

AfriSafeBench: Evaluating LLM Recognition of AI Safety and Governance Risks in African Healthcare AI Deployments

Kimberly Atta-Peters

University of St. Thomas, Minnesota

Regional Tracks: Track 4: Open

Abstract

AfriSafeBench is a benchmark and governance-support tool for testing whether large language models can identify AI safety and governance risks in African healthcare AI deployment scenarios. I evaluated 25 scenarios across seven African countries and ten risk categories using three Groq-accessible models, for 75 total evaluations. The best mean coverage was openai/gpt-oss-20b at 72.00%, followed by llama-3.1-8b-instant at 70.67% and llama-3.3-70b-versatile at 68.00%. The benchmark showed strong recognition of bias and privacy risks but weaker recognition of resource-constrained deployment, vendor dependency, local validation, and monitoring risks. AfriSafeBench also includes a framework-guided report mode that retrieves WHO, NIST, UNESCO, OECD, and African Union guidance to generate governance recommendations from corrected scenario reviews.

1. Introduction

Healthcare AI tools are increasingly deployed in settings where data quality, infrastructure, procurement, clinical staffing, language, and regulatory capacity differ from the environments in which many models are developed. In African healthcare contexts, safety failures may appear not only as technical errors but also as biased triage, weak human oversight, missing local validation, unclear accountability, unsafe medical advice, or vendor dependency that threatens continuity of care.

AfriSafeBench addresses a practical AI safety question: can LLMs identify governance and safety risks in African healthcare AI deployment scenarios before those risks become real-world failures? This matters because many deployment risks are structural rather than obvious. A model may appear clinically useful while still failing because it was not locally validated, cannot be maintained after donor funding ends, gives unsafe advice, or shifts responsibility away from health workers.

Our main contributions are:

- A 25-scenario benchmark focused on African healthcare AI deployment risks across seven countries and ten governance categories.
- A three-model evaluation with scored coverage, missed-risk analysis, and human-audited rescoring flags for semantic matches.
- A framework-guided report mode that turns corrected benchmark reviews into governance recommendations using WHO, NIST, UNESCO, OECD, and African Union guidance.

I reused only the basic application structure and LLM/API plumbing from an earlier independent compliance app that I had created before this hackathon began. During the hackathon, I built AfriSafeBench as a new AI safety benchmark and tool, including the scenario dataset, African healthcare risk taxonomy, scoring workflow, framework-guided report mode, and model comparison results.

2. Related Work

AfriSafeBench builds on AI risk-management and governance frameworks including the NIST AI Risk Management Framework, WHO guidance on AI for health, UNESCO's Recommendation on AI ethics, OECD AI

Principles, and the African Union Continental AI Strategy. It also draws from scenario-based evaluation and retrieval-augmented governance workflows, where model outputs are grounded in external policy or safety documents.

The closest existing approaches are generic AI compliance tools, LLM-as-reviewer systems, and safety benchmarks that test whether models can recognize harmful or unsafe behavior. AfriSafeBench differs by centering African healthcare deployment conditions: local validation, infrastructure constraints, procurement dependency, health-worker oversight, and governance capacity. Someone would use this method when they need to know not just whether a model can discuss AI ethics in general, but whether it can notice risks that appear in realistic African healthcare deployment settings.

3. Methodology

I reused the existing application structure and LLM/API plumbing where possible, but refactored the system into a new AI safety evaluation workflow for AfriSafeBench. The benchmark scenarios, risk categories, scoring logic, result compiler, framework-guided recommendation mode, and analysis artifacts were created for this hackathon. The benchmark dataset contains 25 African healthcare AI deployment scenarios covering Ghana, Kenya, Nigeria, South Africa, Rwanda, Uganda, and Tanzania. Each scenario includes a country, healthcare context, deployment description, severity, expected risk categories, and a researcher-defined explanation. I used scenario evaluation because it tests whether a model can recognize risks in realistic deployment conditions rather than simply define AI safety terms.

The ten risk categories were grounded in AI governance literature and selected because they reflect practical failure modes in healthcare deployment: Bias and Fairness; Human Oversight; Transparency and Explainability; Safety and Reliability; Data Governance and Privacy; Monitoring and Incident Reporting; Distribution Shift and Local Validation; Vendor Dependency; Resource-Constrained Deployment; and Misinformation or Unsafe Medical Advice.

I evaluated three Groq-accessible models: llama-3.1-8b-instant, llama-3.3-70b-versatile, and openai/gpt-oss-20b. This design allowed comparison between a smaller Llama model, a larger Llama model, and a different model family while using the same API workflow. Each model reviewed all 25 scenarios using the same prompt and returned structured JSON with identified risks, severity, scenario-specific explanations, and an overall assessment.

Expected risk categories were compared against model-detected categories using a 0/1/2 rubric: 2 for a scenario-specific match, 1 for a generic or partial match, and 0 for a missed or incorrect risk. I initially found that exact keyword scoring undercounted valid model responses because models often described the same risk using different wording. To address this, I added semantic matching and human-review flags whenever rescoring changed a category. This preserves transparency by recording uncertain scoring decisions rather than silently changing results.

The tool mode uses retrieval-augmented generation. WHO, NIST, UNESCO, OECD, and African Union framework PDFs are extracted, cleaned, chunked, embedded, and stored in a FAISS index. For a corrected scenario review, the system retrieves relevant framework chunks and generates governance recommendations. The dataset, scoring scripts, compiled CSV files, summary JSON files, and generated reports are stored in the project repository so the workflow can be reproduced.

Model	Role in comparison
llama-3.1-8b-instant	Smaller, faster Llama baseline

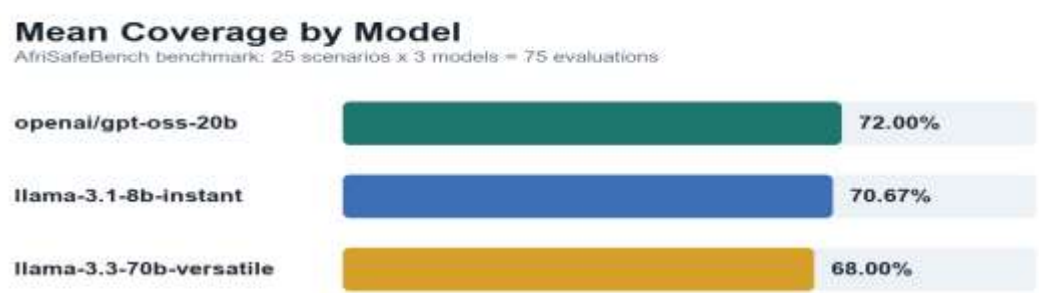
llama-3.3-70b-versatile	Larger Llama model
openai/gpt-oss-20b	Different model family

Table 0. Models evaluated.

4. Results

Observation: the completed benchmark contains 75 evaluations. Overall model performance was close: openai/gpt-oss-20b averaged 72.00% coverage, llama-3.1-8b-instant averaged 70.67%, and llama-3.3-70b-versatile averaged 68.00%. Because the dataset contains 25 scenarios, I interpret these differences as descriptive comparisons rather than statistically significant claims that one model is definitively better.

Figure 1. Mean coverage score by model.



Risk Category	Avg. Coverage	Miss Rate
Bias and Fairness	95.83%	4.17%
Data Governance and Privacy	86.67%	13.33%
Safety and Reliability	83.33%	16.67%
Transparency and Explainability	74.08%	25.92%
Misinformation or Unsafe Medical Advice	73.33%	26.67%
Human Oversight	72.73%	27.27%
Distribution Shift and Local Validation	63.33%	36.67%
Monitoring and Incident Reporting	59.26%	40.74%
Vendor Dependency	58.33%	41.67%
Resource-Constrained Deployment	27.78%	72.22%

Table 1. Category-level coverage and miss rates.

Figure 2. Risk categories with lowest average coverage.

Risk Categories with Lowest Average Coverage

Average across evaluated models. Lower scores indicate harder-to-detect governance risks.



Observation: the strongest categories were familiar AI ethics concepts such as bias, privacy, and safety. The weakest category was Resource-Constrained Deployment, with a 72.22% miss rate. Vendor Dependency, Monitoring and Incident Reporting, and Distribution Shift and Local Validation were also frequently missed. Interpretation: this pattern suggests that models are better at identifying well-known AI ethics risks than structural deployment risks such as infrastructure fragility, sustainability, procurement dependency, and post-deployment monitoring.

A key methodological finding was that 40 of 75 evaluations, or 53.3%, required human-review flags after rescore. Many apparent misses were actually semantic matches expressed in different language. This shows that fully automated scoring can undercount valid model performance unless benchmark systems account for semantic variation transparently.

Robustness checks: I compared performance across scenarios, severity labels, and risk categories. Scenario 017, Community Health Worker Support Tool in Uganda, had the lowest average coverage at 11.11%; its risks were embedded in donor funding, maintenance, and health-worker dependency rather than explicit clinical error. Scenario 002 showed the largest model divergence, with a 100-point spread between models. High-severity scenarios averaged 69.94% coverage while medium-severity scenarios averaged 70.83%, suggesting models did not apply more scrutiny when patient harm risk was higher.

The framework-guided report mode was demonstrated on a representative sample of five scenarios across all three models, generating 15 governance reports with an average of 3.8 recommendations per report. The reports produced recommendations such as local validation, post-deployment monitoring, health-worker training, borderline-case review, and vendor accountability documentation.

5. Discussion, Limitations, and Dual-Use Considerations

Discussion

AfriSafeBench shows that LLMs can identify many AI safety risks in African healthcare scenarios, but their performance is uneven. The most important trend is that models performed better on familiar AI ethics risks, such as bias and privacy, than on structural deployment risks such as infrastructure fragility, procurement dependency, local validation, monitoring, and sustainability. For AI safety, this means a model may sound responsible while still missing the risks that determine whether a healthcare AI system is safe in practice.

Limitations

The main limitations are scope and validation. The dataset contains 25 English-language scenarios, so the results should be treated as an early benchmark rather than a broad claim about all healthcare AI deployments. Expected risks were researcher-defined and should be validated by African healthcare, policy, and legal experts. The scoring also assumes that the expected categories are an appropriate reference standard; if that assumption is wrong or incomplete, model coverage scores may overstate or understate true safety reasoning.

The work does not fully address multilingual deployment, real clinical workflows, country-specific legal requirements, or adversarial attempts to game the benchmark. The framework-guided report mode is not legal or clinical advice.

Dual-Use Considerations

A dual-use risk is that the benchmark could be used to train models to repeat category labels or produce superficial governance reports. Mitigations include requiring scenario-specific explanations, preserving missed-risk outputs, recording human-review flags, showing retrieved framework sources, and requiring expert review for real deployment decisions.

Future Work

Natural next steps are to expand the scenario set, validate expected risks with African healthcare and policy experts, add multilingual and local-language scenarios, test more model families, and strengthen the scoring system with multiple human raters. The tool could also be extended with country-specific legal context, richer framework retrieval, and a public dashboard for comparing models by risk category, country, severity, and healthcare context.

6. Conclusion

AfriSafeBench shows that current LLMs can identify many safety and governance risks in African healthcare AI deployment scenarios, but they remain inconsistent on structural risks such as resource constraints, vendor dependency, monitoring, and local validation. The project contributes both a benchmark for measuring these gaps and a framework-guided tool for converting corrected reviews into governance recommendations.

The main implication is that LLMs should not be treated as complete AI safety reviewers in high-stakes healthcare settings. They can support review workflows, but benchmark results and framework guidance should remain transparent, auditable, and subject to human expert judgment. The AfriSafeBench dataset, scoring scripts and compiled results are openly available to support future benchmark development and community validation.

Figure 3. AfriSafeBench workflow.



Code and Data

- Code repository: github.com/Kimmyatta/AfriSafeBench
- Data/Datasets: [data/afrisafebench_scenarios.json](#); [data/afrisafebench/results/benchmark_results.csv](#); [data/afrisafebench/results/benchmark_summary.json](#)
- Other artifacts: Benchmark results, Guided framework report mode, Result figures and Submission report

References

1. National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce. <https://www.nist.gov/itl/ai-risk-management-framework>
2. World Health Organization. 2021. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. World Health Organization. <https://www.who.int/publications/i/item/9789240029200>
3. UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
4. OECD. 2019. OECD Principles on Artificial Intelligence. Organisation for Economic Co-operation and Development. <https://oecd.ai/en/ai-principles>
5. African Union. 2024. Continental Artificial Intelligence Strategy: Harnessing AI for Africa's Development and Prosperity. African Union. <https://au.int/>

Appendix

Detailed Methodology

Supplementary artifacts include the 25-scenario dataset, compiled benchmark CSV, benchmark summary JSON, framework-guidance summary files, scoring scripts, and generated result figures. The main reproducibility files are: data/afrisafebench_scenarios.json, data/afrisafebench/results/benchmark_results.csv, data/afrisafebench/results/benchmark_summary.json, scripts/compile_afrisafe_results.py, scripts/rescore_afrisafe_results.py, and scripts/compile_framework_guidance_results.py.

A.1 Scenario Fields

The benchmark dataset contains 25 African healthcare AI deployment scenarios. The scenarios cover seven countries: Ghana, Kenya, Nigeria, South Africa, Rwanda, Uganda, and Tanzania. Each benchmark scenario includes scenario_id, title, country, healthcare context, deployment description, expected risk categories, severity, and a researcher-defined explanation. The expected risk categories serve as the reference standard for scoring model outputs.

A.2 Risk Categories

The benchmark uses ten AI safety and governance risk categories: Bias and Fairness; Human Oversight; Transparency and Explainability; Safety and Reliability; Data Governance and Privacy; Monitoring and Incident Reporting; Distribution Shift and Local Validation; Vendor Dependency; Resource-Constrained Deployment; and Misinformation or Unsafe Medical Advice. These categories were grounded in selected governance frameworks but written in practical language suitable for scenario review.

A.3 Models Evaluated

I evaluated three Groq-accessible models: llama-3.1-8b-instant, llama-3.3-70b-versatile, and openai/gpt-oss-20b. Each model reviewed all 25 scenarios using the same scenario text and prompt, producing 75 total benchmark evaluations.

A.4 Evaluation Prompt and Output Schema

Each model received the same scenario text and was asked to identify AI safety and governance risks, explain why each risk was present using scenario-specific evidence, assign severity, and return valid JSON. The expected output schema contained an identified_risks list with risk, explanation, and severity fields, plus an overall_assessment field.

```
{
  "identified_risks": [
    {
      "risk": "brief risk label",
      "explanation": "specific explanation referencing scenario details",
      "severity": "Low | Medium | High"
    }
  ],
  "overall_assessment": "2-3 sentence summary"
}
```

A.5 Scoring Rubric

Each expected category was scored 2 when the model identified the risk with a scenario-specific explanation, 1 when the risk was identified but the explanation was generic or partial, and 0 when the risk was missed or incorrectly identified. Coverage score equals total earned points divided by total possible points. Outputs include matched risks, missed risks, extra risks, raw score, max score, coverage score, and category-level scores.

A.6 Human-Audited Rescoring

Because models often used different wording for the same risk, rescoring included semantic matching and human-review flags. When rescoring changed a category, the result recorded needs_human_review, review_changed_categories, and review_notes. This made score changes auditable instead of hidden.

A.7 Framework Retrieval Workflow

For tool/report mode, WHO, NIST, UNESCO, OECD, and African Union framework PDFs were stored, extracted, cleaned, chunked, embedded, and indexed with FAISS. The framework-guided mode retrieves relevant chunks for a corrected scenario review and generates a framework summary, prioritized recommendations, a governance checklist, limitations and dual-use considerations, and retrieved source references.

This separates evaluation from assistance: benchmark mode measures whether models detect expected risks, while tool/report mode helps a human reviewer respond to identified and missed risks using framework-grounded recommendations.

A.8 Extended Results: Notable Patterns

Hardest scenario: Community Health Worker Support Tool in Uganda (Scenario 017)

The lowest-performing scenario across all three models was Scenario 017, with an average coverage score of 11.11%. Two of the three models scored 0%. The expected risks - Vendor Dependency, Resource-Constrained Deployment, and Human Oversight - were present as subtle structural risks rather than explicit clinical failures. The scenario describes a tool funded by a donor whose cycle ends in eight months with no confirmed maintenance plan, and health workers who report they would not know how to conduct assessments without it. Models consistently described surface-level concerns such as data insecurity and clinical content obsolescence without identifying the underlying sustainability and dependency risks. This suggests models struggle most when risks are embedded in governance and funding structures rather than clinical decision-making workflows.

Biggest model divergence: Maternal Health Risk Scoring in Kenya (Scenario 002)

Scenario 002 showed the greatest model divergence: a 100 percentage point spread. llama-3.1-8b-instant scored 100%, llama-3.3-70b-versatile scored 66.67%, and openai/gpt-oss-20b scored 0%. All three received identical scenario text and prompts. The 0% result was not a parsing failure but a complete mismatch between identified and expected risk categories. A second major divergence occurred in Scenario 025, AI Birth Asphyxia Detection in Rwanda, where llama-3.3-70b-versatile scored 0% while openai/gpt-oss-20b scored 100%. These divergences suggest model-specific reasoning patterns influence risk identification independent of model size.

Severity finding: no meaningful difference

High-severity scenarios averaged 69.94% coverage while Medium-severity scenarios averaged 70.83%, less than one percentage point difference. Models did not apply greater scrutiny to scenarios where patient harm risk was highest. For deployment in real-world healthcare AI governance, this is a significant limitation.

Healthcare context finding: maternal health is the hardest context

Malaria diagnosis and tuberculosis screening scenarios averaged 100% coverage, while maternal health scenarios averaged only 41.67%, the lowest of any context. Maternal health scenarios involve more complex multi-stakeholder risk profiles including community health worker dependency, rural-urban data gaps, and resource continuity risks. The gap between infectious disease contexts, where risks are technically explicit, and maternal health contexts, where risks are systemic and governance-oriented, may reflect a broader pattern in how LLMs reason about structural versus clinical risk.

Country finding: Kenya had the lowest average coverage

Kenya scenarios averaged 63.89%, the lowest of any country, spanning maternal health, medical imaging, clinical decision support, and telemedicine: a diverse set with complex overlapping risk profiles. Ghana averaged the highest at 81.48%. Country-level differences should be interpreted cautiously given the small number of scenarios per country, but the pattern warrants attention in future benchmark expansions.

LLM Usage Statement

I used LLM assistance to help draft and revise report wording, structure the submission document, and support software development tasks. The benchmark data, model outputs, compiled results, figures, and claims reported here were checked against the generated CSV/JSON result files and project code.