutional uptake of audit findings.

6. Conclusion

This report argued that Africa's most realistic AI governance leverage point is technical audit capacity, not compute or mineral resources, and proposed LAAI as a framework for building that capacity by redirecting Africa's existing offensive security talent toward AI evaluation. The regional readiness analysis shows this is not a Nigeria-specific problem: every Sub-Saharan African country in the global top 100 of the Oxford Insights index sits well outside the top tier overall, even where, as in Nigeria's case, policy capacity specifically is strong.

The healthcare audit test matrix developed during the hackathon demonstrates that this methodology can be made concrete and reproducible, even within a weekend's scope. The next step is execution: applying the framework to a real deployed system and testing whether African institutions act on what it finds.

Code and Data

This is a policy and governance project; no code or model training was involved. The LAAI audit framework (Section 3) and the healthcare audit test matrix (Section 4, Table 2) are original methodological artifacts produced during the hackathon and are included in full in this report. No additional data files, datasets, or repositories accompany this submission.

References

Oxford Insights. (2025). Government AI Readiness Index 2025. Oxford Insights.
<https://oxfordinsights.com/ai-readiness/government-ai-readiness-index-2025/>

Segun, S. T., Adams, R., Florido, A., Timcke, S., Shock, J., Junck, L., Adeleke, F., Grossman, N., Alayande, A., Kponyo, J. J., Smith, M., Fosu, D. M., Tetteh, P. D., Arthur, J., Kasaon, S., Ayodele, O., Badolo, L., Plantinga, P., Gastrow, M., Adan, S. N., Wiaterek, J., Abungu, C., Apeagyei, K., Eder, L., & Bissyande, T. (2025). Toward an African Agenda for AI Safety. arXiv. <https://arxiv.org/abs/2508.13179>

LLM Usage Statement

I used Claude (Anthropic) as a drafting and formatting assistant for this submission, not as the source of the ideas. The Sovereign Safety Gap framework and the LAAI concept are work I have been building for months before this hackathon, drawing on my background in offensive security. During the weekend, I made every substantive call myself: choosing healthcare as the pilot deployment context instead of hiring or elections, anchoring the regional argument in Nigeria while generalizing the claim across Sub-Saharan Africa, and deciding what actually counted as new work for this submission. I used Claude to help structure the report against the official template, draft prose from my direction, format the tables, and pull together the reference list. I checked every factual claim myself, including the Oxford Insights Government AI Readiness Index rankings and the cited prior work, against primary sources before including them.