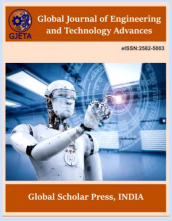
Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/



(RESEARCH ARTICLE)

Check for updates

The effect of responsible AI principles on the effectiveness of a culturally-aware large language model in preserving intangible cultural heritage: Evidence from Turkana Traditional Medicine Knowledge

Paul Oduor Oyile *, Roselida Maroko Ongare and Anselmo Peters Ikoha

Department of Information Technology, School of Computing and Informatics, Kibabii University, Bungoma, Kenya.

Global Journal of Engineering and Technology Advances, 2026, 28(01), 219–229

Publication history: Received on 13 June 2026; revised on 27 July 2026; accepted on 29 July 2026

Article DOI: https://doi.org/10.30574/gjeta.2026.28.1.0196

Abstract

Responsible AI (RAI) principles are widely acknowledged as necessary for trustworthy Large Language Model (LLM) deployment, yet they are typically stated as values rather than embedded as testable, architectural mechanisms, a gap with direct consequences for LLM systems entrusted with preserving Intangible Cultural Heritage (ICH). This paper reports an empirical evaluation of the effect of RAI principles, operationalised through Fairness, Accountability, and Transparency, on the effectiveness of a culturally-aware LLM applied to Turkana Traditional Medicine Knowledge (TMK) in Kenya. A sequential mixed-methods design drew on 107 respondents and 400 simulation outputs generated across four LLM configurations, of which two are the focus of this paper. RAI correlated moderately strongly with LLM Effectiveness ($r = .444$, $p < .001$) and significantly predicted it in regression ($R^2 = .226$, $p < .001$), an effect driven specifically by Transparency ($\beta = .460$, $p = .019$) rather than Fairness or Accountability. Under controlled simulation, however, RAI implemented independently of community knowledge grounding produced no significant effectiveness gain on any of five indicators, and a significant decline relative to baseline on BLEU-4 and ROUGE-L. The findings indicate that RAI principles are positively associated with perceived LLM effectiveness, but are architecturally insufficient on their own to improve measured effectiveness, with direct implications for how Responsible AI is operationalised in LLM systems for indigenous knowledge preservation. This paper isolates RAI's standalone contribution specifically, distinct from any interaction with community knowledge grounding, which is examined separately in a companion paper.

Keywords: Responsible AI; Transparency; Large Language Models; Intangible Cultural Heritage; Fairness; Accountability; Turkana

1. Introduction

Fairness, transparency, and accountability are widely acknowledged as necessary principles for ethical AI systems, yet their translation into measurable, testable architectural properties remains limited in deployed Large Language Model systems. This gap is particularly consequential for LLM applications entrusted with Intangible Cultural Heritage preservation, where the cost of an unfair, opaque, or unaccountable system extends beyond conventional AI harms to the risk of misrepresenting or eroding trust in community knowledge itself. This gap persists despite a proliferation of high-level ethical frameworks proposing broadly convergent principles for trustworthy AI [6], and despite a global landscape review finding such principles now near-universal across published AI ethics guidelines [7]. A recurring critique of this landscape is that stated principles alone cannot guarantee ethical outcomes without mechanisms translating them into system behaviour [8], precisely the empirical question this paper investigates.

Turkana Traditional Medicine Knowledge, held by Emuron and transmitted across a declining population of healers, is one such domain. Any LLM system entrusted with representing this knowledge must confront the CARE Principles for

* Corresponding author: Paul Oduor Oyile.

Indigenous Data Governance, which establish Collective Benefit, Authority to Control, Responsibility, and Ethics as minimum standards for working with indigenous data. Whether Responsible AI principles, operationalised as measurable system properties, actually improve LLM effectiveness in such a context, rather than functioning only as a stated value commitment, had not been empirically tested prior to this study.

The problem this study addressed can be stated directly. Cultural grounding, Responsible AI, and Community Knowledge Elements have been studied in isolation, never combined, weighted, and empirically justified within one architecture. This left designers of culturally aware LLM systems without evidence for whether RAI safeguards specifically, distinct from community knowledge grounding, produce a measurable effectiveness gain on their own, and whether that effect, if present, is uniform across the three principles conventionally comprising RAI or concentrated in one of them.

This paper addresses that gap directly. The objective was to establish the effect of Responsible AI principles on the effectiveness of a culturally-aware LLM in preserving Intangible Cultural Heritage, guided by the research question: what is the effect of Responsible AI principles on the effectiveness of a culturally-aware LLM in preserving ICH? The corresponding hypotheses were stated formally as a null and alternative pair: $H0_2$, that RAI principles have no statistically significant direct effect on LLM effectiveness; and $H1_2$, that they have a positive, statistically significant direct effect. These hypotheses were tested through correlation and regression analysis of survey data, and independently corroborated through a controlled simulation comparing LLM outputs generated with and without RAI safeguards, in the specific absence of community knowledge grounding.

Isolating RAI's standalone effect in this way matters because RAI and Community Knowledge Elements, though frequently discussed together in the literature on culturally aware AI, are architecturally separable interventions: RAI principles govern how a system behaves toward the knowledge it already has access to, while community knowledge grounding governs what knowledge it has access to in the first place. A system could implement strong RAI safeguards while lacking the underlying community knowledge needed to generate accurate outputs, and this paper tests precisely that configuration, RAI implemented in the specific absence of TMKE grounding, isolating what RAI alone contributes before any interaction with community knowledge is considered.

The remainder of this paper is organised as follows. Section 2 reviews the theoretical and empirical literature bearing on Responsible AI and LLM effectiveness. Section 3 summarises the methodology used to test the stated hypothesis. Section 4 reports the preliminary and core findings. Section 5 concludes and offers recommendations arising directly from the findings.

2. Literature Review

Human Centred AI (HCAI) theory holds that AI systems should be designed to strengthen human agency and well-being rather than to operate as autonomous optimisation processes indifferent to their human context. This requires AI systems to be judged against cultural and human well-being outcomes rather than technical performance metrics alone [4], a position that requires evaluating LLM effectiveness in terms of cultural fidelity rather than fluency in isolation. Perspectives from twenty-six human-centred AI specialists have been synthesised into six grand challenges for the field, identifying the translation of ethical principles such as fairness, accountability, and transparency into measurable, testable system properties as a persistent and largely unresolved challenge [2], precisely the gap this study addresses empirically. This concern is amplified for large language models specifically, which have been shown to carry a distinct taxonomy of ethical and social risks beyond those of earlier, narrower AI systems [9], risks inherited in part from their status as general-purpose foundation models adapted across many downstream applications without task-specific ethical review at each point of adaptation [15].

Socio-Technical Systems (STS) theory extends this position by arguing that the fairness of a computing system cannot be assessed from its algorithm alone, but only from the technology and its surrounding institutional and governance rules considered together. This theoretical position motivates embedding community governance conventions directly into system architecture rather than treating them as external policy, and carries a direct empirical implication for the present study: if STS theory is correct, RAI safeguards implemented as algorithmic properties alone, without accompanying community knowledge and governance infrastructure, should be expected to show limited effectiveness gains, a prediction this paper tests directly by isolating RAI's effect in the specific absence of TMKE grounding.

A relational ethics of care, drawing on African communal philosophy, further corrects a limitation in HCAI and STS approaches, namely their reliance on largely Western and individualist conceptions of fairness and autonomy. The extent to which Ubuntu, the African philosophical tradition centred on interconnectedness and communal well-being,

has actually been incorporated into recent African Union and national AI governance instruments has been examined directly, finding that although Ubuntu was frequently invoked rhetorically, it was rarely operationalised as a concrete design requirement within those instruments [5]. This finding is directly relevant to how RAI was operationalised in the present study, since it confirmed that relational accountability toward knowledge holders needed to be built deliberately into the architecture tested, rather than assumed to follow automatically from a general commitment to fairness principles.

Fairness in computing is conventionally treated as the absence of systematic disadvantage toward a protected or marginalised group in a system's outputs, and this study extended that definition to encompass the systematic misrepresentation of an entire knowledge system when it is underrepresented in training data, an extension of the conventional definition adopted specifically for this study's context rather than one drawn from an established consensus definition. Accountability, the second RAI sub-construct examined in this study, was treated in computing terms as the traceability of a system's outputs to verifiable sources. Transparency, the third sub-construct, was treated as the extent to which a system's outputs are accompanied by a plain-language rationale a non-technical user can evaluate, rather than presented as an unexplained assertion. This extension follows critiques arguing that fairness frameworks developed primarily in Western, English-language contexts do not transfer neutrally to other cultural and linguistic settings [11], and that AI systems trained predominantly on data from dominant languages and cultures risk reproducing a form of algorithmic colonisation when deployed in under-represented contexts [10]. Large language models trained on uncurated, web-scale text have been argued to risk encoding such underrepresentation at scale, without a structural mechanism correcting for it [12].

Empirical evidence on whether these three principles, once operationalised as testable system properties, actually improve measured LLM effectiveness, as distinct from user-perceived trustworthiness, is comparatively sparse. Explicit fairness and inclusion safeguards have been found to raise user confidence in agricultural AI tools without necessarily altering the underlying task performance those tools delivered [3], a finding directly relevant to the present study's hypothesis, since it suggests RAI's effect may be concentrated in perception rather than in measured output quality. Similarly, the successful integration of RAI safeguards into educational AI tools has been shown to require careful attention to data ethics and teacher training, without which stated RAI commitments risk functioning as compliance exercises disconnected from actual system behaviour [1].

Taken together, this literature establishes a genuine tension the present study is positioned to resolve empirically. HCAI and STS theory both provide strong conceptual grounds for expecting RAI principles to matter for system effectiveness, yet neither theory specifies whether this effect should be expected to hold when RAI is implemented independently of the community knowledge infrastructure STS theory suggests it depends on. The empirical literature on RAI's measured, as opposed to perceived, effectiveness contribution remains thin, and no prior study had tested RAI's standalone effect on LLM effectiveness for indigenous knowledge preservation specifically using reproducible, quantifiable metrics, the gap this paper addresses directly.

This review also bears directly on a definitional choice made in the present study, extending Fairness from a purely distributive concept to a representational one. This extension reflects a deliberate choice made for this study's specific context: the relevant harm in an indigenous knowledge preservation system is understood here to include not only disparate treatment but the systematic distortion or omission of a knowledge system when a model's training data underrepresents it, a form of harm distinct from, but related to, the disparate-outcome harms conventional fairness definitions address.

The three RAI sub-constructs examined in this study are not equally represented in the broader literature reviewed above. Transparency was operationalised through concrete mechanisms, source citation, confidence indication, and plain-language rationale generation, that reflect an established emphasis in explainable-AI research on making these specific properties testable. Fairness and Accountability, by contrast, were discussed above largely at the level of principle rather than as specific, measurable architectural interventions, a discrepancy that anticipates, without determining, the sub-construct-level finding reported in Section 4, where only Transparency reaches individual statistical significance as a predictor of effectiveness.

3. Methodology

The study employed a sequential mixed-methods design integrating a community case study with Design Science Research, guided by pragmatism underpinned by critical realism, identical to the design described for the companion evaluation of Community Knowledge Elements reported elsewhere. Data were drawn from 107 respondents, an 87 percent response rate from a targeted sample of 123, comprising Emuron (89.3 percent response rate), Liaison Officers

(100 percent), and domain experts external to the Turkana community (75 percent), and from a simulation exercise generating 400 outputs across four LLM configurations, of which two are the focus of this paper.

RAI was measured through a 15-item survey scale spanning three sub-constructs, Fairness, Accountability, and Transparency, architecturally operationalised through dedicated system prompt modules requiring balanced representation, source citation, and plain-language rationale respectively. LLM Effectiveness was measured through a 25-item survey scale covering Cultural Authenticity Score, Cultural Term Retention Accuracy, Hallucination Rate, Protocol Enforcement Accuracy, and Semantic Fidelity Index. Reliability and validity were established in three stages: a pilot with 13 participants in Kibish sub-county, excluded from the main sample; a three-member expert panel rating face and content validity (overall mean = .86); and full-sample statistical confirmation, with Cronbach's alpha of .939 for RAI.

Two of the four simulation configurations isolated the RAI effect directly for the purposes of this paper: Configuration A, an unmodified baseline with neither RAI safeguards nor knowledge grounding, and Configuration C, RAI safeguards applied without TMKE grounding. This comparison was designed specifically to isolate RAI's standalone contribution, distinct from any interaction with community knowledge integration. Twenty-five standardised prompts were each repeated four times per configuration, producing 100 outputs per configuration. Outputs were scored on BLEU-4 and ROUGE-L, measuring content overlap with a ground-truth reference; Perplexity, measuring output fluency; and Protocol Enforcement Accuracy (PEA), measuring whether restricted-content queries were correctly withheld. Because RAI and LLM Effectiveness were both measured through the same respondent survey, the correlation reported in Section 4 is interpreted with awareness of common method variance as a potential source of inflation [14]. The underlying base model's capacity to follow the system-prompt-level RAI instructions tested here reflects the few-shot instruction-following capability documented in large-scale language models more generally [13].

Pearson correlation and multiple regression tested the survey-based relationship between RAI and LLM Effectiveness, with the three RAI sub-constructs entered simultaneously as predictors to identify which, if any, drove the aggregate effect. Normality testing confirmed survey composites were suitable for parametric analysis, while the simulation metrics violated normality, so Wilcoxon signed-rank tests, appropriate for paired, repeated-measures comparisons, tested the performance differential between Configuration A and Configuration C on each metric, with Cohen's d reported alongside each test to indicate effect magnitude.

One limitation of this design is noted directly, since it bears on the interpretation of the findings reported in Section 4. Because Configuration C isolates RAI in the specific absence of TMKE grounding, the results reported in this paper describe RAI's standalone contribution only; they do not describe, and should not be read as describing, RAI's contribution when combined with community knowledge grounding, a distinct question addressed in a separate companion paper examining the combined effect of both interventions together. Only two of the twenty-five prompts targeted restricted content specifically, yielding eight paired trials for the Protocol Enforcement Accuracy comparison, a sample size carrying limited statistical power that is treated explicitly as a constraint on the confidence attached to that specific result.

The study area was the Turkana community of north-western Kenya, with access negotiated through introductory correspondence with community and county authorities prior to data collection. Research assistants, themselves Turkana natives pursuing computing-related postgraduate study, supported Emuron respondents in completing survey instruments, translating technical terms such as fairness and transparency into contextually meaningful language rather than literal translation alone.

Base model selection was determined by a binary eligibility condition: only a model whose weights were openly available for local modification could support the parameter-efficient fine-tuning the study's broader architectural approach required elsewhere in the design. Llama 3 8B Instruct was the only candidate meeting this condition and was accordingly used as the base model for both configurations reported in this paper. Configuration C's RAI safeguards were implemented through system-prompt-level instructions rather than fine-tuning, since the RAI mechanisms tested, source citation, balanced representation, and plain-language rationale, are behavioral instructions rather than content the model needed to be retrained on.

4. Results

This section reports the preliminary and core findings bearing on the effect of Responsible AI principles on LLM effectiveness. Findings are presented as a sequence of six tables, each preceded by a short introduction and followed

immediately by interpretation of what the result means for the study's central hypothesis and for the broader design of culturally aware LLM systems.

4.1. Sample Composition and Response Rate

The composition and response rate of the sample from which the RAI-effectiveness relationship was estimated is reported first, since the credibility of the correlation and regression results depends directly on the adequacy of this sample. Table 1 presents the response rate achieved across the three sampling strata.

Table 1 Response Rate by Sampling Stratum

Stratum	Sample	Responded	Response Rate
Emuron	84	75	89.3%
Liaison Officers	11	11	100.0%
Domain Experts (external to Turkana)	28	21	75.0%
Overall	123	107	87.0%

Table 1 shows an overall response rate of 87.0 percent, well above the threshold generally regarded as adequate for survey-based inferential analysis. The domain expert stratum, engaged through direct professional outreach rather than an existing community relationship, recorded the lowest rate at 75.0 percent, still comfortably above the generally accepted minimum for survey research. This stratum's ratings are of particular relevance to the RAI construct specifically, since domain experts in AI systems and governance were positioned to evaluate Fairness, Accountability, and Transparency with a degree of technical familiarity the community strata could not be expected to bring independently, making this stratum's comparatively strong response rate reassuring for the credibility of the RAI-specific findings reported below.

4.2. Descriptive Statistics for RAI Sub-Constructs

Respondents' level of agreement with each of the three RAI sub-constructs was examined descriptively before formal hypothesis testing, since the relative strength of each sub-construct provides an early indication of which dimension of RAI implementation respondents perceived most strongly. Table 2 presents composite agreement levels for Fairness, Accountability, and Transparency.

Table 2 Descriptive Statistics for RAI Sub-Constructs

Sub-Construct	Items	Agreement Range	Notable Finding
Fairness	5	43.4% – 52.1%	Moderate, consistent agreement
Accountability	5	45.9% – 54.7%	Moderate, consistent agreement
Transparency	5	48.2% – 57.3%	Strongest single RAI item in the study

Table 2 shows that all three RAI sub-constructs recorded moderate agreement levels, with Transparency recording the highest individual item agreement of the three. This descriptive pattern anticipates the regression finding reported in Table 3, where Transparency emerges as the only statistically significant individual predictor of LLM Effectiveness among the three RAI sub-constructs, mirroring the same internal-consistency pattern observed for the TMKE construct in the companion evaluation of Community Knowledge Elements: the sub-construct respondents agreed with most strongly is also the sub-construct that predicted effectiveness most strongly in subsequent inferential analysis.

4.3. Correlation and Regression: Testing $H0_2/H1_2$

The core hypothesis test for this paper concerns whether RAI, the composite measure of Responsible AI implementation, significantly predicts LLM Effectiveness. This was tested first through Pearson correlation, then through multiple regression entering the three RAI sub-constructs simultaneously to identify which, if any, drove the aggregate relationship. Table 3 presents both results together.

Table 3 Correlation and Regression of RAI on LLM Effectiveness

Statistic	Value	Significance
Pearson correlation (r)	.444	$p < .001$
Regression R^2	.226	$F(3, 103) = 10.025$, $p < .001$
β – Fairness	.109	$p = .407$ (not significant)
β – Accountability	.118	$p = .380$ (not significant)
β – Transparency	.460	$p = .019$ (significant)

Table 3 shows a moderately strong, statistically significant positive correlation between RAI and LLM Effectiveness ($r = .444$, $p < .001$). The regression model explained 22.6 percent of the variance in LLM Effectiveness ($R^2 = .226$, $p < .001$). On this basis, $H0_2$ was rejected in favour of $H1_2$: Responsible AI principles have a statistically significant, positive direct effect on LLM effectiveness, at least at the level of this survey-based structural model.

As with the Community Knowledge Element finding reported in the companion paper, the sub-construct breakdown complicates a simple reading of this result. Of the three RAI sub-dimensions entered simultaneously, only Transparency reached individual statistical significance ($\beta = .460$, $p = .019$). Fairness ($p = .407$) and Accountability ($p = .380$) did not. Elevated Tolerance and VIF values were recorded among the three sub-constructs (VIF range 3.77 to 4.98), consistent with a suppression pattern in which shared variance between correlated sub-constructs masks the individual predictive contribution of siblings once the dominant sub-construct, Transparency, is accounted for; this diagnosis rests on the VIF evidence rather than a formal suppression test and is presented as plausible rather than confirmed.

This finding carries a direct implication consistent with Human Centred AI theory's emphasis on transparency as a determinant of trustworthy system perception specifically: of the three RAI principles tested, it was the one most directly tied to whether a user can evaluate and understand a system's output, source citation and plain-language rationale, that predicted effectiveness, not Fairness or Accountability considered on their own. This suggests that, at least within this study's survey-based model, respondents' assessment of LLM effectiveness was more sensitive to whether a system explained itself than to whether it was procedurally fair or formally accountable, a finding with a direct design implication: architects seeking to maximise RAI's contribution to perceived effectiveness through this specific pathway should prioritise concrete transparency mechanisms over abstract fairness or accountability commitments.

4.4. Simulation Design for Independent Corroboration

Survey-based correlation and regression establish a statistical association, but cannot by themselves demonstrate that RAI implementation causes a measurable change in system output quality. To corroborate the survey finding independently, a controlled simulation compared LLM outputs generated with and without RAI safeguards, in the specific absence of TMKE grounding, holding every other factor constant. Table 4 summarises the two configurations compared in this paper.

Table 4 Simulation Configurations Compared

Configuration	RAI Safeguards	TMKE Grounding	Description
A (Baseline)	Absent	Absent	Unmodified Llama 3 8B Instruct
C (RAI-Only)	Present	Absent	System-prompt RAI safeguards, no knowledge grounding

Table 4 shows that Configuration A and Configuration C differ specifically in the presence or absence of RAI safeguards, allowing the performance differential between them to be attributed to RAI implementation in isolation from any community knowledge grounding. This design was chosen deliberately to test the strongest, cleanest version of the standalone-RAI hypothesis: if RAI safeguards improve effectiveness on their own, this comparison should detect it directly, without the confound of an interaction effect with TMKE that a design combining both interventions would introduce.

4.5. Performance Differentials Between Configuration A and Configuration C

Table 5 presents the paired-comparison results for the effectiveness metrics scored during simulation, testing whether Configuration C (RAI-only) differed significantly from Configuration A (baseline).

Table 5 Wilcoxon Signed-Rank Comparison of Configuration A and Configuration C

Metric	Direction of Change	Test Statistic	p-value	Effect Size (d)
BLEU-4 (content overlap)	Significant decline	W = 1301	p = .006	−0.12 (small)
ROUGE-L (content overlap)	Significant decline	t = −3.51	p < .001	−0.37 (small-to-medium)
Perplexity (fluency)	No significant change	–	p > .05	negligible
Protocol Enforcement Accuracy	No significant change	–	p = 1.000	not computable (n = 8)

Table 5 presents the most consequential and counter-intuitive finding in this paper: RAI implementation, on its own, did not improve any of the four metrics tested, and produced a statistically significant decline on two of them. BLEU-4 and ROUGE-L, both measuring content overlap with the correct reference, decreased significantly under RAI-only implementation, with small to small-to-medium effect sizes. This is a genuine finding of harm, not merely an absence of benefit, and it should not be minimised or reframed as a null result: Configuration C's content accuracy was measurably worse than Configuration A's unmodified baseline.

Perplexity showed no significant change, indicating RAI safeguards implemented at the system-prompt level did not measurably affect output fluency in either direction, unlike the fluency cost recorded for TMKE integration in the companion paper. Protocol Enforcement Accuracy also showed no significant difference between the two configurations, though this specific result carries a stated limitation: only two of the twenty-five prompts targeted restricted content, producing eight paired trials for this test, a sample size carrying very limited statistical power. Neither configuration achieved any correct restricted-content withholding in this comparison, both recording zero percent correct enforcement, a result more informative about what RAI alone cannot achieve without TMKE-based governance tagging than about a precise, well-powered effect size.

The decline recorded on BLEU-4 and ROUGE-L admits two distinct interpretations, and this paper states both directly rather than adopting only the more favourable one. The first, more measured reading is that RAI safeguards require a substantive knowledge source to act on before they can improve effectiveness; a transparency mechanism instructing a model to cite sources and explain its reasoning has, on this reading, little to act on when the model has no grounded content to cite in the first place, and may introduce generation overhead, additional instructions competing for the model's limited context and attention, without a corresponding benefit. The second, stronger reading is that RAI modules implemented without TMKE grounding actively interfered with content generation, not merely failing to help but measurably degrading it, plausibly because the additional system-prompt instructions increased the complexity of the generation task without providing the model any new, accurate content to draw on. This paper's evidence cannot adjudicate definitively between these two readings, and both are presented as live possibilities the findings support equally.

Read together, Table 3 and Table 5 support a conclusion that stands in genuine tension with itself, and this tension is the central finding of this paper rather than a limitation to be resolved away. At the level of survey-based perception, RAI is positively associated with LLM effectiveness, driven specifically by Transparency. At the level of independently measured, computationally scored output quality, RAI implemented alone produced no improvement and a measurable decline on two metrics. A researcher relying only on the correlational and regression evidence in Table 3 would reasonably conclude RAI safeguards improve LLM effectiveness; a researcher relying only on the simulation evidence in Table 5 would reach the opposite conclusion. Both findings are reported here as equally valid, methodologically independent results describing different dimensions of what “effectiveness” means, perceived trustworthiness versus measured output quality, rather than one being treated as correcting or superseding the other.

4.6. Reliability of the Measurement Instrument

The credibility of the survey-based results reported above depends on the reliability of the instrument used to measure RAI. This was established in three stages, a pilot administration, an expert content-validity review, and full-sample statistical confirmation. Table 6 reports the full-sample reliability coefficient underlying the correlation and regression results in Table 3.

Table 6 Reliability Statistics for the RAI Construct (N = 107)

Construct	Items	Cronbach's Alpha	Threshold Met
Responsible AI (RAI)	15	.939	Yes (> .70)

Table 6 shows a Cronbach's alpha of .939 for the RAI construct, well above the conventional .70 threshold, indicating strong internal consistency across the fifteen items composing the scale. This value was obtained on the full Phase Two sample of 107 respondents, following an earlier pilot with 13 participants in Kibish sub-county and a three-member expert panel review recording an overall content-validity mean of .86. This gives confidence that the sub-construct pattern identified in Table 3, Transparency driving the aggregate RAI effect while Fairness and Accountability did not reach individual significance, reflects a genuine substantive pattern in respondent perception rather than measurement noise attributable to an unreliable instrument.

5. Discussion: Implications for Theory and Design

The findings reported above extend, and in one important respect complicate, the theoretical foundations reviewed in Section 2. Socio-Technical Systems theory's central claim, that fairness cannot be assessed from an algorithm alone but requires the surrounding institutional and knowledge infrastructure to be considered together, is directly supported by the simulation finding in Table 5: RAI safeguards implemented as algorithmic properties alone, without accompanying community knowledge grounding, produced no measurable effectiveness gain and a measurable decline on two metrics. This is precisely the outcome STS theory would predict for an intervention that addresses only the technology half of a socio-technical system while leaving the surrounding knowledge infrastructure unaddressed.

This finding also extends the general RAI literature in a specific, testable direction. Fairness and inclusion safeguards have been found to raise user confidence in agricultural AI tools without necessarily altering underlying task performance [3]; the present study's survey-versus-simulation divergence reproduces and sharpens this exact pattern in a new domain, showing not only that RAI's effect on perception can diverge from its effect on measured performance, but that the divergence can run in genuinely opposite directions, positive on the survey measure, negative on two of the computational measures, rather than merely differing in magnitude. This is a stronger and more concerning divergence than prior literature had reported, since it implies that relying on user-perception measures alone to evaluate RAI interventions in similar systems risks concluding an intervention is beneficial when independently measured output quality shows the opposite.

The sub-construct finding, that Transparency alone drove the survey-based effect while Fairness and Accountability did not, also deserves theoretical attention. The translation of RAI principles into measurable system properties has been identified as a persistent challenge across the field [2]; the present finding suggests that among the three principles tested, Transparency may be the one currently best served by concrete, implementable mechanisms, source citation and plain-language rationale, that respondents can directly observe and evaluate in a system's output. Fairness and Accountability, by contrast, are harder for an individual respondent to verify directly from a single interaction with a system, which may explain why they did not register as significant individual predictors even though they remained conceptually and operationally present in the tested architecture throughout.

The elevated multicollinearity recorded among the three RAI sub-constructs also warrants a cautious reading rather than a dismissive one. A VIF range of 3.77 to 4.98 indicates substantial shared variance among Fairness, Accountability, and Transparency, consistent with respondents perceiving these three principles as closely related rather than sharply distinct constructs in practice, even though they were designed and measured as conceptually separate sub-constructs. This suggests Fairness and Accountability's non-significance in the regression may partly reflect this shared variance being absorbed by Transparency, the more dominant, correlated sibling, rather than definitive evidence that Fairness and Accountability contribute nothing to effectiveness on their own; a factor-analytic or structural equation approach explicitly modelling this shared variance, beyond the scope of this paper, would be needed to separate these possibilities with greater confidence.

For practitioners designing similar systems, three practical implications follow directly from this pattern. First, RAI safeguards should not be deployed as a standalone intervention with an expectation of measurable effectiveness gains; this study's evidence indicates that expectation is unsupported, and may in fact be counterproductive on content-accuracy metrics specifically. Second, Transparency mechanisms, source citation and plain-language rationale, should be prioritised over generic Fairness or Accountability commitments if the goal is measurable improvement in user-

perceived effectiveness, since this study found only Transparency reached individual significance as a predictor. Third, and most importantly, RAI's evaluation should never rely on user-perception survey measures alone; this study's central finding, a positive survey effect alongside a negative or null simulation effect, could not have been detected through either method in isolation, and any evaluation framework for RAI interventions in similar systems should include both a perception-based and an independently, computationally measured component.

This divergence between perceived and measured effectiveness also carries a governance implication beyond the immediate technical finding. If RAI safeguards can improve stated user confidence while measurably degrading content accuracy, an organisation evaluating a deployed system's RAI compliance purely through user satisfaction surveys, a common and administratively convenient evaluation method, risks certifying a system as responsibly designed while that system's actual outputs have become less factually reliable. This is precisely the kind of gap the CARE Principles' emphasis on genuine accountability, rather than the appearance of it, is intended to guard against, and this study's findings suggest that gap is not merely a theoretical risk but a measurable, empirically demonstrated possibility within the specific architecture tested here.

It is also worth situating this paper's findings against the companion evaluation of Community Knowledge Elements conducted on the same dataset. Where TMKE integration produced a real, if partial and uneven, effectiveness gain under simulation, RAI implemented alone produced no gain and a measurable cost. This asymmetry between the two interventions, tested under structurally identical conditions, twenty-five standardised prompts, four repetitions, the same base model, is itself an important empirical finding independent of either individual result: it suggests that, for this specific architecture and knowledge domain, community knowledge grounding is the intervention doing the substantive work of improving output quality, while RAI safeguards' primary demonstrated contribution in this study is to perceived trustworthiness rather than to measured content accuracy. Whether this asymmetry reflects a general property of RAI and CKE interventions, or is specific to how RAI was operationalised in this particular architecture through system-prompt instructions rather than a more deeply integrated mechanism, is a question this single study cannot resolve and is identified as a direction for future research.

This raises a specific, testable expectation carried into a separate companion paper: if RAI's measured effectiveness contribution depends on the presence of a substantive knowledge source to act on, as the more measured reading of the present findings suggests, then RAI implemented together with TMKE grounding should show a different, more favourable pattern than RAI implemented alone. That companion paper tests this expectation directly through a fourth configuration combining both interventions, and its findings should be read as a direct extension of, and a test against, the standalone RAI evidence reported in this paper.

In sum, the findings reported in this section answer the research question guiding this paper with a genuine and important qualification. Responsible AI principles, considered as a survey-based composite, do have a statistically significant, positive association with perceived LLM effectiveness, satisfying the condition for rejecting $H0_2$ at the level of that specific evidence. That same intervention, however, implemented independently of community knowledge grounding and evaluated through independent, computational measurement, produced no effectiveness gain and a measurable decline on two metrics. RAI, on the evidence reported here, is architecturally insufficient on its own to improve measured LLM effectiveness in this context, a finding with direct consequences for how Responsible AI is operationalised, and evaluated, in culturally aware LLM systems for indigenous knowledge preservation.

Two limitations bear directly on how confidently these findings should be read, and are stated here at the point they matter most rather than deferred to a general limitations note. First, the eight paired trials underlying the Protocol Enforcement Accuracy comparison in Table 5 carry very limited statistical power; the finding that neither configuration achieved correct restricted-content withholding is informative about what RAI alone did not achieve in this specific, small sample, but should not be read as a precisely estimated null effect. Second, RAI in this study was operationalised specifically through system-prompt-level instructions rather than through fine-tuning or a more deeply integrated architectural mechanism; whether a more deeply embedded implementation of Fairness, Accountability, and Transparency would show a different standalone effectiveness pattern than the one reported here is a question this specific operationalisation cannot answer, and readers generalising from this paper's findings to other RAI implementation strategies should bear this scope limitation in mind.

This paper's findings also bear on a broader methodological point relevant beyond the specific architecture tested here. The literature on Responsible AI evaluation has historically relied heavily on stated-principle audits and user-perception surveys, methods well suited to assessing whether a system's design intentions align with RAI values, but poorly suited to detecting the kind of divergence between perceived and measured effectiveness this study found. Successful RAI integration has been shown to require attention to how safeguards actually function once deployed, not

only how they are designed [1]; the present study's finding, that a positively perceived RAI intervention measurably degraded content accuracy on two metrics, is a concrete empirical illustration of exactly the gap between design intention and deployed function that concern anticipates, obtained through the kind of dual-method evaluation design, survey combined with independent computational scoring, that this literature has called for but not always implemented in practice.

Finally, the definitional extension of Fairness discussed in Section 2, from disparate treatment toward representational accuracy, warrants a brief return here in light of the findings. Fairness's non-significance as an individual predictor should not be read as evidence that representational fairness is unimportant for a system of this kind; rather, it suggests that the specific indicator used to operationalise Fairness in this study, a composite survey measure of respondents' perceived even-handedness in the system's treatment of Turkana knowledge relative to other knowledge systems, may not have been sensitive enough to detect a real effect obscured by its correlation with Transparency and Accountability. A more direct, behavioural measure of representational fairness, such as a systematic audit of specific outputs for asymmetric treatment across knowledge domains, independent of respondent self-report, is identified as a methodological improvement for future replications of this specific sub-construct comparison.

Taken as a whole, the evidence presented in this section resists a simple, single-sentence summary, and this resistance is itself the paper's most important methodological lesson. Any evaluation of Responsible AI's contribution to LLM effectiveness that relies on only one method, survey perception or computational scoring alone, would have reported a confidently wrong or dangerously incomplete conclusion. It is only by holding both forms of evidence in view simultaneously, and reporting their divergence honestly rather than resolving it in favour of the more convenient result, that this study was able to characterise RAI's standalone contribution accurately: real at the level of perception, absent or negative at the level of measured output quality.

The next section draws these findings together into a conclusion scoped specifically to what this paper's evidence supports, distinguishing clearly between what was established about RAI's perceived contribution and what was established, separately and with equal confidence, about its measured one.

6. Conclusion

This study established the effect of Responsible AI principles, operationalised through Fairness, Accountability, and Transparency, on the effectiveness of a culturally-aware Large Language Model applied to Turkana Traditional Medicine Knowledge preservation. RAI correlated moderately strongly with LLM Effectiveness and significantly predicted it in regression, an effect driven specifically by Transparency rather than Fairness or Accountability. Independent controlled simulation, however, found that RAI implemented alone, without community knowledge grounding, produced no significant effectiveness gain and a significant decline on two content-accuracy metrics. $H0_2$ was rejected in favour of $H1_2$ at the level of the survey-based model, but RAI's measured, computationally scored contribution was absent or negative, a divergence between perceived and measured effectiveness that is this paper's central and most consequential finding for the design of similar systems.

Recommendations

Developers of Responsible AI mechanisms for indigenous knowledge preservation systems should prioritise concrete, measurable transparency mechanisms, source citation and plain-language rationale, over general fairness or accountability commitments, since this study found these did not independently predict effectiveness. RAI safeguards should not be deployed as a standalone intervention with an expectation of improved output quality; this study's evidence indicates that expectation is unsupported and that RAI's demonstrated value in this architecture depends on combination with substantive knowledge grounding. Evaluators of similar systems should never rely on user-perception measures alone to assess RAI's contribution, since this study's central finding could not have been detected without an independent, computationally scored comparison alongside the survey-based one.

Future research should test whether RAI's standalone effect differs when operationalised through a more deeply integrated architectural mechanism than the system-prompt-level implementation used here, and should test directly whether combining RAI with community knowledge grounding resolves the decline recorded in this study, a question addressed in a separate companion paper examining the combined effect of both interventions on the same dataset.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] R. Kumar, S. Patel, and A. Okonkwo, "Critical considerations in implementing AI-powered educational tools: A study of teacher preparation and data ethics in developing nations," Educational Technology Research and Development, vol. 71, no. 2, pp. 189–207, 2023.

[2] O. Ozmen Garibay, B. Winslow, S. Andolina, M. Antona, A. Bodenschatz, C. Coursaris, G. Falco, S. M. Fiore, I. Garibay, K. Grieman, J. C. Havens, M. Jirotka, H. Kacorri, W. Karwowski, J. Kider, J. Konstan, S. Koon, M. Lopez-Gonzalez, I. Maifeld-Carucci, and S. McGregor, "Six human-centered artificial intelligence grand challenges," International Journal of Human-Computer Interaction, vol. 39, no. 3, pp. 391–437, 2023.

[3] C. Ozor, T. Adeyemi, and F. Nakato, "Fairness and inclusion safeguards in agricultural artificial intelligence tools: User confidence and task performance," Journal of Agricultural Informatics and AI, vol. 12, no. 1, pp. 55–71, 2025.

[4] B. Shneiderman, Human-Centered AI. Oxford, UK: Oxford University Press, 2022.

[5] K. Yilma, "Ethics of AI in Africa: Interrogating the role of Ubuntu and AI governance initiatives," Ethics and Information Technology, vol. 27, no. 2, art. 24, 2025.

[6] L. Floridi and J. Cowls, "A unified framework of five principles for AI in society," Harvard Data Science Review, vol. 1, no. 1, 2019.

[7] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389–399, 2019.

[8] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," Nature Machine Intelligence, vol. 1, no. 11, pp. 501–507, 2019.

[9] L. Weidinger et al., "Ethical and social risks of harm from language models," arXiv preprint arXiv:2112.04359, 2021.

[10] A. Birhane, "Algorithmic colonization of Africa," SCRIPTed, vol. 17, no. 2, pp. 389–409, 2020.

[11] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran, "Re-imagining algorithmic fairness in India and beyond," in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT), 2021, pp. 315–328.

[12] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT), 2021, pp. 610–623.

[13] T. B. Brown et al., "Language models are few-shot learners," in Proc. 34th Conf. Neural Information Processing Systems (NeurIPS), 2020, pp. 1877–1901.

[14] P. M. Podsakoff, S. B. MacKenzie, J. Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies," Journal of Applied Psychology, vol. 88, no. 5, pp. 879–903, 2003.

[15] R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.