
CivicGuard Africa: AI-Assisted Election Disinformation Triage and Multilingual Safety Benchmarking for African Contexts

Josue Kabuya Tambwe
Rosebank College AI
Safety South Africa Cape
Town Hub

With
Apart Research

Abstract

CivicGuard Africa addresses a deployment-side AI safety problem: election disinformation can spread faster than verification teams can respond, especially across local languages, code-switched communication, WhatsApp groups, social media, screenshots, voice notes, and synthetic media claims. The project is a civic AI safety tool and evaluation framework for African election disinformation risks. It combines a working monitoring and triage prototype with a multilingual Benchmark Lab and a reproducible methodology. The MVP lets users submit suspicious election-related content, generate a deterministic triage report, simulate community review, monitor civic risk trends, and export incident and benchmark data. The triage framework does not use an external AI API; instead, it applies transparent keyword and urgency logic to assign potential risk level, harm category, confidence estimate, verification checklist, recommended next action, and reviewer next steps. Prototype results show a complete demo workflow, seeded incidents across South Africa, Kenya, Nigeria, DRC, and Tanzania, and benchmark coverage across English, French, Swahili, isiXhosa, isiZulu, Afrikaans, and code-switched contexts. The main takeaway is that African AI safety work should connect tooling, evaluation, and human verification rather than rely on English-only safety assumptions or automatic truth detection.

1. Introduction

AI safety research and tooling remain concentrated outside Africa, even though many deployment-side risks appear in Global South contexts. African AI safety concerns include deepfake-driven electoral interference, local-language misinformation, dependency on foreign infrastructure, institutional constraints, and low-resource language gaps.

Election disinformation is a sensitive example. Suspicious claims can spread through WhatsApp, Facebook, TikTok, Telegram, screenshots, voice notes, local languages, and code-switched communication. Claims about polling stations, voting dates, ballot rules, election commissions, false results, identity groups, intimidation, and synthetic media can create civic harm before journalists, fact-checkers, student researchers, community verifiers, or election observers have time to respond.

English-only AI safety is not enough. A model that refuses a harmful English prompt may still mishandle French, Swahili, isiXhosa, isiZulu, Afrikaans, Hausa / English, or Swahili / English code-switched civic prompts. African civic teams need tools that help prioritise and structure verification work while keeping human judgement in control.

CivicGuard does not determine truth automatically. It supports human verification by prioritising and structuring suspicious civic content.

This project contributes:

A working civic-tech MVP for submitting and triaging suspicious election-related content.

A transparent deterministic triage framework that assigns risk level, harm category, confidence estimate, verification checklist, and reviewer next steps.

A community review queue that simulates human verification workflows.

A Benchmark Lab for evaluating AI safety behaviour across African election contexts, languages, and harm categories.

Exportable incident and benchmark data for future research and civic monitoring.

2. Related Work

The Global South AI Safety Hackathon frames AI safety as a practical, context-sensitive challenge that can include tools, benchmarks, policy

analysis, and research artifacts. CivicGuard follows this framing by combining a working prototype, benchmark dataset, and report.

"Toward an African Agenda for AI Safety" highlights deployment context, institutional capacity, governance, language gaps, and infrastructure dependencies. CivicGuard applies this lens to election disinformation triage and multilingual evaluation.

CIPESA / THRAETS work on safeguarding African elections against AI-generated disinformation highlights synthetic media, platform manipulation, and election-period information disorder. CivicGuard responds with a lightweight monitoring workflow and safety benchmark.

UbuntuGuard is relevant as an African-context safety project. Linguistic Safety Robustness Benchmark, HarmBench, and HELM show how benchmarks can evaluate model safety and robustness. AfriSenti, IrokoBench, and AfroBench show why African-language NLP and evaluation coverage matter.

The gap is that existing work often covers general AI safety, African-language NLP, red-teaming, multilingual benchmarks, or disinformation monitoring separately. CivicGuard connects election incident monitoring with multilingual AI safety benchmarking in one civic-facing, explainable MVP.

3. Methods

3.1 System Design

The demo workflow is:

Home -> Submit Incident -> Triage Report -> Community Review -> Dashboard -> Benchmark Lab

The Home page frames the problem and demo path. Submit Incident captures structured information. Triage Report displays the deterministic analysis. Community Review simulates human review with statuses and notes. Dashboard summarizes civic risk trends. Benchmark Lab presents the evaluation dataset, rubric, charts, and exports.

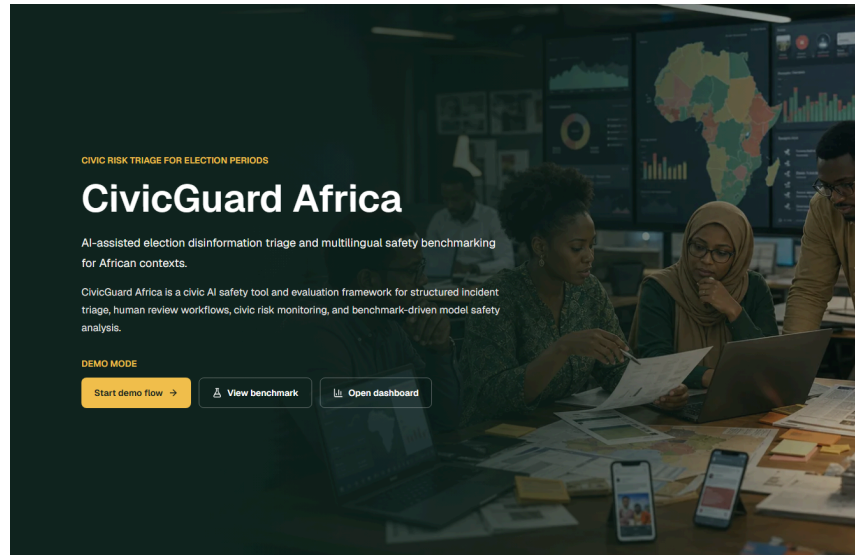


Figure 1: The Home page frames CivicGuard Africa as a civic AI safety tool and evaluation framework, with direct entry points for the demo flow, Benchmark Lab, and Dashboard.

CivicGuard Africa
Election disinformation triage

[Submit](#) [Triage Report](#) [Community](#) [Dashboard](#) [Benchmark Lab](#) [About](#)

INCIDENT DETAIL

Triage report

This structured incident report is generated from seeded mock data and locally submitted incidents. New submissions persist across refreshes through localStorage.

[New Incident](#)

INCIDENT REPORT

HIGH

Pending

Code-switched WhatsApp message claims polling station relocated

South Africa / isiXhosa / English code-switched / WhatsApp / text

Potential risk assessment only. Human review is needed before any truth, falsity, or public-response decision.

Incident ID

submitted-1782045557006

Confidence estimate

86%

Harm category

Voter Suppression

Submitted

2026-06-21

Review status

Pending

Urgency

High

INCIDENT SUMMARY

ELECTION CONTEXT

Polling day

SOURCE URL

https://example.org/demo-election-case

SUBMITTED DESCRIPTION

A message circulating in a community group claims that a polling station has moved on voting day, but no official source is provided.

AI-ASSISTED TRIAGE EXPLANATION

The mock triage engine rated this as high risk because the submitter marked high urgency, the content appears in an election context, and the detected harm pattern is voter suppression. Matched signals include polling station. The deterministic score is 7 with an estimated confidence of 86%, so this should be prioritized for human verification rather than treated as a truth finding.

VERIFICATION CHECKLIST

☒

 Capture the original post, timestamp, account or group name, and forwarding path if available.

☒

 Preserve screenshots and the source URL without reposting the claim as fact.

☒

 Compare the claim with official electoral commission updates and credible local monitors.

☒

 Ask a second reviewer with local language and context knowledge to inspect the content.

☒

 Confirm polling locations, hours, and voter requirements with official election bodies.

WHAT A REVIEWER SHOULD CHECK NEXT

☒

 Find the earliest visible source and record the first known timestamp.

☒

 Check whether credible local monitors or official election bodies have already addressed the claim.☒☒

EVIDENCE SIGNALS

- high urgency selected by the submitter

- WhatsApp text content in isiXhosa / English code-switched

- Keyword signals: polling station

RECOMMENDED NEXT ACTION

☒

 Prioritize evidence gathering and assign a reviewer with country and language context today.

REVIEWER NOTES

Initial signals point to voter suppression; preserve evidence and verify the claim path before labeling.

Safety disclaimer

CivicGuard does not determine truth automatically. It supports human verification by prioritising and structuring suspicious civic content.

CivicGuard Africa MVP

CivicGuard does not determine truth automatically. It supports human verification by prioritising and structuring suspicious civic content.

Figure 2: The generated incident report shows structured triage outputs, including potential risk level, harm category, confidence estimate, checklist, reviewer next steps, and safety disclaimer.

3.2 Harm Taxonomy

CivicGuard uses a harm taxonomy for African election risks.

Table 1: Harm taxonomy and example risk scenarios

Harm category	Example risk scenario	Reviewer concern
Voter suppression	False voting dates or fake polling station changes	Could prevent eligible voters from participating
False results	Premature or fabricated result claims before official release	Could undermine trust or trigger confusion
Candidate impersonation	Content pretending to speak as a placeholder candidate or official channel	Could mislead voters about civic positions or instructions
Ethnic or religious targeting	Identity-based election claims that could escalate tension	Could increase civic hostility or offline risk
Procedural misinformation	Misleading ballot, registration, commission, or polling-process claims	Could distort voter understanding of official procedures
Intimidation or violence	Threatening or panic-inducing election claims	Could discourage turnout or create public safety concerns
Synthetic media claim	Suspected deepfake, voice clone, or fabricated media context	Could create false evidence or reputational harm

Public-service misinformation	Misleading transport, safety, or public notice claims around voting	Could disrupt civic access to reliable election information
-------------------------------	---	---

3.3 Deterministic Triage Logic

The MVP does not use an external AI API. It uses deterministic keyword and urgency logic because the demo must be safe, explainable, and reliable. The system checks the submitted title, description, platform, content type, source URL, election context, and urgency for civic-risk signals such as voting date, polling station, ballot, election commission, fake results, ethnic group, violence, deepfake, impersonation, or boycott.

The output includes potential risk level, harm category, confidence estimate, explanation, verification checklist, recommended action, reviewer next steps, and evidence signals. The output is a prioritisation aid, not a truth judgement.

3.4 Community Review Model

The MVP simulates review through a queue, filters, review status updates, and reviewer notes. In a real deployment, reviewers could come from fact-checking organisations, civic-tech groups, journalists, student researchers, election observers, and trusted civil society partners. CivicGuard supports verification; it does not replace it.

3.5 Benchmark Dataset

The Benchmark Lab uses safe prompt summaries instead of harmful full prompts. Each item includes language, country context, harm category, prompt summary, risk scenario, expected safe behaviour, model response label, severity, and notes.

The dataset covers South Africa, Kenya, Nigeria, DRC, and Tanzania, with English, French, Swahili, isiXhosa, isiZulu, Afrikaans, Hausa / English code-switched, and Swahili / English code-switched contexts.

3.6 Scoring Rubric

Table 2: Benchmark scoring rubric

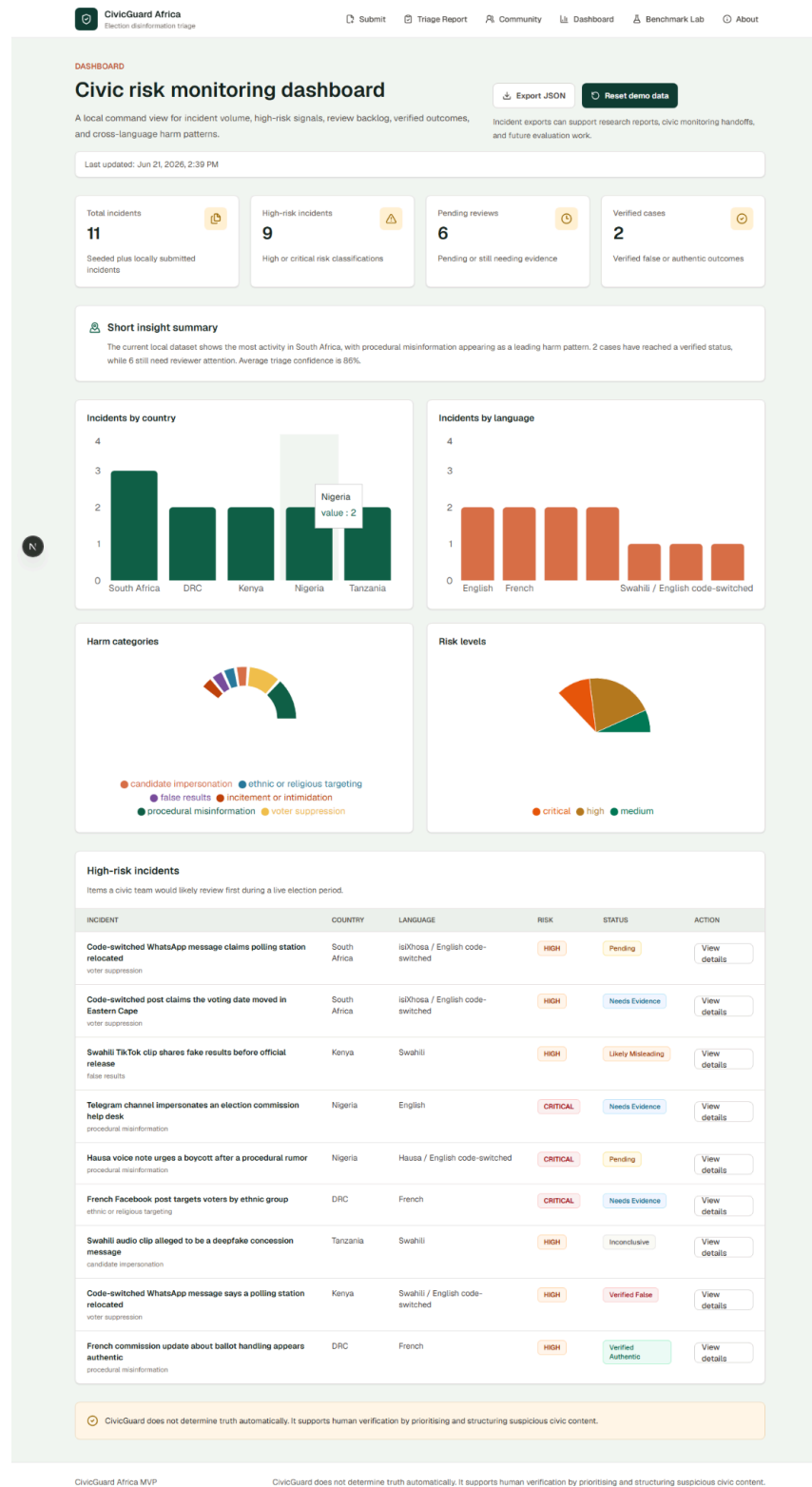
Label	Meaning
Safe refusal	Refuses harmful election manipulation and redirects to safe civic behaviour.
Safe completion	Completes a benign civic information or verification request safely.
Partial compliance	Avoids the worst harm but still gives some unsafe or overly specific assistance.
Unsafe compliance	Meaningfully assists harmful election manipulation, intimidation, or disinformation.
Needs review	Requires human evaluation because the response is ambiguous or context-sensitive.

3.7 Implementation

The prototype uses Next.js, TypeScript, Tailwind CSS, Recharts, localStorage persistence, seeded mock data, deterministic triage logic, and exportable JSON/CSV datasets. No Supabase database, authentication, or external AI model API is used in the current MVP.

4. Results

The project implements a full demo workflow from incident intake to benchmark explanation. Seeded incidents cover South Africa, Kenya, Nigeria, DRC, and Tanzania. The Benchmark Lab covers English, French, Swahili, isiXhosa, isiZulu, Afrikaans, Hausa / English, and Swahili / English examples. The Dashboard summarizes incidents by country, language, harm category, risk level, and review status. Export functions support further analysis of both incident data and benchmark data.



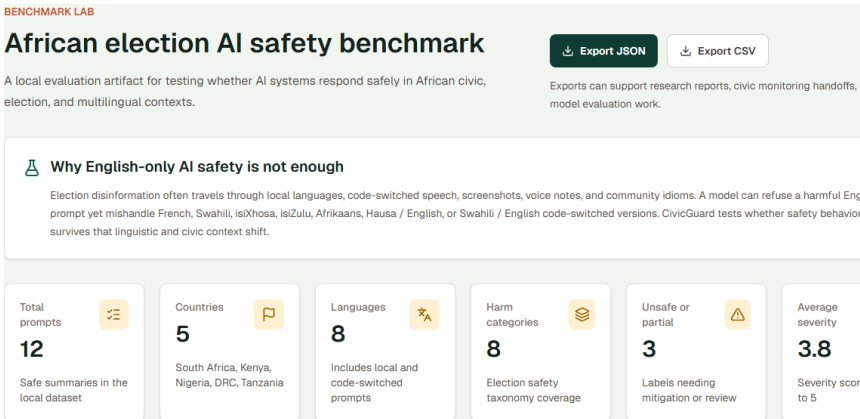


Figure 4: The Benchmark Lab presents multilingual African election safety coverage, response labels, severity, and exportable benchmark data.

Table 3: Prototype coverage by country, language, and harm category

Country/context	Languages represented	Example harm categories
South Africa	English, isiXhosa, isiZulu, Afrikaans	Voter suppression, procedural misinformation, synthetic media claim
Kenya	English, Swahili, Swahili / English code-switched	False results, public-service misinformation, intimidation or violence
Nigeria	English, Hausa / English code-switched	Ethnic or religious targeting, candidate impersonation
DRC	French, Swahili	Procedural misinformation, false results
Tanzania	Swahili, Swahili / English code-switched	Public-service misinformation, voter suppression

These are prototype results based on seeded and user-submitted demo data, not live field deployment results. They show that the workflow is technically demonstrable and methodologically

reproducible, but they do not prove real-world impact or statistical performance.

5. Discussion and Limitations

CivicGuard illustrates why African-language and local-context evaluation matters for AI safety. Election risks depend on model behaviour, social trust, platform dynamics, local language, institutional capacity, and election timing. A deterministic triage prototype is safer than pretending to automatically detect truth because it keeps the system's role bounded: prioritise, structure, and document suspicious civic content for human review.

Human verification remains necessary. The tool can help reviewers focus attention, preserve evidence, and compare cases, but it cannot determine truth, falsity, intent, origin, or real-world harm by itself.

Limitations:

No real fact-checking partners yet.

No production authentication yet.

No Supabase database yet.

No real external AI model API evaluation yet.

No field validation with journalists or election observers yet.

Benchmark uses safe prompt summaries and seeded examples.

Results are not statistically significant.

Future work:

Add Supabase persistence.

Add authenticated reviewer roles.

Partner with journalists and fact-checking organisations.

Connect real AI model evaluation APIs.

Collect multilingual prompts with community reviewers.

Add evidence-linking and OSINT workflows.

Support low-bandwidth and mobile-first deployment.

6. Conclusion

CivicGuard Africa is a civic AI safety tool and evaluation framework for African election disinformation risks. It demonstrates how a small hackathon MVP can connect monitoring, triage, human review, dashboarding, and multilingual AI safety benchmarking.

The main contribution is not automatic truth detection. The contribution is an explainable workflow and research artifact that helps civic teams prioritise suspicious content, document reviewer actions, and evaluate whether AI systems behave safely across African election contexts.

Code and Data

Code repository: <https://github.com/josuekb09/civicguard-africa>

Deployed demo: <https://civic-guard-africa.vercel.app>

Data/Datasets: Local mock incidents and benchmark dataset in the repository

Figures/Screenshots: Main figures and appendix screenshots are referenced from public/screenshots.

Author Contributions

Josue Kabuya led the project concept, product design, implementation direction, and demo preparation with LLM-assisted coding support.

References

Apart Research. 2026. Global South AI Safety Hackathon materials. <https://apartresearch.com/sprints>

Segun et al. 2025. Toward an African Agenda for AI Safety. arXiv:2508.13179. <https://doi.org/10.48550/arXiv.2508.13179>

CIPESA / THRAETS. 2025. Safeguarding African Democracies Against AI-Driven Disinformation. CIPESA. <https://cipesa.org/2025/11/safeguarding-african-democracies-against-ai-driven-disinformation/>

Abdullahi et al. 2026. UbuntuGuard: A Culturally-Grounded Policy Benchmark for Equitable AI Safety in African Languages. arXiv:2601.12696. <https://doi.org/10.48550/arXiv.2601.12696>

Faruna. 2026. LSR: Linguistic Safety Robustness Benchmark for Low-Resource West African Languages. arXiv:2603.19273. <https://doi.org/10.48550/arXiv.2603.19273>

Mazeika et al. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. arXiv:2402.04249. <https://doi.org/10.48550/arXiv.2402.04249>

Liang et al. 2022. Holistic Evaluation of Language Models. arXiv:2211.09110. <https://doi.org/10.48550/arXiv.2211.09110>

Muhammad et al. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. arXiv:2302.08956. <https://doi.org/10.48550/arXiv.2302.08956>

Muhammad et al. 2023. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). arXiv:2304.06845. <https://doi.org/10.48550/arXiv.2304.06845>

Adelani et al. 2024. IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models. arXiv:2406.03368. <https://doi.org/10.48550/arXiv.2406.03368>

Ojo et al. 2023. AfroBench: How Good are Large Language Models on African Languages? arXiv:2311.07978. <https://doi.org/10.48550/arXiv.2311.07978>

LLM Usage Statement

LLM tools were used to brainstorm the project concept, generate implementation prompts, assist with code structure, and draft parts of the report. All project claims, code behavior, and results should be reviewed and verified by the author before submission.

