



# The effect of community knowledge elements 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 Anselemo Peters Ikoha

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

World Journal of Advanced Engineering Technology and Sciences, 2026, 20(01), 375–386

Publication history: Received on 22 June 2026; revised on 27 July 2026; accepted on 30 July 2026

Article DOI: <https://doi.org/10.30574/wjaets.2026.20.1.0389>

## Abstract

Large Language Models (LLMs) increasingly encode and generate knowledge at scale, yet consistently underrepresent indigenous and community knowledge systems, a gap with acute consequences for critically endangered Intangible Cultural Heritage (ICH) such as Turkana Traditional Medicine Knowledge (TMK) in Kenya. Cultural grounding, Responsible AI, and Community Knowledge Elements have been studied in isolation, never combined and empirically justified within one architecture, leaving designers of culturally aware LLM systems without evidence for how Community Knowledge Elements specifically should be weighted. This paper reports an empirical evaluation of the effect of Community Knowledge Elements (CKEs), operationalised as Traditional Medicine Knowledge Elements (TMKEs), on the effectiveness of a culturally-aware LLM in preserving ICH. A sequential mixed-methods design combined a community case study with Design Science Research, drawing on 107 respondents from the Turkana community and domain experts, and 400 simulation outputs generated across two controlled LLM configurations. TMKE Integration Quality correlated strongly with LLM Effectiveness ( $r = .524$ ,  $p < .001$ ) and significantly predicted it in regression ( $R^2 = .290$ ,  $p < .001$ ), an effect driven specifically by Knowledge Integration Effectiveness ( $\beta = .389$ ,  $p = .020$ ). Under controlled simulation, TMKE integration produced significant gains in content overlap with the correct reference on BLEU-4 and ROUGE-L, at a significant cost to output fluency, and no significant improvement in literal Turkana-term retention. The findings establish that Community Knowledge Element integration produces a real, statistically supported, but partial and uneven effect on LLM effectiveness, with direct implications for how knowledge-grounded LLM architectures for indigenous heritage preservation should be designed and evaluated.

**Keywords:** Community Knowledge Elements; Large Language Models; Intangible Cultural Heritage; Traditional Medicine Knowledge; Knowledge Graph; Retrieval-Augmented Generation; Turkana

## 1. Introduction

Artificial Intelligence, particularly through Large Language Models, has transformative potential for natural language understanding and knowledge preservation. However, LLM application in safeguarding Intangible Cultural Heritage remains limited, since insufficient cultural grounding renders these systems poorly suited to preserving Traditional Medicine Knowledge, a critically endangered ICH domain held primarily by Emuron, traditional healers of the Turkana community of north-western Kenya. The transmission of TMK depends on a declining population of knowledge holders. Each loss of an Emuron without a trained successor represents a permanent, irreversible loss to the knowledge base, not a temporary gap, a pattern already documented among indigenous medicinal knowledge custodians in South Africa, where elders were found to no longer be alive to pass their knowledge on while younger generations showed diminishing interest in receiving it [6]. This capability rests on the transformer architecture underlying contemporary

\* Corresponding author: Paul Oduor Oyile

language models [7], demonstrated at scale in early bidirectional pretraining approaches [8], and on the broader shift toward large, general-purpose foundation models adapted to downstream tasks [10].

Grounding techniques capable of anchoring LLM outputs in external knowledge sources already existed in the literature, but had not been tested within a community-governed framework specific to indigenous medicinal knowledge. Traditional Medicine Knowledge Elements, comprising medicinal plant knowledge, remedy compositions, preparation and dosage protocols, contraindication knowledge, and practitioner governance conventions, had not been formalised into a structured, machine-usable input for an LLM architecture. This left an empirical gap: whether integrating Community Knowledge Elements measurably improves LLM effectiveness for ICH preservation had not been tested using reproducible, quantifiable metrics.

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 Community Knowledge Element integration specifically, distinct from Responsible AI safeguards, produces a measurable effectiveness gain, and if so, how large and how consistent that effect is across different dimensions of system performance.

This paper addresses that gap directly. The objective was to evaluate the effect of Community Knowledge Elements on the effectiveness of a culturally-aware LLM in preserving Intangible Cultural Heritage, guided by the research question: what is the effect of Community Knowledge Elements on the effectiveness of a culturally-aware LLM in preserving ICH? The corresponding hypotheses were stated formally as a null and alternative pair:  $H_{01}$ , that Community Knowledge Elements have no statistically significant direct effect on LLM effectiveness; and  $H_{11}$ , 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 Traditional Medicine Knowledge Element grounding.

The remainder of this paper is organised as follows. Section 2 reviews the theoretical and empirical literature bearing on Community Knowledge Elements and LLM effectiveness. Section 3 summarises the methodology used to test the stated hypothesis. Section 4 reports the preliminary and core findings, presented as a sequence of tables with accompanying interpretation. Section 5 concludes and offers recommendations arising directly from the findings.

---

## 2. Literature Review

Indigenous Knowledge Theory positions traditional knowledge systems as structurally distinct bodies of knowledge requiring representation on their own epistemological terms, rather than as supplementary data appended to an otherwise unmodified model. It has been argued that the epistemological foundations of mainstream AI development are rooted in a narrow rationalist tradition that systematically excludes non-Western ways of knowing, motivating an Indigenous-led research agenda built on the premise that Indigenous Knowledge systems should shape the conceptual architecture of AI rather than being retrofitted into it after the fact [5]. This theoretical position justifies treating Community Knowledge Elements as architectural design inputs rather than incidental training augmentation, and directly supports the decision in the present study to treat TMKEs as structural inputs to the LLM configuration tested, rather than as supplementary content appended after training.

Retrieval-augmented generation and related grounding techniques have been shown to reduce hallucination and improve factual accuracy in general-purpose LLM applications. A growing body of work demonstrates that anchoring generation in an external, retrievable knowledge source measurably improves factual consistency relative to relying on parametric knowledge alone [3]. This finding extends specifically to structured-output settings, where retrieval grounding has been shown to reduce fabricated content when the underlying knowledge source is well-curated [1]. However, the application of these techniques to structured indigenous knowledge, governed by community-specific access and representation conventions, remained empirically untested prior to this study; existing retrieval-augmented generation literature evaluates grounding against general-domain or enterprise knowledge bases, not against a knowledge source elicited through community participatory processes and encoding cultural governance rules as part of its structure. The retrieval-augmented generation approach used in this study follows the general architecture first proposed for combining a parametric language model with a non-parametric retrieval mechanism over an external knowledge source [9].

A participatory approach, informed by a decolonial epistemological lens, was adopted in this study as the methodology for formalising indigenous knowledge into digital structures, positioning knowledge-holders as co-designers rather than passive data sources. The reasoning behind this choice is that the process by which indigenous knowledge is

formalised for digital use matters as much as the technical grounding mechanism itself; knowledge extracted without community participation risks both inaccuracy and a breach of the governance conventions the knowledge is embedded within. This motivated the present study's use of participatory design workshops with Emuron during TMKE formalisation, rather than eliciting knowledge through interviews alone. This choice follows established guidance that indigenous data collection methodologies themselves, not only the resulting data, must be accountable to the communities they concern [15].

Indigenous Data Sovereignty scholarship further establishes that indigenous communities retain authority over how their knowledge is represented and used in digital systems. The CARE Principles for Indigenous Data Governance establish Collective Benefit, Authority to Control, Responsibility, and Ethics as minimum standards for working with indigenous data [2], standards that apply directly to any digitisation of Traditional Medicine Knowledge regardless of the technical grounding mechanism used. This principle was operationalised in the present study through tiered access permissions embedded directly in the underlying knowledge graph structure, ensuring that community-designated restricted knowledge remained inaccessible to the retrieval mechanism regardless of how a query was phrased. This scholarship traces to a broader movement establishing indigenous authority over data concerning indigenous peoples as a matter of self-determination, not only research ethics [12].

Empirically, the closest available precedent for combining structured community knowledge with an LLM interface is a community-governed AI workflow developed for the Hawaiian-medium KĀ'EO assessment programme [4], in which a document-grounded language model was deployed under an explicit governance framework combining human oversight and knowledge grounding. That study reported that grounding accelerated expert analysis while preserving linguistic and cultural authority, offering a working example of community knowledge integration functioning within a governed LLM system, though it did not report the kind of controlled, counterfactual comparison, with and without knowledge grounding, that would isolate the specific magnitude of the knowledge-grounding effect. This is precisely the gap the present study's Configuration A versus Configuration B comparison was designed to close.

Taken together, this literature establishes a strong theoretical case for expecting Community Knowledge Element integration to improve LLM effectiveness in an indigenous knowledge preservation context, grounded in Indigenous Knowledge Theory's insistence on structural rather than supplementary integration, retrieval-augmented generation's demonstrated general-purpose effectiveness, participatory design's methodological guidance for formalising the knowledge itself, and Indigenous Data Sovereignty's requirement that governance be embedded structurally. What this literature does not yet establish, and what the present study tests directly, is whether this theoretical case holds empirically when measured against reproducible, quantifiable effectiveness metrics in a specific indigenous medicinal knowledge context, and whether the resulting effect is uniform across the different dimensions, content accuracy, fluency, and literal terminological fidelity, that constitute LLM effectiveness for this purpose.

This distinction matters specifically because Community Knowledge Elements and Responsible AI principles, though frequently discussed together in the broader literature on culturally aware AI, are conceptually and architecturally separable interventions. RAI principles govern how an LLM behaves toward the knowledge it already has access to, through transparency, fairness, and accountability mechanisms, while Community Knowledge Elements govern what knowledge the LLM has access to in the first place. A system could plausibly implement strong RAI safeguards while still lacking the underlying community knowledge needed to generate accurate outputs, or conversely possess rich community knowledge grounding without any accompanying governance safeguard. This paper isolates the second of these two interventions, evaluating Community Knowledge Element integration on its own empirical merits, independent of the Responsible AI evaluation reported in a companion paper addressing the same dataset.

A further consideration bears directly on the specific mechanism by which TMKEs were formalised for this study. Structured elicitation, rather than passive corpus mining, was adopted as the formalisation approach, since medicinal and ritual knowledge of this kind is rarely documented in text form accessible to conventional web scraping or corpus-based extraction pipelines. This shifted the methodological burden toward direct, sustained engagement with knowledge-holders, treated in this study's design not as a constraint on scale but as a precondition for validity: knowledge elicited without sustained community engagement risks both factual inaccuracy and governance violations that downstream technical safeguards cannot fully correct for after the fact.

A fluency-accuracy trade-off was anticipated as a plausible outcome of this study's design, consistent with the general expectation that retrieval-augmented systems evaluated against domain-specific corpora often show content accuracy gains alongside some cost to output naturalness, particularly when the retrieved content includes vocabulary or phrasing patterns underrepresented in the base model's original training distribution, a pattern especially plausible for a base model such as the one used in this study, trained predominantly on high-resource languages with minimal

exposure to Turkana-language medicinal terminology. Whether this trade-off held, and at what magnitude, for Traditional Medicine Knowledge Elements specifically had not been previously reported and is addressed directly in the findings that follow.

### 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. Phase One formalised TMKEs through participatory design workshops with Emuron, producing the Turkana Medical Knowledge Graph (TMKG) and an annotated corpus of Traditional Medicine Knowledge, structured around five categories: medicinal plant knowledge, remedy knowledge, preparation and dosage knowledge, contraindication and safety knowledge, and practitioner governance knowledge. Phase Two comprised a structured survey administered to 107 respondents, an 87 percent response rate from a targeted sample of 123, drawn from Emuron (89.3 percent response rate), Liaison Officers (100 percent), and domain experts external to the Turkana community (75 percent), and a simulation exercise generating 400 outputs across four LLM configurations, of which two are the focus of this paper.

TMKE Integration Quality was measured through a 15-item survey scale spanning three sub-dimensions: Knowledge Representation, Governance Compliance, and Knowledge Integration Effectiveness. 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 for TMKE and LLM Effectiveness both exceeding .93. This threshold follows established psychometric convention for acceptable internal consistency in a newly developed scale [14].

Two of the four simulation configurations isolated the TMKE effect directly for the purposes of this paper: Configuration A, an unmodified baseline with no knowledge grounding, and Configuration B, retrieval-augmented generation grounded in the TMKG. Twenty-five standardised prompts, spanning plant-identification and ailment-treatment templates, 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 constructed by matching each prompt to its source TMKG entry; Perplexity, measuring output fluency, with higher values indicating lower fluency; and Cultural Term Retention Accuracy (CTRA), a strict, literal measure of whether specific Turkana-language terms drawn from a Lexical Database appeared in the output.

Pearson correlation and multiple regression tested the survey-based relationship between TMKE Integration Quality and LLM Effectiveness, with the three TMKE sub-dimensions entered simultaneously as predictors to identify which, if any, drove the aggregate effect. Normality testing confirmed survey composites were suitable for parametric analysis. The simulation metrics violated normality, so Wilcoxon signed-rank tests, a non-parametric procedure appropriate for paired, repeated-measures comparisons, tested the performance differential between Configuration A and Configuration B on each of the four metrics, with Cohen's *d* reported alongside each test to indicate effect magnitude. Because both TMKE Integration Quality and LLM Effectiveness were measured through the same self-report survey, the correlation reported below is read with awareness of common method variance as a potential source of inflation, a limitation inherent to single-source survey designs of this kind [13].

Two limitations of this design are noted directly, since they bear on the interpretation of the findings reported in Section 4. First, the TMKE Registry's content was elicited entirely from Emuron self-report, without independent triangulation against external botanical or ethnobotanical sources; this registry subsequently functioned as the ground truth for BLEU-4 and ROUGE-L scoring, so any error in the registry would propagate into the computational results. Second, Configuration B was built using parameter-efficient fine-tuning in addition to retrieval grounding, while Configuration A was not, meaning the reported effect reflects TMKE grounding as implemented through this specific fine-tuning approach, not TMKE content in complete isolation from the training methodology used to integrate it.

The study area was the Turkana community of north-western Kenya, and access was negotiated through introductory correspondence with community and county authorities prior to data collection, following the standard ethical clearance procedure required for research involving indigenous knowledge in this context. Research assistants, themselves Turkana natives pursuing computing-related postgraduate study, supported Emuron respondents in completing survey instruments, bridging both language and technical literacy gaps while translating items faithfully rather than paraphrasing their intent. This arrangement was adopted specifically because the technical vocabulary used

in the LLM Effectiveness scale, terms such as hallucination and semantic fidelity, required contextual explanation that a self-administered questionnaire could not reliably provide to respondents without a computing background.

Base model selection for the simulation strand was determined by a binary eligibility condition rather than a weighted trade-off among candidate models: only a model whose weights were openly available for local modification could support the parameter-efficient fine-tuning this study's architectural approach required. Llama 3 8B Instruct was the only candidate meeting this condition among the models compared, and was accordingly selected as the base model for both configurations reported in this paper, a constraint discussed further as a limitation on generalisability in Section 5.

## 4. Results and Discussion

This section reports the preliminary and core findings bearing on the effect of Community Knowledge Elements on LLM effectiveness. Findings are presented as a sequence of five 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

Before reporting the substantive relationship between Community Knowledge Elements and LLM Effectiveness, the composition and response rate of the sample from which this relationship was estimated are reported, since the credibility of any correlation or regression result depends directly on the adequacy of the sample underlying it. 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 Liaison Officer stratum achieved a full response rate, consistent with their engagement through an existing institutional channel, while the Emuron stratum, engaged through community relationships built during Phase One knowledge elicitation, achieved 89.3 percent. The domain expert stratum recorded the lowest rate at 75.0 percent, consistent with a professionally engaged, time-constrained population recruited through direct outreach rather than an existing community relationship. This lower rate still exceeded the generally accepted minimum for survey research and is noted here as a property of this specific stratum rather than a threat to the overall analysis, since Emuron, the primary holders of the Traditional Medicine Knowledge this study centres, achieved a substantially higher rate. The resulting sample of 107 respondents is the basis for every correlation, regression, and descriptive statistic reported in the remainder of this section.

### 4.2. Descriptive Statistics for TMKE Sub-Constructs

Before testing the hypothesised relationship formally, the level of agreement respondents expressed toward each of the three TMKE sub-constructs was examined descriptively, since the relative strength of each sub-construct provides an early indication of which dimension of Community Knowledge Element integration respondents perceived most strongly. Table 2 presents composite agreement levels for Knowledge Representation, Governance Compliance, and Knowledge Integration Effectiveness.

**Table 2** Descriptive Statistics for TMKE Sub-Constructs

| Sub-Construct                       | Items | Agreement Range | Notable Finding                         |
|-------------------------------------|-------|-----------------|-----------------------------------------|
| Knowledge Representation            | 5     | 42.1% – 51.4%   | Moderate, consistent agreement          |
| Governance Compliance               | 5     | 44.8% – 53.2%   | Moderate, consistent agreement          |
| Knowledge Integration Effectiveness | 5     | 47.6% – 59.8%   | Strongest single TMKE item in the study |

Table 2 shows that all three TMKE sub-constructs recorded moderate agreement levels, with no sub-construct reaching a clearly dominant level of consensus, consistent with a generally cautious but not sceptical respondent orientation toward the TMKE construct as a whole. Knowledge Integration Effectiveness recorded the highest individual item agreement of the three, reaching 59.8 percent on its strongest item, a result that anticipates the regression finding reported in Table 3, where this same sub-construct emerges as the only statistically significant individual predictor of LLM Effectiveness among the three. This descriptive pattern, the sub-construct respondents agreed with most strongly also being the sub-construct that predicted effectiveness most strongly in the subsequent inferential analysis, provides a first, non-inferential signal of internal consistency between how respondents perceived TMKE integration and how that integration actually related to system effectiveness, though this descriptive pattern alone cannot establish the statistical relationship formalised in the next subsection.

#### 4.3. Correlation and Regression: Testing $H_{01}/H_{11}$

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

**Table 3** Correlation and Regression of TMKE Integration Quality on LLM Effectiveness

| Statistic                                     | Value | Significance                   |
|-----------------------------------------------|-------|--------------------------------|
| Pearson correlation ( $r$ )                   | .524  | $p < .001$                     |
| Regression $R^2$                              | .290  | $F(3, 103) = 14.047, p < .001$ |
| $\beta$ – Knowledge Representation            | .142  | $p = .251$ (not significant)   |
| $\beta$ – Governance Compliance               | -.017 | $p = .911$ (not significant)   |
| $\beta$ – Knowledge Integration Effectiveness | .389  | $p = .020$ (significant)       |

Table 3 shows a strong, statistically significant positive correlation between TMKE Integration Quality and LLM Effectiveness ( $r = .524, p < .001$ ), classified as a moderately strong relationship under conventional correlation-strength bands. The regression model explained 29.0 percent of the variance in LLM Effectiveness ( $R^2 = .290, p < .001$ ), a substantial share for a single composite predictor entered alongside its own sub-dimensions. On this basis,  $H_{01}$  was rejected in favour of  $H_{11}$ : Community Knowledge Elements have a statistically significant, positive direct effect on LLM effectiveness.

The sub-construct breakdown, however, complicates a simple reading of this result. Of the three TMKE sub-dimensions entered simultaneously, only Knowledge Integration Effectiveness reached individual statistical significance ( $\beta = .389, p = .020$ ). Knowledge Representation, the completeness and structural accuracy of how medicinal knowledge was encoded in the TMKG, did not reach significance individually ( $p = .251$ ), nor did Governance Compliance, the extent to which access-restriction conventions were respected ( $p = .911$ , with a near-zero, slightly negative coefficient). This pattern indicates that the aggregate TMKE effect established in the correlation and overall regression is not evenly distributed across the construct's three dimensions; it is concentrated specifically in how effectively knowledge was integrated into the generation process, not in how completely it was represented in the underlying graph or how strictly governance rules were observed.

This finding carries a direct theoretical implication consistent with Indigenous Knowledge Theory's emphasis on structural integration over passive representation. A knowledge graph can be complete and accurate in its content, satisfying Knowledge Representation, without that content being effectively surfaced and applied during generation,

the specific function Knowledge Integration Effectiveness measures. The present finding suggests that respondents, and by extension the effectiveness outcomes this study measured, were more sensitive to whether TMKE content was actually and correctly used in system outputs than to whether the underlying knowledge base was comprehensive. This has a practical design implication: system architects seeking to maximise the TMKE effect on effectiveness should prioritise the integration and retrieval mechanism, ensuring stored knowledge reaches generation reliably, over expanding the raw completeness of the knowledge base alone, though both remain necessary preconditions for a functioning system.

Governance Compliance's non-significance merits a further, more cautious reading than a simple absence of effect. Multicollinearity diagnostics computed alongside this regression showed elevated Tolerance and VIF values among the three sub-constructs, consistent with a suppression pattern in which shared variance between correlated sub-constructs masks the individual predictive contribution of one once a more dominant, correlated sibling is accounted for. This diagnosis rests on the VIF evidence rather than a formal suppression test isolating the specific mechanism, and is presented as a plausible explanation rather than a settled one; Governance Compliance's true individual contribution to effectiveness may be larger than this specific regression coefficient suggests, masked by its correlation with Knowledge Integration Effectiveness rather than genuinely absent.

#### 4.4. Simulation Design for Independent Corroboration

Survey-based correlation and regression establish a statistical association, but cannot by themselves demonstrate that TMKE integration causes a measurable change in system output quality, since respondents were rating their perceptions of a deployed system rather than directly comparing controlled variants of it. To corroborate the survey finding independently, a controlled simulation compared LLM outputs generated with and without TMKE grounding, holding every other factor constant. Table 4 summarises the two configurations compared in this paper.

**Table 4** Simulation Configurations Compared

| Configuration       | TMKE Grounding | Description                                                                                                          | Outputs Generated |
|---------------------|----------------|----------------------------------------------------------------------------------------------------------------------|-------------------|
| A (Baseline)        | Absent         | Unmodified Llama 3 8B Instruct, no knowledge grounding                                                               | 100               |
| B (TMKE-Integrated) | Present        | Retrieval-augmented generation grounded in the Turkana Medical Knowledge Graph, with parameter-efficient fine-tuning | 100               |

Table 4 shows that Configuration A and Configuration B differ specifically in the presence or absence of TMKE grounding, allowing the performance differential between them to be attributed, with the qualification noted in Section 3, to the TMKE integration itself. Twenty-five standardised prompts were generated four times under each configuration, producing 100 outputs per configuration and 88 matched pairs after excluding three prompts whose general, governance-oriented phrasing lacked a single traceable ground-truth entry in the Knowledge Graph for content-overlap scoring purposes. This design allowed each of the four effectiveness metrics reported in Table 5 to be compared directly between the two configurations using a paired, repeated-measures statistical procedure, providing a form of evidence structurally independent of the survey data underlying Table 3, since simulation outputs were scored computationally against objective linguistic and lexical criteria rather than through respondent perception.

#### 4.5. Performance Differentials Between Configuration A and Configuration B

Table 5 presents the paired-comparison results for all four effectiveness metrics scored during simulation, testing whether Configuration B (TMKE-integrated) differed significantly from Configuration A (baseline) on each dimension of output quality.

**Table 5** Wilcoxon Signed-Rank Comparison of Configuration A and Configuration B

| Metric                               | Direction of Change   | Test Statistic | p-value  | Effect Size (d)        |
|--------------------------------------|-----------------------|----------------|----------|------------------------|
| BLEU-4 (content overlap)             | Significant increase  | W = 306        | p < .001 | 0.63 (medium)          |
| ROUGE-L (content overlap)            | Significant increase  | W = 1423       | p = .026 | 0.40 (small-to-medium) |
| Perplexity (fluency; higher = worse) | Significant increase  | W = 1854       | p = .021 | 0.19 (small)           |
| CTRA (literal term retention)        | No significant change | –              | p = .792 | 0.07 (negligible)      |

Table 5 shows a pattern that is genuinely mixed rather than uniformly favourable to TMKE integration, and each metric is interpreted in turn below before a synthesised reading is offered.

BLEU-4 and ROUGE-L, both measuring content overlap between the generated output and the ground-truth reference drawn from the Knowledge Graph, increased significantly under TMKE integration, with a medium effect size on BLEU-4 and a small-to-medium effect on ROUGE-L. This is the clearest evidence in this paper that TMKE grounding improved the factual, content-level accuracy of generated outputs: Configuration B's outputs more closely matched the correct medicinal information than Configuration A's did, a result directly consistent with the retrieval-augmented generation literature reviewed in Section 2, which established that anchoring generation in an external knowledge source improves factual consistency relative to relying on parametric knowledge alone [1], [3]. This finding extends that general literature specifically to a structured indigenous medicinal knowledge context for the first time.

Perplexity increased significantly under TMKE integration, and because a higher Perplexity value indicates lower fluency, this result represents a genuine cost, not a benefit, of TMKE integration. This finding is theoretically important and should not be minimised: it indicates that grounding generation in Turkana-language medicinal content measurably reduced the fluency of the base model's output, plausibly because the base model's original training distribution contained minimal exposure to this vocabulary and phrasing, so retrieved content, however factually accurate, was harder for the model to integrate smoothly into fluent text. This is a real trade-off that architects of similar systems for other low-resource indigenous languages should budget for explicitly rather than assume away; content grounding and output fluency were not jointly optimised by the intervention tested here, and improving one measurably cost the other.

CTRA showed no statistically significant difference between the two configurations, with a negligible effect size. This is the most theoretically consequential null result in this paper. CTRA is a strict, literal measure, counting only the exact presence of specific Turkana-language terms drawn from a Lexical Database, with no credit for paraphrase or partial recall, so some portion of this null result is attributable to the strictness of the scoring method itself rather than a true absence of cultural grounding in the response as a whole. Even accounting for this, the finding carries real weight for any claim that TMKE integration alone achieves cultural-linguistic fidelity: TMKE grounding measurably improved general content accuracy, as shown by BLEU-4 and ROUGE-L, without producing a corresponding, measurable improvement in whether the specific Turkana-language terminology expected of a culturally grounded output actually appeared in that output. This tension, improved general accuracy without improved literal cultural-term fidelity, is a genuine limitation on how far TMKE integration alone, absent an accompanying mechanism specifically targeting terminology retention, can be said to achieve cultural preservation in the fullest sense the term implies.

Read together, Table 3 and Table 5 support a single, consistent conclusion, stated here as the central finding of this paper: Community Knowledge Element integration produces a real, statistically supported, but partial and uneven effect on LLM effectiveness. The effect is partial because it is concentrated in one TMKE sub-dimension, Knowledge Integration Effectiveness, rather than distributed evenly across all three. It is uneven because, at the level of computational output quality, it improves content accuracy significantly while simultaneously reducing fluency significantly and leaving literal cultural-term retention statistically unchanged. A system designer reading only the aggregate correlation and  $R^2$  in Table 3 would reasonably conclude TMKE integration is straightforwardly beneficial; a designer reading the full pattern in Table 5 would correctly conclude that TMKE integration involves real trade-offs that must be actively managed, not a costless improvement across every dimension of system performance.

Two limitations bear directly on how confidently these findings should be read, both stated in Section 3 and revisited here at the point they matter most. First, the ground-truth reference used for BLEU-4 and ROUGE-L scoring was constructed from the TMKE Registry itself, content elicited entirely from Emuron self-report without independent triangulation against external botanical or ethnobotanical sources. Any error in that registry would propagate directly into what this study treated as a correct answer, meaning the content-accuracy gains reported here should be

understood as gains relative to the community's own account of correct medicinal knowledge, not independently externally verified fact. Second, Configuration B combined TMKE grounding with parameter-efficient fine-tuning, a training-methodology difference absent from Configuration A, so the reported effect reflects TMKE integration as implemented through this specific fine-tuning approach, not TMKE content in complete isolation from the method used to integrate it. Both limitations bound the specificity of the causal claim this paper can support, while leaving the core statistical pattern, a real, partial, and uneven effect, intact.

#### 4.6. Discussion: Implications for Theory and Design

The findings reported above extend, and in one respect complicate, the theoretical foundations reviewed in Section 2. Indigenous Knowledge Theory's insistence on structural integration over supplementary representation is directly supported by the sub-construct finding in Table 3: it was specifically Knowledge Integration Effectiveness, the dimension most closely tied to how knowledge is functionally used during generation rather than how completely it is stored, that predicted effectiveness. This is precisely the distinction drawn in the literature between indigenous knowledge shaping a system's conceptual architecture and indigenous knowledge being retrofitted onto an unmodified one [5]; the present findings suggest that even within a study explicitly designed around the former principle, the specific mechanism of integration, not merely its presence, is what determines the resulting effectiveness gain.

The comparison with the KĀ'EO programme [4] is instructive precisely because of where the present findings diverge from that precedent. That account reported that community-governed knowledge grounding accelerated expert analysis while preserving linguistic and cultural authority, without reporting a controlled comparison isolating the magnitude or evenness of that effect across different output dimensions. The present study's controlled design reveals what such an uncontrolled account cannot: that a grounding intervention can simultaneously improve expert-facing content accuracy while leaving literal linguistic fidelity, in this study's terms, Cultural Term Retention Accuracy, statistically unchanged. This suggests that reported successes in the literature for culturally grounded AI systems may understate the internal unevenness of the effect achieved, since qualitative or expert-satisfaction-based evaluations, of the kind the KĀ'EO account draws on, are less likely to surface a null result on a narrow, strict metric like CTRA than a quantitative, multi-metric design of the kind used in this paper.

The fluency cost recorded on Perplexity also warrants theoretical attention beyond its immediate practical implication. A plausible explanation is that content grounding in vocabulary underrepresented in a base model's training distribution measurably taxes that model's fluency, even as it improves the accuracy of what is being said. For Turkana, a language with minimal digital footprint relative to the high-resource languages dominating contemporary LLM training corpora, this cost is unlikely to be specific to the base model used in this study; it plausibly reflects a structural property of applying retrieval-augmented generation to genuinely low-resource linguistic content generally, a proposition this single-model study cannot confirm but which future work comparing multiple base models could test directly. Turkana sits among the languages Joshi et al. classify as having minimal or no dedicated NLP resources, a category shown to be systematically underserved by contemporary language model training pipelines [11].

Finally, the CTRA null result bears on Indigenous Data Sovereignty considerations in a way not immediately obvious from the statistic alone. If TMKE integration improves general content accuracy without reliably surfacing the specific Turkana-language terminology community members would expect to see, a deployed system might satisfy an external, content-accuracy-based evaluation of its usefulness while still failing a community-defined standard of what it means for output to be genuinely and recognisably Turkana. This distinction, between accuracy as judged by an external content-matching metric and authenticity as judged by literal linguistic presence, is precisely the kind of gap the CARE Principles' emphasis on community authority over representation is meant to guard against, and this study's findings suggest that gap can persist even in a system explicitly designed with community knowledge grounding at its core, unless language-specific retention is measured and optimised for directly rather than assumed to follow automatically from content-level grounding.

#### 4.7. Reliability of the Measurement Instruments

The credibility of every result reported above depends on the reliability of the survey instruments used to measure TMKE Integration Quality and LLM Effectiveness. 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 coefficients underlying the correlation and regression results in Table 3.

**Table 6** Reliability Statistics for Core Constructs (N = 107)

| Construct                | Items | Cronbach's Alpha | Threshold Met |
|--------------------------|-------|------------------|---------------|
| TMKE Integration Quality | 15    | .933             | Yes (> .70)   |
| LLM Effectiveness        | 25    | .963             | Yes (> .70)   |

Table 6 shows Cronbach's alpha values well above the conventional .70 threshold for both constructs central to this paper, indicating strong internal consistency across the items composing each scale. These values were obtained on the full Phase Two sample of 107 respondents, following an earlier pilot with 13 participants in Kibish sub-county, deliberately excluded from the main sample, and a three-member expert panel review of face and content validity that recorded an overall mean of .86. This three-stage process gives confidence that the correlation and regression results reported in Table 3 rest on instruments that reliably and validly measured the constructs they were designed to capture, and that the sub-construct pattern identified there, Knowledge Integration Effectiveness driving the aggregate TMKE effect while Knowledge Representation and Governance Compliance did not reach individual significance, reflects a genuine substantive pattern in the data rather than measurement noise attributable to an unreliable instrument.

Taken as a whole, the evidence presented in this section supports a single, carefully qualified conclusion. Community Knowledge Elements, measured reliably and validly across a well-powered sample, exert a real and statistically significant effect on the effectiveness of a culturally aware Large Language Model, corroborated independently through controlled simulation. That effect is neither uniform across the TMKE construct's three sub-dimensions nor uniform across the dimensions of output quality it improves. Architects and researchers building on this finding should treat Community Knowledge Element integration as a necessary but insufficient design principle on its own, one that measurably improves content accuracy while introducing a fluency cost and leaving literal cultural terminology retention unresolved, three distinct outcomes requiring three distinct design responses rather than a single, undifferentiated claim of improvement.

For practitioners designing similar systems for other indigenous or low-resource knowledge domains, three practical implications follow directly from the pattern reported above, each corresponding to one of the three outcomes just named. First, since Knowledge Integration Effectiveness rather than Knowledge Representation drove the aggregate TMKE effect, development effort is better directed toward the retrieval and integration mechanism, ensuring stored knowledge is reliably surfaced and correctly applied during generation, than toward maximising the raw size or completeness of the underlying knowledge base alone; a smaller, well-integrated knowledge source may outperform a larger, poorly integrated one on the effectiveness dimensions this study measured. Second, the fluency cost recorded on Perplexity should be budgeted for explicitly in any deployment plan, whether through additional fine-tuning on domain-appropriate fluent examples, through post-generation fluency correction, or through transparent communication to end users that grounded outputs may read less smoothly than an ungrounded model's fluent but potentially inaccurate output, a trade-off users should be equipped to weigh rather than have obscured from them. Third, given that CTRA showed no significant improvement under TMKE integration alone, systems aiming specifically for cultural linguistic fidelity, not only factual accuracy, should not assume this outcome follows automatically from knowledge grounding and should consider a dedicated mechanism, such as explicit term-preservation prompting or post-generation terminology verification against a lexical database, layered on top of retrieval grounding rather than relying on retrieval grounding to achieve it unassisted.

It is also worth situating these findings against the broader retrieval-augmented generation literature reviewed in Section 2, since the pattern reported here both confirms and refines that literature's general expectations. Retrieval grounding has been established to generally improve factual consistency [1], [3], a finding this study's BLEU-4 and ROUGE-L results directly reproduce in a new, indigenous-language context. Neither of those general studies, however, reported a fluency cost of the magnitude observed here, nor did they measure anything resembling the strict, literal terminology-retention construct captured by CTRA, since neither general-domain retrieval-augmented generation nor enterprise document grounding typically involves a language as underrepresented in training data as Turkana. This suggests the fluency cost and the terminology-retention null result reported in this paper may be specific consequences of applying retrieval grounding to a genuinely low-resource language, rather than general properties of retrieval-augmented generation as such, a distinction with direct implications for how transferable these findings are: architects working with higher-resource indigenous languages, where the base model's training exposure is less minimal, might reasonably expect a smaller fluency cost, though this expectation would itself require empirical testing rather than being assumed from the present single-language study.

In sum, the findings reported in this section answer the research question guiding this paper directly and with appropriate nuance. Community Knowledge Elements do have a statistically significant, positive effect on the effectiveness of a culturally aware Large Language Model, confirmed independently through survey-based inference and controlled simulation, satisfying the condition for rejecting  $H_{01}$ . That effect, however, is neither large enough nor even enough across every dimension of effectiveness to be reported without qualification. The next section draws these threads together into a conclusion scoped specifically to what this paper's evidence supports.

## 5. Conclusion

This study evaluated the effect of Community Knowledge Elements, operationalised as Traditional Medicine Knowledge Elements, on the effectiveness of a culturally-aware Large Language Model applied to Turkana Traditional Medicine Knowledge preservation. TMKE Integration Quality correlated strongly with LLM Effectiveness and significantly predicted it in regression, an effect driven specifically by Knowledge Integration Effectiveness rather than Knowledge Representation or Governance Compliance. Independent controlled simulation corroborated this finding, showing significant gains in content overlap under TMKE grounding, at a significant cost to output fluency, and no significant improvement in literal cultural-term retention.  $H_{01}$  was rejected in favour of  $H_{11}$ : Community Knowledge Elements have a statistically significant, positive effect on LLM effectiveness, though that effect is partial and unevenly distributed across the dimensions of system performance it improves, a finding with direct implications for how such systems should be designed and evaluated.

### *Recommendations*

Developers of culturally grounded LLM systems should prioritise the knowledge integration mechanism, ensuring stored community knowledge is reliably surfaced and correctly applied during generation, over expanding the raw completeness of the underlying knowledge base alone, since it was integration effectiveness specifically, not representational completeness, that predicted effectiveness in this study. System evaluators should report fluency and literal cultural-term retention alongside content overlap metrics, since these findings show the three do not necessarily move together, and a system reported as successful on content accuracy alone may still be failing on fluency or terminological fidelity. Community-governed knowledge graph co-design should be treated as a prerequisite for Traditional Medicine Knowledge Element formalisation, not an optional participatory add-on.

Future research should independently triangulate community-elicited knowledge graph content against external botanical or ethnobotanical sources before relying on it as ground truth for computational scoring, and should test whether the fluency cost and terminology-retention null result reported here persist across base models with greater training exposure to the target indigenous language, since the present study's reliance on a single base model leaves this an open, empirically testable question rather than an established property of retrieval-augmented generation applied to indigenous knowledge domains generally.

## Compliance with ethical standards

### *Disclosure of conflict of interest*

No conflict of interest to be disclosed.

## References

- [1] O. Ayala and P. Bechard, "Reducing hallucination in structured outputs via retrieval-augmented generation," in Proc. NAACL Industry Track, 2024, pp. 228–238.
- [2] S. C. Carroll, I. Garba, O. L. Figueroa-Rodriguez, J. Holbrook, R. Lovett, S. Materechera, M. Parsons, K. Raseroka, D. Rodriguez-Lonebear, R. Rowe, R. Sara, J. D. Walker, J. Anderson, and M. Hudson, "The CARE Principles for Indigenous Data Governance," Data Science Journal, vol. 19, no. 1, p. 43, 2020.
- [3] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2023.
- [4] W. Kūkea-Shultz and K. Brockmann, "Community-governed artificial intelligence for indigenous language education: The KĀ'EO assessment programme," Journal of Indigenous Language Technology, vol. 3, no. 1, pp. 14–31, 2025.

- [5] B. Lewis, N. Abdilla, N. Arista, K. Baker, S. Benesiinaabandan, M. Brown, M. Cheung, M. Coleman, A. Cordes, S. Davison, A. Duncan, S. Garzon, D. Harrell, P. Kealiikanakaoleohaililani, N. Kelleher, S. Kite, O. L. Lagon, J. Leigh, M. Levesque, and J. Walker, "Indigenous protocol and artificial intelligence position paper," Indigenous AI Working Group, 2024.
- [6] N. F. Mzimela and S. Moyo, "Custodianship and continuity: Indigenous medicinal knowledge transmission in South Africa," *Journal of Ethnobiology and Ethnomedicine*, vol. 20, no. 1, pp. 1–15, 2024.
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Conf. Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
- [8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. 34th Conf. Neural Information Processing Systems (NeurIPS)*, 2020, pp. 9459–9474.
- [10] R. Bommasani et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.
- [11] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 6282–6293.
- [12] T. Kukutai and J. Taylor, Eds., *Indigenous Data Sovereignty: Toward an Agenda*. Canberra, Australia: ANU Press, 2016.
- [13] 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.
- [14] J. C. Nunnally, *Psychometric Theory*, 2nd ed. New York, NY: McGraw-Hill, 1978.
- [15] M. Walter and D. Suina, "Indigenous data, indigenous methodologies and indigenous data sovereignty," *International Journal of Social Research Methodology*, vol. 22, no. 3, pp. 233–243, 2019.