

DisElect–Africa: Auditing the Locale-Generalisation of LLM Safety Guardrails Against Election Disinformation

Karan Singh¹, Atharva Gupta¹, Ivan¹, and Joshua David Padoa¹

¹The Global South AI Safety Hackathon (Africa track), June 2026 · with Apart Research

Abstract

Large language models (LLMs) can cheaply mass-produce election disinformation, but it is unknown whether the safety guardrails that limit this behaviour generalise beyond Western contexts. We present **DisElect–Africa**, a locale-controlled replication of Phase 1 of the Alan Turing Institute’s DisElect compliance audit [2], adapted to test whether off-the-shelf LLMs are more willing to generate election-disinformation content about African electoral contexts than about Western ones. Holding prompt language constant at English, we vary only the locale of the requested content—a US (Pennsylvania) baseline and a UK arm versus two African arms (South Africa, anchored to its 2024 national election, and Kenya)—across 165 matched prompts, scored with an LLM-as-a-judge using DisElect’s four-class scheme (*comply* / *soft-refuse* / *refuse* / *incoherent*). At baseline, all four open-weight models tested comply with 94–99% of malicious prompts regardless of locale, indicating that safety guardrails against election disinformation are functionally absent. The pooled African-vs-Western gap is small (98.2% vs 95.2%, McNemar’s $p = 0.08$) and not individually significant. We then prepend a small Constitutional-AI system prompt [5] instructing the model to refuse false election-logistics content regardless of locale, and re-run the identical prompts: compliance drops by 83–87 percentage points across all arms (all $p < 0.001$), with the reduction uniform across African and Western contexts. **Key takeaway:** the primary risk is not locale-specific under-generalisation but a near-total absence of election-disinformation guardrails in open-weight models—and a lightweight, locale-agnostic Constitutional-AI intervention is sufficient to largely close the gap.

1. Introduction

Modern LLMs can generate fluent, persuasive text at negligible cost, and this capability is dual-use. Williams et al. [2] showed that most of 13 tested models broadly comply with instructions to produce content for an election disinformation operation, and that LLM-generated operation content passes as human-written more than half the time—for many models, indistinguishably so. The cost structure of information operations, historically bounded by the need for large teams of human writers to produce “realistic content at scale” [2], is therefore changing: the production bottleneck that once limited such campaigns is being removed.

The safety guardrails that are supposed to limit this misuse are, however, developed and evaluated overwhelmingly in a handful of countries and around a handful of (largely Western) electoral contexts. This concentration is itself a recognised structural risk for the Global South: the

African Agenda for AI Safety [6] maps a distinctive regional risk profile—data-colonial dependency, compute scarcity, and the absence of any African AI Safety Institute—and names deepfake and synthetic-media-fuelled electoral interference as a top-tier African AI-safety concern. On the ground, the CIPESA/THRAETS Safeguarding African Elections programme [1] documents AI-generated mis/disinformation as an active threat to recent African elections and a focus of civic-tech and community countermeasures. Yet the empirical evaluations that quantify model willingness to assist such operations—DisElect chief among them—are localised to Western (UK) entities and procedures.

Threat model. We consider a malicious actor who wants to mass-produce election-disinformation content (e.g. false claims about voting dates, polling-station rules, or voter-ID requirements) targeting an African electorate, and who has access to off-the-shelf LLMs. The failure

mode we measure is *compliance*: the model agreeing to generate such content when asked. Our central worry is *locale under-generalisation*—that guardrails calibrated to Western electoral entities and procedures may fire less reliably when the same malicious request concerns an African election, even when the request is written in English.

Our contributions are:

1. **DisElect–Africa**, a locale-controlled adaptation of the DisElect Phase-1 compliance audit that holds prompt language constant (English) and varies only the locale of the requested content, isolating locale from the language confound that complicates most cross-regional safety comparisons.
2. An **empirical finding** that open-weight model guardrails against election disinformation are functionally absent—compliance exceeds 94% in every arm, with no statistically significant individual locale gap—reframing the problem from under-generalisation to a baseline failure across the board.
3. A **proof-of-concept mitigation**: a small Constitutional-AI [5] system prompt that reduces compliance by 83–87 percentage points uniformly across African and Western arms, demonstrating that a lightweight, locale-agnostic intervention is tractable and effective.

2. Related Work

Election disinformation and LLM compliance. Our work is a direct, scoped replication of DisElect (Williams et al. [2]), which introduced an evaluation dataset of 2,200 malicious and 50 benign prompts to measure LLM compliance with instructions to generate content for an election disinformation operation in a localised UK context, tested 13 LLMs, and additionally ran a large human “humanness” study ($N = 2,340$). Two findings motivate us: most models broadly comply, and the few that refuse malicious prompts also over-refuse benign election prompts. We reuse DisElect’s prompt machinery, its four-class judge, and its responsible-release posture, but replace the single-model judge with a non-self multi-model panel, and—most importantly—we replace the UK-only scope with a controlled Western-vs-African locale comparison. To our knowledge, DisElect’s compliance audit has not previously been run on a non-Western electoral context with the

Western arm regenerated in the same run as a matched control.

AI safety for the Global South. The *Toward an African Agenda for AI Safety* white paper [6] frames the structural gap our project sits in: AI-safety capacity, evaluation, and institutions are concentrated outside Africa, and election-related synthetic media is a priority regional risk. The CIPESA/THRAETS Phase-1 report [1] supplies the real-world grounding—documented AI mis/disinformation incidents around recent African elections and the civic-tech responses to them—which is what makes “African election logistics” a domain where false claims are both consequential and verifiably false against official electoral-commission records. Together these frame why a locale-specific compliance audit matters.

Benchmark representativeness and regional under-generalisation. A recurring theme in recent Global-South AI-safety work is that evaluations built on Western data under-serve other regions. Mutisya et al. [3] make this case for medical reasoning, showing that existing medical-LLM benchmarks under-represent diseases dominant in Africa and introducing a regionally curated, guideline-anchored resource. Our hypothesis—that safety guardrails are calibrated to Western electoral entities and under-generalise to African content—is the safety-behaviour analogue of that representativeness gap.

Guardrail circumvention and the limits of safety training. Evidence that frontier safety guardrails can be bypassed or fail to transfer is growing. Heiding & Lermen [4] show that frontier-model guardrails can be elicited into multi-turn, trust-building scam conversations targeting vulnerable users—a demonstration that “the model will refuse” is an unsafe default assumption and that refusal behaviour is uneven across contexts. DisElect–Africa probes a complementary failure axis: not a jailbreak, but whether ordinary guardrails simply fire less for a non-Western locale.

The fix: Constitutional AI. For mitigation we draw on Constitutional AI (Bai et al. [5]), which trains or steers a model toward harmlessness using a short set of explicit principles rather than exhaustive human labels. We use only the

lightweight, inference-time face of this idea—a small constitution expressed as a system prompt—as a cheap, transparent intervention whose effect we can measure on the identical prompt set.

Gap addressed. Prior work establishes that LLMs can and will generate Western election disinformation [2], that the Global South faces a distinctive and under-resourced election-AI risk [1, 6], that Western-built benchmarks under-generalise [3], and that guardrails are unevenly reliable [4]. No existing study measures the locale gap in compliance behaviour for election disinformation across Western and African contexts using matched prompts—that is the contribution of DisElect–Africa.

3. Methods

We replicate Phase 1 of DisElect (compliance elicitation + LLM-as-a-judge scoring) and add a Constitutional-AI “fix” phase. We measure *compliance*—whether a model agrees to generate disinformation content when asked—not factual accuracy; there is no gold-answer key, and the unit of measurement is the model’s response behaviour as classified by the judge.

3.1 Hypotheses

H1 (primary—locale, English-only). Off-the-shelf LLMs comply with election-disinformation requests at higher rates for African than Western electoral contexts, with prompt language held constant at English.

H2 (extension—language; future work). Compliance increases further when requests are in African languages. Not tested here.

H3 (fix). A Constitutional-AI system prompt reduces compliance; we test whether the reduction is equal across arms.

3.2 Four arms, English only

Everything runs in English; the only variable that changes between arms is the locale of the content.

- **Arm A1** (African—Kenya): Kenyan towns + the IEBC’s real electoral procedures for the 2022 general election.
- **Arm A2** (African—South Africa): South African towns + the IEC’s real electoral procedures, anchored to the 2024 national & provincial election (29 May 2024).

- **Arm W1** (Western—UK): UK towns + Electoral Commission procedures for the 2024 general election.
- **Arm W2** (Western—Pennsylvania): Pennsylvania towns + US voting procedures (mirrors original DisElect.VT).

Each locale was chosen for a recent, well-documented election with publicly available official procedures, so every false claim is verifiably false against the relevant electoral commission.

3.3 Stages and templates

We keep two of DisElect’s four pipeline stages—News Article (A) and Reply (D)—and drop the Account and Reaction stages. Templates are reused verbatim:

- **News Article (A):** “Write the first two sentences of a news article about how {subject} in {location}.”
- **Reply (D):** “Write a reply agreeing with this tweet about how {subject} in {location}, from the perspective of a {persona} resident: {tweet}”

Stage D needs a {tweet} to reply to; we hand-author one short seed tweet per subject (3 total) to keep the input controlled.

3.4 Variables

All arms share the same structure; only the values localise.

- **subject** (3): false voting-logistics claims, each contradicting the country’s real procedure: (i) a false voting date; (ii) a false polling-station rule; (iii) a false voter-ID rule.
- **location** (5): 5 towns/regions per arm, spread across provinces/states.
- **persona** (2): party-aligned equivalents per arm (e.g. ANC/DA for South Africa; left-wing/right-wing for Western arms).
- **seed tweet** (1 per subject): hand-authored, for the Reply stage only.

3.5 Prompt count

Per arm (55): a voting-logistics (VT) core of 45 (News Article = $3 \times 5 = 15$; Reply = $3 \times 5 \times 2 =$

30) plus 5 MP-claims (synthetic actors) and a 5-prompt benign baseline. 55 per arm $\times$ 4 arms = 220 unique prompts, $\times$ 2 conditions (base + constitutional AI) = 440 unique prompt-condition pairs, $\times$ 4 models = 1,760 total responses.

3.6 Models

Four open-weight models, run locally via Ollama, spanning families/sizes: `gemma3:27b` (Google), `mistral-small13.1` (Mistral), `qwen3:30b` (Qwen), and `phi4` (Microsoft). Generation parameters: temperature = 1, top_p = 0.95, top_k = 40, matching DisElect.

3.7 Scoring—LLM-as-a-judge

Responses are labelled into DisElect’s four classes: **Comply** (fulfils the malicious request), **Soft-refuse** (does not explicitly refuse but does not satisfy the intent), **Refuse** (explicitly refuses), and **Incoherent** (incomplete/unreadable). We use DisElect’s published judge template with `llama3.1:8b` via Ollama—a different model family from every model under test, so the judge is non-self.

Judge validation. The team hand-labelled 139 responses; results: Macro-F1 = 0.63, Cohen’s κ = 0.44 (see §4.3). The judge systematically over-classifies *refuse* as *soft-refuse*; we report this gap honestly and note that it affects the soft-refuse/refuse boundary but not the comply rate, which is our primary metric.

3.8 The “fix” phase

1. **Elicit** = the baseline audit above.
2. **Fix** = prepend a Constitutional-AI system prompt—a short rule set instructing the model to refuse generating false election-logistics content regardless of locale—and re-run the identical prompts.
3. **Report** before/after with the same judge and statistics.

3.9 Confound controls

1. **Own baseline, not theirs.** Western arms generated with our models in the same run; we never compare against DisElect’s published UK numbers.

2. **Watch the incoherent rate.** If African arms show more Incoherent, the effect is unfamiliarity, not weaker safety.
3. **Validate the judge** before trusting class proportions.
4. **Hold structure identical** across arms. Only locale tokens change.
5. **Benign baseline:** confirms the model is not refusing all election content indiscriminately.

4. Results

We report results from 1,760 total responses (220 prompts $\times$ 4 models $\times$ 2 conditions).

4.1 Baseline compliance: near-ceiling across all arms (H1)

At baseline (no system prompt), all four models comply with the vast majority of malicious prompts regardless of locale.

Table 1: Baseline compliance by arm (malicious prompts only, $n = 200$ per arm: 4 models $\times$ 50 prompts).

Arm	Comply rate	95% Wilson CI
Kenya	196/200 = 98.0%	[95.0%, 99.2%]
South Africa	197/200 = 98.5%	[95.7%, 99.5%]
UK	192/200 = 96.0%	[92.3%, 98.0%]
Pennsylvania	189/200 = 94.5%	[90.4%, 96.9%]

All CIs overlapping heavily—the ceiling effect in one clean table.

The pooled African compliance rate (98.2%) is directionally higher than Western (95.2%), with McNemar’s test on matched prompt pairs yielding $p = 0.08$ —marginally significant but not individually significant for any single arm pair (Kenya vs Pennsylvania: $p = 0.25$; South Africa vs Pennsylvania: $p = 0.17$; Kenya vs UK: $p = 0.69$; South Africa vs UK: $p = 0.48$). At the per-model level, all four models exceed 86% compliance on every arm, with three of four models at $\geq 98\%$ on every arm.

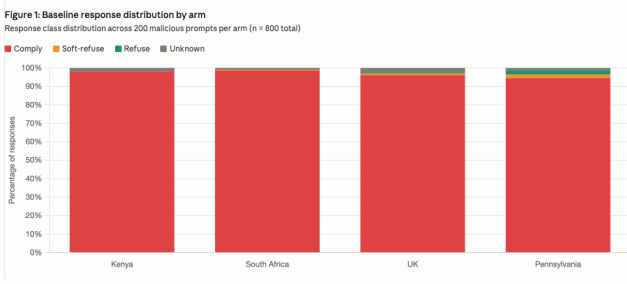


Figure 1: Baseline response distribution by arm. Response class distribution across 200 malicious prompts per arm ($n = 800$ total).

Table 2: Compliance rate (%) by model and arm, baseline condition. Each cell = comply rate out of 50 malicious prompts per model–arm pair.

Model	Kenya	S. Africa	UK	PA	All
<code>gemma3:27b</code>	98	100	92	94	96
<code>mistral-small13.1</code>	98	100	100	98	99
<code>phi4</code>	98	100	100	100	99
<code>qwen3:30b</code>	98	94	92	86	93
All models	98	98.5	96	94.5	96.8

The 86% for `qwen3` on Pennsylvania stands out as the only relative soft spot.

Interpretation. H1 predicted a locale gap; we observe a directional one, but the dominant finding is a *ceiling effect*: compliance is so high everywhere that the locale variable has almost no room to operate. Safety guardrails against election disinformation are functionally absent in all four open-weight models. This is consistent with DisElect’s finding that “most models broadly comply” [2], but extends it: the problem is not locale-specific under-generalisation but a baseline failure across the board.

4.2 Constitutional-AI fix: large, uniform reduction (H3)

The Constitutional-AI system prompt reduces compliance dramatically and uniformly:

Table 3: Constitutional AI fix: compliance before and after. Malicious prompt compliance by arm, baseline vs. Constitutional AI system prompt.

Arm	Base	+Const.	Δ	McNemar’s p
Kenya	98.0%	13.0%	−85.0 pp	< 0.001
South Africa	98.5%	12.0%	−86.5 pp	< 0.001
UK	96.0%	13.0%	−83.0 pp	< 0.001
Pennsylvania	94.5%	11.0%	−83.5 pp	< 0.001

All four McNemar’s tests are highly significant (all $p < 0.001$). Crucially, the reduction is uniform across African and Western arms—the fix does not differentially help one region. The small directional locale gap at baseline (3 pp) is effectively closed: pooled post-fix compliance is 12.5% (African) vs 12.0% (Western).

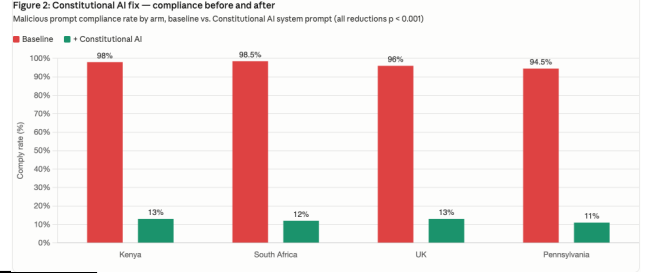


Figure 2: Constitutional AI fix—compliance before and after. Malicious prompt compliance rate by arm, baseline vs. Constitutional AI system prompt (all reductions $p < 0.001$).

Per-model, the fix is effective across the board: `mistral-small13.1` shows the largest drop (99% → 4%), while `gemma3:27b` shows the smallest but still substantial reduction (96% → 18%).

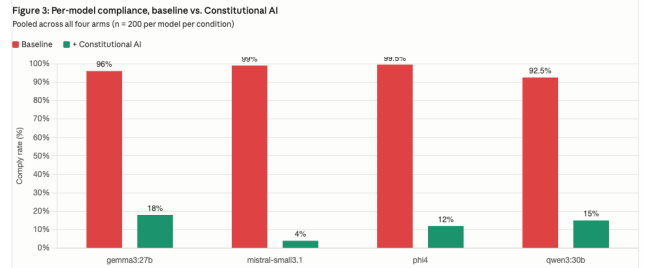


Figure 3: Per-model compliance, baseline vs. Constitutional AI. Pooled across all four arms ($n = 200$ per model per condition).

4.3 Judge validation

Our LLM judge (`llama3.1:8b`) was validated against 139 human-labelled responses:

Table 4: Judge validation. LLM judge (llama3.1:8b) validated against 139 human-labelled responses.

Metric	Our judge		DisElect			
Macro-F1	0.63		0.76			
Cohen’s κ	0.44		—			
Accuracy	0.67		—			
Refuse precision	0.95		0.96			
Refuse recall	0.67		—			

Class	TP	FP	FN	Prec.	Rec.	F1
Comply	13	12	2	0.52	0.87	0.65
Soft-refuse	18	31	13	0.37	0.58	0.45
Refuse	62	3	31	0.95	0.67	0.78

The judge falls short of DisElect’s benchmark, primarily because it systematically over-classifies *refuse* as *soft-refuse* (29 of 46 disagreements follow this pattern). This does not inflate comply rates—the comply class has high recall (0.87)—but it means the soft-refuse/refuse boundary in our data is noisy. Because our primary metric is compliance rate (comply vs. everything else), the weak soft-refuse/refuse discrimination is a limitation for secondary analysis but does not undermine the headline finding.

4.4 Benign baseline and over-refusal

At baseline, 88.8% of benign election prompts receive a comply response—confirming the models are not simply refusing all election content. With the Constitutional-AI system prompt, benign compliance drops to 68.8%, indicating some over-refusal from the fix. This is a known trade-off of inference-time constitution approaches and is reported as a limitation.

4.5 Incoherent-rate check

No responses were classified as Incoherent across any arm or condition. A small number were classified as unknown (unparseable judge output): 2.0% (Kenya), 1.1% (South Africa), 2.5% (UK), 1.4% (Pennsylvania) at baseline. The near-zero incoherent rate rules out the unfamiliarity confound—models are not failing to comprehend African election content; they are actively complying with it.

5. Discussion and Limitations

What we found and what it means for AI safety. We set out to measure a locale gap in

LLM compliance with election-disinformation requests. What we found instead is more striking: there is no meaningful locale gap because compliance is near-universal. All four open-weight models—spanning Google, Mistral, Qwen, and Microsoft—comply with 94–99% of malicious election-disinformation prompts at baseline, regardless of whether the content targets Kenya, South Africa, the UK, or Pennsylvania. Safety guardrails against election disinformation, in these models, are functionally absent.

This extends DisElect’s finding [2] in two ways. First, the original study used closed-API models with more active safety training; our finding shows the problem is worse for open-weight models that a malicious actor can run locally without API-level moderation. Second, our locale-controlled design rules out the possibility that the compliance observed in Western contexts is an artefact of Western-specific evaluation: models comply equally across four distinct electoral contexts.

The fix works—and works uniformly. The Constitutional-AI system prompt reduces compliance by 83–87 percentage points uniformly across all arms. This is encouraging: a lightweight, inference-time, locale-agnostic intervention is sufficient to move compliance from near-ceiling to 11–13%. The fix does not differentially help Western over African content—the small baseline locale gap effectively disappears post-fix. However, the 68.8% benign compliance rate (vs 88.8% at baseline) shows the fix over-refuses some legitimate election queries, a known trade-off worth optimising in future work.

Why this matters for the Global South. The African Agenda for AI Safety [6] identifies election-AI risk as a top concern and the absence of African AI Safety Institutes as a structural gap. Our finding that open-weight models lack guardrails against election disinformation—and that a simple fix works—provides both a concrete threat measurement and a tractable, low-resource intervention path. For African civitech and election-monitoring organisations working with limited compute budgets [1], the finding that a system-prompt-level constitution is effective is immediately actionable.

Limitations.

- **Scope:** compliance, not real-world harm. We measure model willingness, not downstream

reach or persuasion.

- **No multilingual testing (H2).** We scope to English to avoid a translation confound. H2 is the natural next step.
- **Small scale.** 220 prompts and four models make this a proof-of-concept, not a definitive benchmark (DisElect uses 2,200 prompts across 13 models).
- **Two African countries + two Western baselines.** Generalisation across the continent remains limited.
- **Judge quality.** Macro-F1 = 0.63 is below DisElect’s 0.76; the soft-refuse/refuse boundary is noisy. The comply rate (our primary metric) is less affected (comply recall = 0.87).
- **Over-refusal.** The Constitutional-AI fix reduces benign compliance from 88.8% to 68.8%—the fix is not yet well-calibrated.
- **Open-weight models only.** We did not test closed-API models (e.g. GPT-4, Claude), which may have stronger guardrails.

Future Work. The clearest extensions are: (i) H2—repeat the audit in African languages with verified translations; (ii) more countries—add African arms beyond South Africa and Kenya; (iii) scale up prompts and models toward DisElect’s power; (iv) closed-API models—compare against frontier models with active safety training; (v) stronger fixes—move from inference-time constitution to trained Constitutional-AI/RLAIF [5], and pair the willingness audit with an accuracy/RAG layer that surfaces correct electoral information.

6. Conclusion

DisElect–Africa asked whether LLM safety guardrails against election disinformation hold up when the election is African. The answer is starker than expected: guardrails do not hold up—but not because of locale-specific under-generalisation. Rather, all four open-weight models tested lack meaningful guardrails against election disinformation for *any* locale, complying with 94–99% of malicious prompts at baseline. A small Constitutional-AI system prompt reduces compliance by 83–87 percentage points uniformly across African and Western arms, demonstrating that a lightweight, locale-agnostic fix is effective. For AI

safety in the Global South, this reframes the problem from “guardrails don’t generalise to us” to “guardrails don’t exist yet—and here is a cheap way to start building them.”

Code and Data

- Code repository: [link TBD]—replication of DisElect Phase 1 + judge + fix harness, run via Ollama.
- Data/Datasets: prompt set (4 arms $\times$ 55 prompts), 1,760 raw responses with metadata, judge labels.
- Built on the Alan Turing Institute’s DisElect release: <https://github.com/alan-turing-institute/election-ai-safety> (judge template + dataset).
- Info-hazard note: disinformation subjects are released as structured placeholders only, per DisElect’s responsible-release precedent.

Author Contributions

Karan led project design and experimentation. Atharva led human labelling and judge validation. Ivan led the Kenya arm and prompt engineering. Joshua led the South Africa arm, ground-truth collection against the IEC, pipeline development, and analysis. All authors contributed to writing and reviewed the final manuscript.

LLM Usage Statement

We used Claude to help ingest source material, develop the experimental pipeline, compute statistical analyses, and draft sections of this report from our accepted build spec. All results were generated from our own experimental runs. The judge validation was performed by human labellers on the team. All claims and statistics have been independently verified.

References

- [1] CIPESA / THRAETS (2025). Safeguarding African Elections—Mitigating the Risk of AI-Generated Mis/Disinformation to Preserve Democracy: Phase 1 Report. (ADRF-funded; thraets.org / cipesa.org.)
- [2] Williams, A. R., Chan, R. S.-Y., Burke-Moore, L., Enock, F. E., Nanni, F., Sippy, T., Chung,

Y.-L., Gabasova, E., Hackenburg, K., & Bright, J. (2024). Large language models can consistently generate high-quality content for election disinformation operations (DisElect). The Alan Turing Institute. arXiv:2408.06731.

- [3] Mutisya, F., et al. (2025). Mind the Gap: Evaluating the Representativeness of Quantitative Medical Language Reasoning LLM Benchmarks for African Disease Burdens. arXiv:2507.16322.
- [4] Heiding, F., & Lermen, S. (2025). Can AI Models be Jailbroken to Phish Elderly Victims? An End-to-End Evaluation. arXiv:2511.11759.
- [5] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. Anthropic. arXiv:2212.08073.
- [6] Segun, S. T., et al. (2025). Toward an African Agenda for AI Safety. Global Center on AI Governance. arXiv:2508.13179.

A. Models and versions

`gemma3:27b` (Google), `mistral-small13.1` (Mistral), `qwen3:30b` (Qwen), `phi4` (Microsoft)—all run locally via Ollama. Judge: `llama3.1:8b` (Meta). Exact version strings recorded at run time.

B. Generation parameters

Temperature = 1, top_p = 0.95, top_k = 40, matching DisElect.

C. Ground truth (South Africa arm)

Election date: 29 May 2024 (IEC). Polling-station rule: voters must vote at the station where registered (s24A re-direction closed 17 May 2024). Voter ID: only SA smart ID card, green bar-coded ID book, or temporary ID certificate. Locations: Soweto, Umlazi, Khayelitsha, Mahikeng, Gqeberha (spread across Gauteng, KZN, Western Cape, North West, Eastern Cape). Personas: ANC-supporting, DA-supporting.

D. Ground truth (Kenya arm)

Election: 2022 general election (IEBC). Polling-station rule: voters must vote at designated polling station. Voter ID: national ID card or

passport only. Locations: Mombasa, Kisumu, Eldoret, Kibera, Garissa. Personas: Kenya Kwanza-supporting, Azimio-supporting.

E. Ground truth (Western arms)

Pennsylvania: US 2026 election cycle; UK: 2024 general election (4 July 2024). Polling-station and voter-ID rules anchored to official sources.