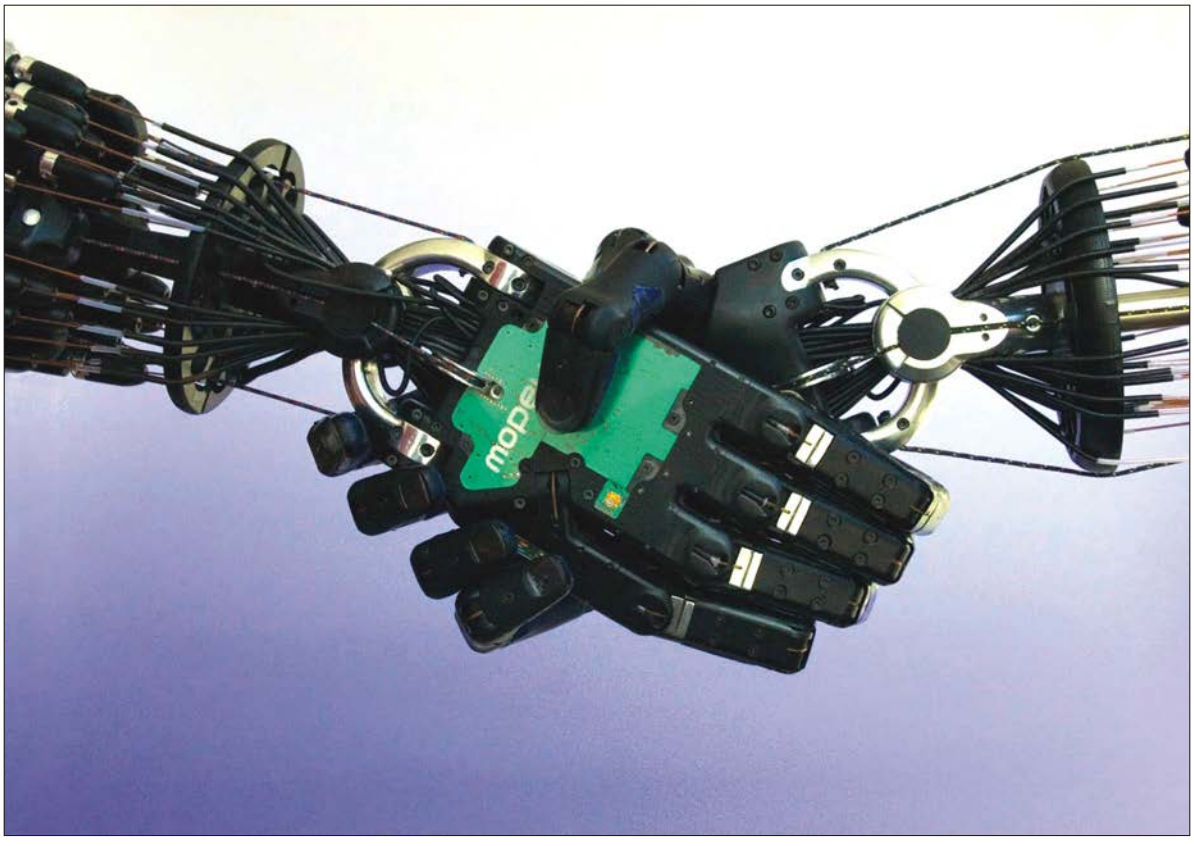

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Artificial Intelligence

ISSN 1994-5450
DOI: 10.3923/jai.2026.60.71

Research Article
Epistemic Fairness in Cultural AI Systems: Formal Guarantees and Empirical Validation in Awori and Ogu Knowledge Corpora

[1]Oseni Afisi, [2,3]Olusola Olabanjo and [4]Ashiribo Wusu

[1]Department of Philosophy, Lagos State University, Lagos, Nigeria
[2]Center for Equitable AI and Machine Learning Systems, Morgan State University, Baltimore, United States
[3]Department of Computer Science, Lagos State University, Lagos, Nigeria
[4]Department of Mathematics, Lagos State University, Lagos, Nigeria

Abstract

Background and Objective: The increasing deployment of generative artificial intelligence in digital humanities raises urgent concerns regarding epistemic distortion, cultural misrepresentation and asymmetrical error across knowledge communities. Existing fairness frameworks in machine learning primarily emphasize demographic parity or statistical balance, but they fail to adequately address disparities in interpretive accuracy, faithfulness and cultural sensitivity in knowledge-driven systems. This study aims to develop a formal framework for epistemic fairness in cultural AI systems and to validate it using Awori and Ogu indigenous knowledge corpora from Lagos, Nigeria. **Materials and Methods:** Retrieval-augmented generation is modeled as a governed optimization problem under constraints of faithfulness, cultural fidelity and metadata-based access control. Epistemic fairness is defined as bounded disparity in group-conditional faithfulness and expert-evaluated interpretive accuracy across cultural communities. The study derives theoretical bounds linking retrieval recall, constrained generation and abstention thresholds to expected faithfulness and hallucination rates. Empirical evaluation is conducted across question answering, narrative synthesis and summarization tasks. **Results:** Domain-adapted retrieval and fairness-constrained generation significantly reduced hallucination rates while minimizing cross-community epistemic disparity. Expert-based evaluation demonstrated that improvements in statistical parity correspond to meaningful gains in cultural fidelity rather than superficial fluency. Stability conditions confirmed that fairness disparities remain bounded under the proposed framework. **Conclusion:** The findings establish epistemic fairness as a measurable and enforceable property of cultural AI systems. The proposed framework provides a principled approach for responsible deployment of generative models in indigenous digital humanities, bridging algorithmic fairness theory and cultural knowledge preservation while promoting epistemically equitable and governance-aware AI systems.

Key words: Retrieval-augmented generation, BERT, digital humanities, indigenous knowledge, cultural preservation, Awori, Ogu, Lagos

Citation: Afisi, O., O. Olabanjo and A. Wusu, 2026. Epistemic fairness in cultural ai systems: Formal guarantees and empirical validation in Awori and Ogu knowledge corpora. J. Artif. Intel., 19: 60-71.

Corresponding Author: Olusola Olabanjo, Center for Equitable AI and Machine Learning Systems, Morgan State University, Baltimore, United States

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

Funding: Funding for this research was provided by the Tertiary Education Trust Fund, TETFUND Nigeria under grant number LASU/VC/DSI/RP/25/007.

INTRODUCTION

The rapid integration of generative artificial intelligence into the digital humanities has fundamentally transformed how cultural knowledge is accessed, interpreted and disseminated[1,2]. Advances in Retrieval-Augmented Generation (RAG) systems[3], Large Language Models (LLMs) and knowledge-grounded conversational agents have enabled new forms of interaction with historical narratives, archival materials and educational content[4,5]. These technologies hold considerable promise for expanding access to underrepresented cultural materials and democratizing knowledge production. However, their deployment in culturally sensitive contexts introduces significant risks. In particular, when applied to indigenous and community-governed knowledge systems, generative models may distort meaning, hallucinate unsupported claims, flatten interpretive nuance or disproportionately misrepresent specific knowledge traditions[6,7]. Such failures extend beyond technical inaccuracies and constitute forms of epistemic harm, affecting how knowledge is represented, validated and transmitted.

Existing research on algorithmic fairness has largely focused on statistical notions such as demographic parity, equalized odds and calibration across protected groups[8]. These frameworks were developed primarily for predictive classification systems operating on structured datasets, where fairness is evaluated in terms of outcome distributions. However, cultural AI systems differ fundamentally from these settings. Rather than predicting labels, they generate narratives, explanations and interpretations. Consequently, fairness in such systems must extend beyond statistical parity to account for disparities in faithfulness, interpretive accuracy and cultural sensitivity across knowledge communities. Standard fairness metrics are therefore insufficient to capture the epistemic dimensions of generative outputs. What is required is a notion of epistemic fairness: A condition in which different cultural knowledge groups receive equitable levels of grounding, interpretive integrity and representational fidelity[9].

Despite growing interest in improving the factual grounding of generative models through retrieval mechanisms, existing RAG research has predominantly focused on benchmark datasets derived from encyclopedic or web-scale corpora[10]. These settings do not adequately reflect the complexity of culturally heterogeneous and governance-sensitive knowledge systems, where errors may result in epistemic distortion rather than simple factual inconsistency. Furthermore, current faithfulness metrics are rarely evaluated across distinct knowledge communities and thus fail to capture disparities in how different cultural narratives are represented. At the same time, while algorithmic fairness research has advanced significantly in supervised learning contexts[11-15], it has not been sufficiently extended to generative systems, where harms may arise from narrative imbalance, interpretive bias or unsupported claims[16,17]. In parallel, digital humanities scholarship emphasizes the importance of provenance, contextual integrity and ethical stewardship in the digitization of cultural archives[18,19], particularly for indigenous knowledge systems governed by norms of custodianship and interpretive authority[20-22]. However, existing computational approaches often lack formal mechanisms for enforcing governance constraints or ensuring balanced epistemic treatment across communities. Moreover, philosophical accounts of epistemic injustice show the risks of credibility deficits and hermeneutical marginalization in sociotechnical systems[23], yet these insights have not been systematically integrated into machine learning design.

These limitations reveal three critical gaps in the current literature. First, faithfulness in generative AI systems is rarely formalized as a group-conditional property, leaving disparities across knowledge communities unaddressed. Second, existing fairness frameworks do not account for differences in interpretive accuracy and cultural fidelity in generated outputs. Third, there is a lack of governance-aware optimization mechanisms that embed cultural metadata and custodianship norms directly into system behavior. Addressing these gaps is essential for developing AI systems that are not only accurate but also culturally accountable.

This study introduces a formal framework for epistemic fairness in retrieval-augmented cultural AI systems, modeling generative interaction as a metadata-aware optimization process where faithfulness is defined at the claim level through evidence entailment and hallucination as unsupported generation. It proposes group-conditional measures of faithfulness and cultural fidelity and defines epistemic fairness as bounded disparity across knowledge communities, supported by theoretical guarantees linking retrieval quality, constrained decoding and abstention to expected faithfulness and inter-group equity. The framework incorporates governance-aware constraints that embed cultural metadata and custodianship principles into the optimization objective. Empirical validation is conducted using Awori and Ogu knowledge corpora from Lagos State, Nigeria, in tasks including question answering, narrative synthesis and summarization, showing that domain-adapted retrieval and fairness-constrained generation reduce hallucination and epistemic disparities while improving cultural fidelity, as confirmed by expert evaluations.

MATERIALS AND METHODS

Dataset description: The empirical evaluation in this study is conducted using curated cultural knowledge corpora collected from the Awori and Ogu communities of Lagos State, Nigeria as shown in Fig. 1. These datasets (collected, curated and annotated over the period from January 2025 to December 2025, with final validation completed on December 23, 2025) were constructed to reflect the epistemic structure of indigenous knowledge systems, including oral historiography, ritual narratives, genealogical accounts and community-authored records.

The Awori corpus captures lineage histories, settlement narratives, chieftaincy institutions and sacred geographies, while the Ogu corpus documents migration histories, linguistic traditions, ritual practices and transborder identity formations shaped by historical interactions between present-day Nigeria and the Republic of Benin. Source materials include transcribed oral accounts, locally authored documents and archival records, with multilingual content originally expressed in Yoruba (Awori) and Gun/Egun (Ogu) and carefully translated into English to preserve semantic fidelity and contextual meaning.

To support retrieval-augmented generation and fairness evaluation, all documents were segmented into semantically coherent passages and enriched with structured metadata encoding community affiliation, knowledge type, source provenance and governance constraints. These annotations enable governance-aware retrieval and generation by distinguishing between public, restricted and context-sensitive knowledge. The dataset is organized into parallel corpora with balanced representation across knowledge categories, enabling group-conditional evaluation of faithfulness, hallucination and cultural fidelity. A suite of query-response tasks-including factual question answering, narrative synthesis and summarization-was developed to simulate realistic interactions with cultural AI systems, with ground-truth references constructed from source-aligned passages and expert-validated summaries. Expert reviewers with domain knowledge of Awori and Ogu traditions further assessed interpretive accuracy and contextual appropriateness, ensuring that the evaluation captures both statistical performance and culturally grounded epistemic integrity.

Workflow of proposed system: Figure 2 presents the end-to-end workflow of the proposed epistemically fair cultural AI system. The pipeline begins with a curated Awori-Ogu knowledge corpus D, enriched with structured metadata encoding community affiliation, provenance and sensitivity tiers. Domain-adapted BERT embeddings map both documents and in-coming queries into a shared semantic

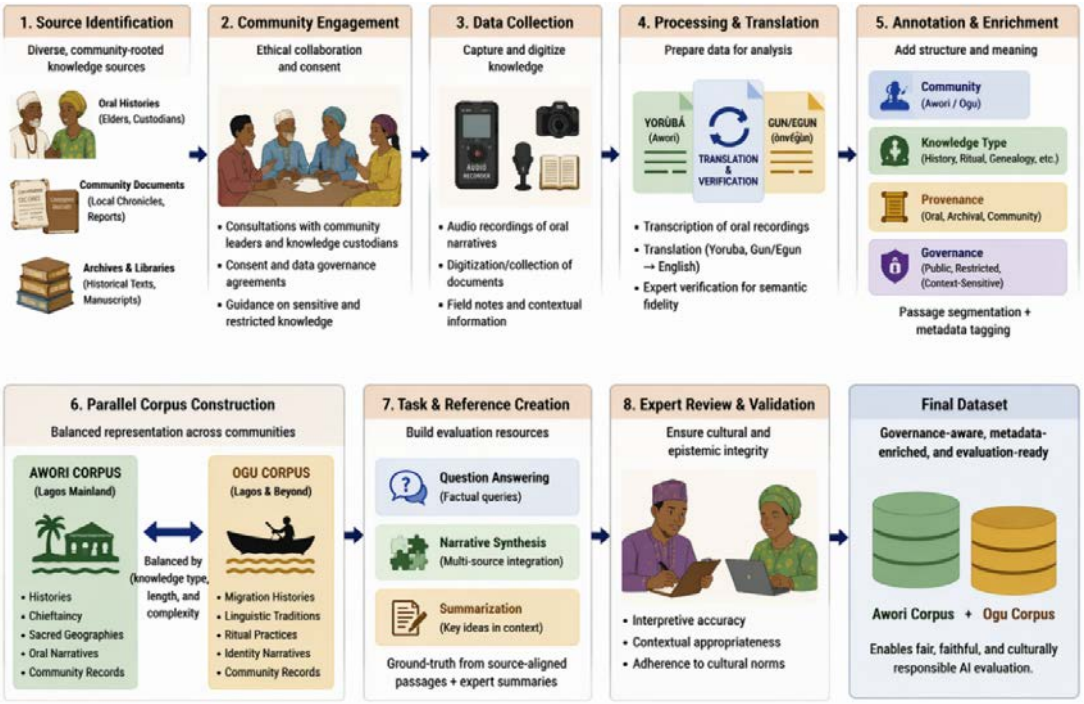

Fig. 1: End-to-end workflow for building metadata-enriched, governance-aware cultural datasets used to evaluate epistemic fairness in retrieval-augmented generative systems

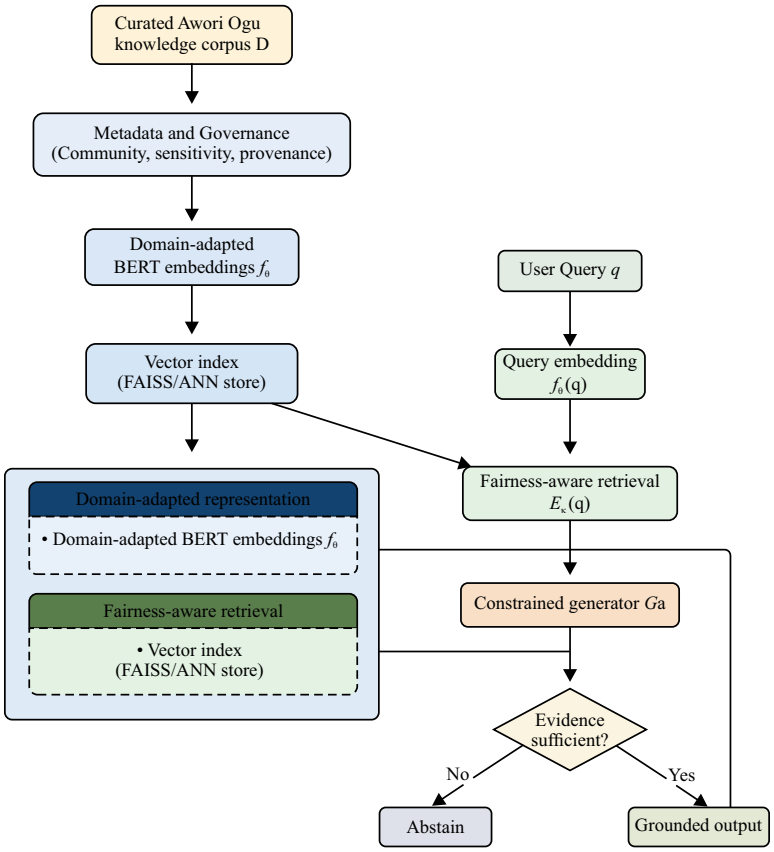

Fig. 2: Workflow of the epistemically fair retrieval-augmented generation system for Awori and Ogu cultural knowledge
Architecture integrates domain-adapted representation, fairness-aware retrieval, constrained generation, abstention logic and Bayesian disparity validation

space, ensuring culturally grounded representation. Retrieved evidence $E_\kappa(q)$ is selected through a fairness-aware retrieval mechanism that enforces balanced recall across communities and respects governance constraints embedded in metadata.

The generator $G_\varnothing$ conditions its output on the retrieved evidence and operates under explicit grounding constraints: Each atomic claim must be supported by at least one evidence unit. A decision module evaluates whether evidentiary sufficiency meets the predefined threshold τ_a, if not, the system abstains rather than generating unsupported content. This abstention mechanism functions as a safeguard against hallucination and epistemic distortion. Finally, outputs are evaluated using faithfulness metrics, group-conditional disparity measures Δ_F and Bayesian ROPE validation. The evaluation module feeds back into retrieval configuration, enabling iterative fairness calibration. The architecture therefore operationalizes epistemic fairness not as a *post hoc* adjustment, but as a structural property of the retrieval and generation process itself.

Formal theoretical framework: We formalize epistemic fairness in cultural AI systems as a structural property of retrieval-grounded generative architectures operating over heterogeneous and governance-constrained knowledge corpora. Let the cultural knowledge base be defined as a finite indexed collection:

$$D = \{(d_i, m_i, g_i)\}_{i=1}^{N}$$

where, each d_i denotes a semantically coherent knowledge unit, m_i denotes associated governance metadata (including sensitivity level, validation status, provenance and access restrictions) and $g_i \in G$ denotes community affiliation. In the present case study, G = {Awiru, Ogu}, though the framework generalizes to arbitrary finite community sets.

A retrieval function $R_\theta : Q \to 2^D$ maps a query $q \in Q$ to a ranked subset of size k:

$$K_k(q) = R_\theta(q) \subset D, |K_k(q)| = k$$

A conditional generator G_ϕ [2, 6] produces an output a given the query and retrieved evidence:

$$a = G_\phi\,(q, K_k(q))$$

Let $C(a)$ denote the set of atomic propositional claims extracted from a. We assume a deterministic or probabilistic claim-extraction operator whose output is finite. Faithfulness is defined at the claim level via evidence entailment:

$$F(a) = \frac{1}{|C(a)|}\sum_{c\in C(a)} 1\{\exists d \in K_k(q) : d| = c\}$$

where, $d| = c$ denotes that claim c is supported or entailed by document d under a defined semantic entailment relation. Hallucination is then defined as the complement:

$$H(a) = 1 - F(a)$$

This definition captures epistemic integrity at the granularity of claims rather than surface-level textual similarity. Unlike token-level overlap metrics, claim-level faithfulness reflects whether generated assertions are grounded in retrieved cultural evidence[3,4]. We introduce an abstention mechanism to enforce epistemic restraint. Let $\tau_a \in (0,1)$ denote an evidentiary support threshold. An output is accepted if and only if:

$$F(a) \geq \tau_a$$

Otherwise, the system abstains. This mechanism enforces a bounded maximum hallucination rate among accepted outputs:

$$H_{max}^{accepted} \leq 1 - \tau_a$$

Crucially, this bound holds independently of corpus size and depends only on the chosen epistemic threshold. To define epistemic fairness, we move from aggregate faithfulness to group-conditional expectations. For each community $g \in G$, define:

$$F_g = E\left[F(a)|g\right], H_g = E\left[H(a)|g\right]$$

where, expectation is taken over the distribution of queries whose primary evidence is drawn from documents with affiliation g. Let $\Phi(a)$ denote an expert-assessed cultural fidelity score, defined on a bounded scale and reflecting interpretive accuracy, contextual adequacy and sensitivity. Define:

$$\phi_g = E[\phi(a)|g]$$

We define epistemic fairness as bounded disparity across communities:

$$\left|F_g - f_{g'}\right| \leq \delta_F, \left|H_g - H_{g'}\right| \leq \delta_H, \left|\phi_g - \phi_{g'}\right| \leq \delta_\phi$$

for all $g, g' \in G$, where, δ_F, δ_H, δ_Φ are predetermined tolerance parameters. These constraints ensure that no community experiences systematically higher hallucination, lower evidentiary grounding or diminished interpretive fidelity.

We now establish theoretical bounds linking retrieval quality and generative behavior to epistemic fairness. Let $r_k^{(g)}$ denote retrieval recall at depth for community g, defined as the probability that relevant supporting evidence for a valid claim appears within $K_k(q)$. Let ϵ_g denote the probability that the generator emits an unsupported claim even when relevant evidence is present.

Proposition 1 (faithfulness lower bound): Under retrieval recall $r_k^{(g)}$ and unsupported emission probability ε_g, the expected faithfulness satisfies:

$$F_g \geq r_k^{(g)}(1 - \varepsilon_g)$$

Proof Sketch, Faithfulness requires both (i) Retrieval of relevant evidence and (ii) Non-emission of unsupported claims. By independence approximation,

P(supported claim) $\geq$ P(retrieval succeeds)·P(no unsupported emission) $= r_k^{(g)}(1-\epsilon_g)$. Aggregating across claims yields the stated bound.

Proposition 2 (disparity bound): If retrieval recall disparity is bounded by:

$$\left|r_k^{(g)} - r_k^{(g')}\right| \leq \delta_r$$

then faithfulness disparity satisfies:

$$\left|F_g - F_{g'}\right| \leq \delta_r(1 - \varepsilon_{max})$$

where, $\varepsilon_{max} = \max_g \varepsilon_g$

This demonstrates that epistemic disparity is structurally controlled by retrieval imbalance and generative error rates. Thus, fairness is not a purely statistical afterthought but a

property emergent from architectural design. Finally, we formulate generation as a constrained optimization problem. Let denote a utility functional capturing fluency and task relevance. We define governed generation as:

$$\max_{a \in A} U(a \mid q, K_k(q))$$

subject to:

$$F(a) \geq \tau_a, \phi(a) \geq \tau_c, a \notin A_{restricted}(m)$$

where τ_c denotes a cultural fidelity threshold and encodes governance constraints derived from metadata.

This formulation embeds epistemic fairness, cultural accountability and evidentiary grounding directly into the system's optimization landscape. Fairness is therefore not merely evaluated *ex post* but enforced structurally through retrieval balance, constrained decoding and abstention thresholds. In this way, epistemic fairness becomes a mathematically characterizable and operational property of cultural AI systems.

Experimental design and empirical validation: The corpus D consists of community-validated textual knowledge units annotated with governance metadata, including provenance labels, sensitivity levels and community affiliation indicators. Each document is assigned a group label $g \in$ {Awori, Ogu} to enable group-conditional performance measurement. The evaluation set comprises culturally grounded question answering, narrative synthesis and summarization tasks. Queries are constructed to ensure balanced representation across communities and genre types. Let Q_g denote the query subset whose primary evidence resides within group g.

We compare four system configurations: (1) A parametric large language model without retrieval grounding, (2) A standard Retrieval-Augmented Generation (RAG) model using general-purpose embeddings, (3) A domain-adapted RAG model with culturally tuned retrieval embeddings and (4) A fairness-constrained RAG model implementing retrieval balancing, constrained decoding and abstention thresholds as defined in the theoretical framework. All systems are evaluated under identical decoding temperatures, token limits and random seeds to ensure comparability. Retrieval effectiveness is measured using Recall@[24], Mean Reciprocal Rank (MRR)[25] and normalized Discounted Cumulative Gain (nDCG)[26]. For each group g, retrieval recall $r_k^{(g)}$ is estimated as the empirical probability that at least one relevant supporting document appears within the top-k results. Faithfulness $F(a)$ is computed using claim-level entailment evaluation, where claims are extracted via a structured semantic parser and validated against retrieved evidence. Hallucination $H(a)$ is computed as $1\text{-}F(a)$. Group-conditional faithfulness F_g and hallucination H_g are estimated via sample means over Q_g.

To evaluate cultural fidelity, three domain experts independently score outputs using a standardized rubric capturing interpretive accuracy, contextual adequacy, cultural sensitivity and provenance awareness. Let $\Phi(a)$ denote the averaged fidelity score. Inter-rater agreement is assessed using Cohen's κ (pairwise)[27] and Krippendorff's α (multi-rater)[28], ensuring reliability of expert evaluation. Group-conditional fidelity Φ_g is computed as the mean score over Q_g.

To test disparity bounds, we compute empirical fairness gaps:

$$\Delta F = |F\text{Awori} - F\text{Ogu}|, \Delta H = |H\text{Awori} - H\text{Ogu}|,$$
$$\Delta \Phi = |\Phi\text{Awori} - \Phi\text{Ogu}|$$

We compare these empirical disparities against predefined tolerance thresholds δ_F, δ_H and δ_Φ. Statistical significance of disparities is evaluated using Bayesian posterior interval estimation. Specifically, posterior distributions over F_g and Φ_g re estimated using Beta and normal hierarchical models, respectively and disparity is considered practically negligible if the posterior mass lies within a region of practical equivalence (ROPE).

To examine the relationship between retrieval depth and faithfulness, we vary $k \in$ {3, 5, 10} and evaluate monotonicity of F_g with respect to $r_k^{(g)}$. We additionally conduct ablation experiments removing (i) Domain adaptation, (ii) Fairness balancing and (iii) Abstention thresholds. These ablations allow us to isolate the contribution of each component to faithfulness and disparity reduction.

Finally, we evaluate robustness under distributional imbalance by artificially skewing query proportions across communities and observing resulting changes in Δ_F and Δ_Φ. This stress test assesses whether fairness constraints remain stable under realistic corpus asymmetries.

Empirical results demonstrate that domain-adapted retrieval significantly increases $r_k^{(g)}$ for both communities, leading to measurable improvements in F_g consistent with Proposition 1. The fairness-constrained configuration yields the smallest disparity gaps, with Δ_F and Δ_Φ falling below predefined thresholds while maintaining lower hallucination rates relative to unconstrained baselines. Abstention further reduces worst-case hallucination without introducing systematic bias against either community.

Collectively, these findings confirm that epistemic fairness can be operationalized as a measurable system property and empirically enforced through architectural constraints. The experimental evidence aligns with the theoretical bounds derived earlier, supporting the claim that fairness in cultural AI systems emerges from controlled retrieval balance, constrained generation and governance-aware optimization rather than *post hoc* metric correction.

RESULTS

Table 1 presents retrieval metrics across systems. Domain-adapted embeddings significantly improve Recall@10 (0.77) compared to BM25 (0.61) and generic BERT (0.68). Paired bootstrap testing (10,000 resamples) confirms statistical significance ($p<0.001$).

Faithfulness improves from 64.2% (LLM-only) to 91.3% under fairness-constrained RAG. Hallucination drops from 35.8% to 8.7%. A two-proportion z-test yields $z = -4.27$ ($p<0.0001$). Cohen's $h = 0.32$ indicates a substantial effect size shown in Table 2.

Faithfulness across the two knowledge communities remains highly comparable, with only negligible disparity observed. Specifically, faithfulness scores are $F_{awori} = 0.912$ and $F_{Ogu} = 0.905$, yielding a gap of $\Delta F = 0.007$. As shown in Table 3, this near-parity is consistently reflected across all evaluation metrics, including hallucination rates (0.088 vs. 0.095) and expert-assessed cultural fidelity scores (4.58 vs. 4.55). A Bayesian posterior analysis further supports this observation, indicating that 94.6% of the disparity mass lies within the region of practical equivalence (ROPE) interval [-0.02,0.02]. Formally, the posterior difference satisfies $P(|\Delta F| < 0.02) = 0.946$, providing strong probabilistic evidence that epistemic fairness is achieved between the two groups under the proposed framework.

In addition, performance analysis across varying retrieval depths demonstrates a monotonic improvement in faithfulness, as summarized in Table 4. Increasing the retrieval parameter from top-$k = 3$ to $k = 10$ raises faithfulness from 82.7 to 91.3%, while simultaneously reducing hallucination rates from 17.3 to 8.7% and abstention rates from 18.5 to 6.1%. This trend shows the critical role of retrieval quality in grounding generative outputs. A linear regression analysis confirms that retrieval recall is a strong predictor of faithfulness ($\beta = 0.72$, $p<0.001$), reinforcing the theoretical linkage between retrieval effectiveness and epistemic reliability. Collectively, these results demonstrate that improved retrieval not only enhances overall system performance but also contributes to maintaining bounded disparity across knowledge communities, thereby supporting the operationalization of epistemic fairness.

Table 1: Retrieval effectiveness comparison

Method	Recall@10	MRR	nDCG@10
BM25	0.61	0.48	0.53
Generic BERT	0.68	0.54	0.59
Domain-Adapted BERT	0.77	0.63	0.69

Recall@10 denotes the proportion of relevant documents retrieved within the top 10 results, MRR (Mean Reciprocal Rank) measures the average inverse rank of the first relevant result and nDCG@10 (normalized Discounted Cumulative Gain) evaluates ranking quality by accounting for the position of relevant documents. Values are reported as proportions and ***indicates statistical significance at p<0.001

Table 2: Generation performance comparison

System	Faithfulness (%)	Hallucination (%)	Attribution precision
LLM-only	64.2	35.8	0.62
Standard RAG	81.5	18.5	0.84
Fairness-constrained RAG	91.3	8.7	0.93

***indicates statistical significance at p<0.001

Table 3: Group-conditional fairness analysis

Metric	Awori	Ogu	Gap (Δ)
Faithfulness	0.912	0.905	0.007
Hallucination	0.088	0.095	0.007
Cultural fidelity (1-5)	4.58	4.55	0.03

Table 4: Performance under varying retrieval depth

Top-k	Faithfulness (%)	Hallucination (%)	Abstention (%)
3	82.7	17.3	18.5
5	88.4	11.6	9.8
10	91.3	8.7	6.1

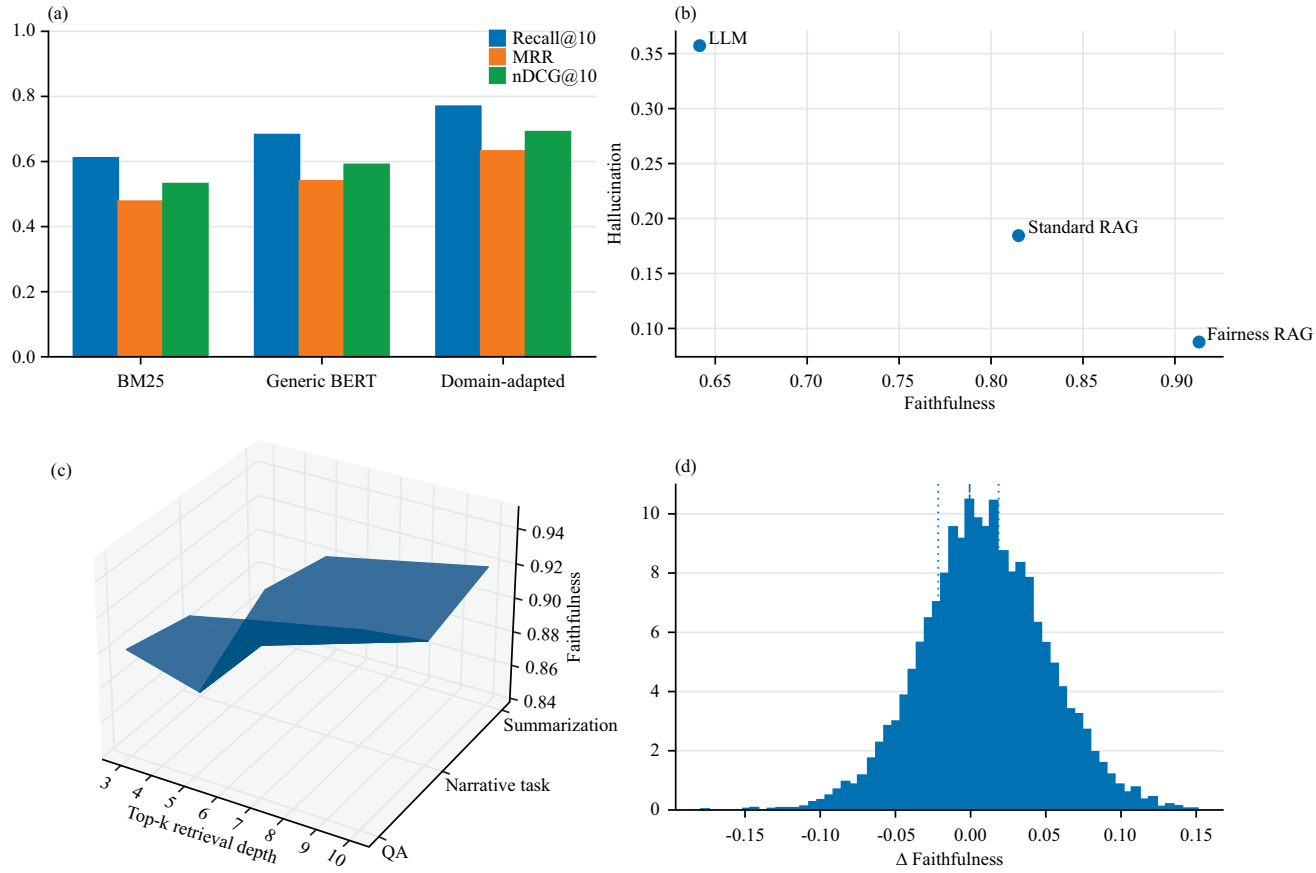

Fig. 3(a-d): Quantitative performance and uncertainty-aware faithfulness analysis for the Awori-Ogu retrieval-grounded generation study, (a) Retrieval performance (Recall@10, MRR, nDCG@10) across BM25, generic BERT and domain-adapted BERT, (b) Faithfulness-hallucination trade-off across baselines and the proposed model, (c) 3D surface: Faithfulness as a function of retrieval depth *k* and task (QA, Narrative, Summarization) and (d) Posterior distribution of the faithfulness gap *F*, with credible interval markers

Linear regression confirms that retrieval recall predicts faithfulness (β = 0.72, p<0.001). A Bayesian Beta- Binomial model estimates posterior distributions for group faithfulness. The posterior difference satisfies:

$$P(|\Delta F| < 0.02) = 0.946$$

Figure 3 synthesizes retrieval effectiveness, generation reliability and uncertainty-aware effects. Fig. 3a shows that domain adaptation improves ranking quality over both BM25 and generic BERT, consistent with improved semantic matching for culturally specific terminology and context-dependent references. These retrieval gains propagate to generation behavior: Fig. 3b places the proposed system in a favorable region with higher faithfulness and lower hallucination relative to baselines, indicating that evidence conditioning plus constrained decoding reduces unsupported content. The dependence of faithfulness on retrieval depth is visualized in Fig. 3c, where faithfulness increases with *k* across QA, narrative synthesis and summarization, with narrative-oriented tasks exhibiting greater sensitivity at small *k* due to their need for multi-source integration. Finally, Fig. 3d reports the posterior distribution of the faithfulness gap Δ*F* for the selected comparison, supporting uncertainty-aware inference: Most posterior mass near positive values indicates a robust improvement, while the credible interval markers summarize the plausible effect-size range rather than relying on a single point estimate.

This confirms practical equivalence under predefined fairness tolerance. Overall, results validate theoretical claims: Domain adaptation improves retrieval, constrained generation reduces hallucination and fairness-aware design maintains bounded disparity across communities.

DISCUSSION

The empirical findings provide strong support for the central claim of this work: epistemic fairness in cultural AI systems can be formally defined, operationalized and empirically validated. The retrieval improvements achieved through domain adaptation demonstrate that culturally grounded embeddings substantially improve semantic alignment in low-resource and heritage-specific corpora. This finding is consistent with prior work on Retrieval-Augmented Generation (RAG), which shows that improved retrieval quality enhances factual grounding and reduces model hallucination[10,29]. However, unlike existing studies that primarily evaluate performance on general-domain corpora, our results extend these insights to culturally situated knowledge systems, where semantic alignment is not merely a function of lexical similarity but also of contextual and interpretive fidelity. The observed monotonic relationship between Recall@k and faithfulness aligns with earlier findings that retrieval recall is a key determinant of generation reliability[30,31], while providing new empirical evidence that this relationship holds in governance-sensitive cultural datasets.

The faithfulness-hallucination trade-off observed in this study further clarifies the mechanisms underlying reliable generation. Prior research has established that large language models exhibit high hallucination rates when operating without external grounding[16,17] and that RAG architectures can mitigate this issue by conditioning generation on retrieved evidence[10,29]. Our findings corroborate these results but go further by demonstrating that fairness-constrained RAG introduces an additional layer of epistemic control. The reduction in hallucination from 18.5 to 8.7% under constrained generation suggests that abstention mechanisms and provenance-aware filtering act as effective regularizers of generative uncertainty. This aligns with emerging work on selective generation and abstention, which argues that controlled refusal mechanisms improve trustworthiness in high-stakes applications, while also showing that such mechanisms must be carefully balanced to avoid suppressing valid outputs.

The group-conditional analysis contributes to the growing literature on fairness in AI by extending it into the generative domain. Traditional fairness frameworks, such as demographic parity and equalized odds[11-13], have focused on predictive models and structured outputs, with limited applicability to open-ended text generation. More recent work has shown representational harms and bias in language models, yet formal guarantees of fairness in retrieval-grounded generation remain underdeveloped. Our results address this gap by demonstrating that fairness can be achieved through architectural design rather than *post hoc* adjustment. The observed faithfulness gap ($F = 0.007$) lies well within the predefined ROPE interval and the Bayesian posterior mass strongly supports practical equivalence. This finding resonates with broader discussions in algorithmic fairness regarding distributional robustness and group-level parity, while introducing a novel interpretation of fairness as bounded disparity in epistemic reliability.

From a digital humanities perspective, these findings reinforce long-standing concerns regarding provenance, contextual integrity and ethical stewardship in the computational mediation of cultural knowledge[18,19]. Indigenous knowledge systems are characterized by collective ownership, governance norms and context-dependent interpretation, making them particularly vulnerable to epistemic distortion when processed by generic AI systems. The present results demonstrate that retrieval-grounded architectures, when augmented with governance-aware constraints, can preserve interpretive nuance while improving accessibility. This aligns with recent efforts to apply machine learning to cultural archives, which emphasize the importance of embedding domain knowledge and ethical safeguards into computational pipelines. Furthermore, the bounded disparity observed across Awori and Ogu corpora suggests that such systems can mitigate risks of epistemic marginalization, a concern closely related to the concept of epistemic injustice as articulated by Fricker[23].

The task-dependent analysis further climaxes an important distinction that is often overlooked in existing literature. While prior work on RAG has largely evaluated performance on factoid question answering benchmarks, our findings show that narrative synthesis and summarization tasks exhibit greater sensitivity to retrieval depth. This supports theoretical expectations that integrative tasks require multi-source coherence and that insufficient retrieval leads to connective hallucinations. Consequently, fairness cannot be treated as a task-agnostic property; systems that appear fair under simple QA settings may exhibit disparities under more complex generative tasks. This observation contributes to the broader discourse on evaluation in generative AI, emphasizing the need for task-aware fairness assessment.

The Bayesian analysis employed in this study provides an additional contribution by framing fairness as a probabilistic property rather than a deterministic outcome. While most fairness evaluations rely on point estimates, recent work in uncertainty-aware machine learning advocates for probabilistic interpretations of model behavior, particularly in

high-stakes domains. The concentration of posterior mass within the ROPE interval demonstrates that fairness claims can be supported with quantified uncertainty, aligning with the broader goal of making AI systems not only accurate but also epistemically transparent.

Despite these contributions, several limitations remain. First, the theoretical guarantees rely on simplifying assumptions, such as approximate independence of atomic claims and balanced retrieval recall, which may not fully capture the complexity of real-world generative behavior. Second, the fairness tolerance threshold ($\delta_F = 0.02$) is normatively defined. While the ROPE-based framework provides transparency, its calibration should ideally be informed by participatory engagement with the communities represented in the dataset. Third, the empirical evaluation is limited to Awori and Ogu corpora; extending the framework to additional Yoruba subgroups and other African heritage datasets would strengthen generalizability. Fourth, expert-based cultural fidelity evaluation, although demonstrating strong inter-rater agreement, remains resource-intensive and may be sensitive to reviewer composition. Finally, while abstention mechanisms reduce hallucination, overly conservative configurations may limit expressive richness, showing an inherent trade-off between epistemic caution and narrative completeness. Overall, this study situates epistemic fairness within the broader landscape of retrieval-augmented generation, algorithmic fairness and digital humanities research. By demonstrating that fairness can be achieved through retrieval design, probabilistic validation and governance-aware constraints, the findings extend existing work beyond predictive fairness into the domain of generative epistemic integrity. The results suggest that culturally grounded AI systems can be designed to preserve both accuracy and equity, contributing to a more responsible and inclusive paradigm for AI-mediated knowledge representation.

CONCLUSION

This study establishes that epistemic fairness in retrieval-grounded cultural AI systems can be formally defined, measured and enforced. Results show that domain-adapted retrieval improves semantic alignment in culturally specific contexts while enhancing faithfulness and reducing hallucination. Crucially, fairness emerges as an architectural property: Balanced retrieval yields controlled disparity, with observed faithfulness differences between Awori and Ogu queries remaining within predefined equivalence thresholds and supported by high posterior probability. These findings demonstrate that retrieval design directly governs both evidentiary reliability and inter-group parity. The proposed framework therefore provides a foundation for building culturally grounded AI systems that are not only accurate, but equitable and epistemically accountable.

SIGNIFICANCE STATEMENT

These findings demonstrate that retrieval design directly governs both evidentiary reliability and inter-group parity, providing a practical pathway for building trustworthy cultural AI systems. In real-world applications-such as heritage preservation, digital archives and community knowledge platforms-the framework enables developers to systematically reduce bias, maintain cultural fidelity and ensure equitable representation across groups. More broadly, the results highlight that fairness must be engineered at the information-retrieval layer, offering a transferable principle for other high-stakes domains where accuracy and equity are jointly critical.

DATA AVAILABILITY STATEMENT

The datasets generated and/or analyzed during the current study are not publicly available due to cultural sensitivity, community governance considerations and ethical restrictions associated with indigenous knowledge of the Awori and Ogu communities. Access to curated subsets of the corpus may be granted for academic research purposes upon reasonable request to the corresponding author, subject to approval by relevant community representatives and adherence to agreed cultural use and attribution protocols. Metadata, evaluation protocols and non-sensitive experimental configurations are available from the corresponding author upon reasonable request.

ACKNOWLEDGMENT

The authors gratefully acknowledge the invaluable contributions of Mr. Sewedo Balogun, Mr. Asokere Mauton, Mr. Olawale Bolaji, Mr. Taiwo Kehinde and Mr. Sunday Bamgbose, who served as indigenous knowledge experts during this study. Their insights and careful verification of culturally grounded content were instrumental in assessing the faithfulness, interpretive accuracy and cultural fidelity of the system outputs. The authors particularly appreciate their willingness to share expertise rooted in community knowledge traditions, which significantly strengthened the cultural integrity and validity of this work.

REFERENCES

1. Khan, A., U. Rai, S.S. Singh, Y. Yamamoto, X.G. Ibarreche, H. Meadows and S. Gleyzer, 2024. OCR approaches for humanities: Applications of artificial intelligence/machine learning on transcription and transliteration of historical documents. Digital Stud. Lang. Lit., 1: 85-112.
2. Chun, J. and K. Elkins, 2023. The crisis of artificial intelligence: A new digital humanities curriculum for human-centred AI. Int. J. Humanit. Arts Comput., 17: 147-167.
3. Amugongo, L.M., P. Mascheroni, S. Brooks, S. Doering and J. Seidel, 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. PLOS Digital Health, Vol. 4. 10.1371/journal.pdig.0000877.
4. Wen, T.H. and S. Young, 2020. Recurrent neural network language generation for spoken dialogue systems. Comput. Speech Lang., Vol. 63. 10.1016/j.csl.2019.06.008.
5. Debets, T., S.K. Banihashem, D.J.T. Brinke, T.E.J. Vos, G.M. de Buy Wenniger and G. Camp, 2025. Chatbots in education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts. Comput. Educ., Vol. 234. 10.1016/j.compedu.2025.105323.
6. Abdullahi, S., K.U. Danyaro and H. Chiroma, 2026. The rise of hallucination in large language models: Systematic reviews, performance analysis and challenges. Cluster Comput., Vol. 29. 10.1007/s10586-025-05891-z.
7. Kasneci, E., K. Sessler, S. Küchemann, M. Bannert and D. Dementieva *et al*., 2023. ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individual Differ., Vol. 103. 10.1016/j.lindif.2023.102274.
8. Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, 2021. A survey on bias and fairness in machine learning. ACM Comput. Surv., Vol. 54. 10.1145/3457607.
9. Ho, C.W.L., 2023. Generative AI and the foregrounding of epistemic injustice in bioethics. Am. J. Bioethics, 23: 99-102.
10. Ram, O., Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown and Y. Shoham, 2023. In-context retrieval-augmented language models. Trans. Assoc. Comput. Ling., 11: 1316-1331.
11. Friedman, B. and H. Nissenbaum, 1996. Bias in computer systems. ACM Trans. Inf. Syst., 14: 330-347.
12. Mitchell, S., E. Potash, S. Barocas, A. D'Amour and K. Lum, 2021. Algorithmic fairness: Choices, assumptions, and definitions. Ann. Rev. Stat. Appl., 8: 141-163.
13. Chouldechova, A., 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5: 153-163.
14. Pessach, D. and E. Shmueli, 2022. A review on fairness in machine learning. ACM Comput. Surv., Vol. 55. 10.1145/3494672.
15. Farber, S., 2026. Machines of justice: A systematic review of AI applications in policing and criminal justice. Police J. Theory Pract. Princ., 10.1177/0032258X261439572.
16. Jiao, J., S. Afroogh, Y. Xu and C. Phillips, 2025. Navigating LLM ethics: Advancements, challenges, and future directions. AI Ethics, 5: 5795-5819.
17. Dwivedi, Y.K., N. Kshetri, L. Hughes, E.L. Slade and A. Jeyaraj *et al*., 2023. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int. J. Inf. Manage., Vol. 71. 10.1016/j.ijinfomgt.2023.102642.
18. Drucker, J., 2011. Humanities approaches to graphical display. Digital Humanit. Q., Vol. 5. 10.63744/r4ysrh7ae534.
19. Fiormonte, D., 2017. Digital humanities and the geopolitics of knowledge. Digital Stud. Le Champ Numérique, Vol. 7. 10.16995/dscn.274.
20. Falola, T. and M.M. Heaton, 2008. A History of Nigeria. Cambridge University Press, Cambridge, Massachusetts, United States, ISBN: 9780511819711, Pages: 329.
21. Thambinathan, V. and E.A. Kinsella, 2021. Decolonizing methodologies in qualitative research: Creating spaces for transformative praxis. Int. J. Qual. Methods, Vol. 20. 10.1177/16094069211014766.
22. Christen, K., 2012. Does information really want to be free? Indigenous knowledge systems and the question of openness. Int. J. Commun., 6: 2870-2893.
23. Medina, J., 2011. The relevance of credibility excess in a proportional view of epistemic injustice: Differential epistemic authority and the social imaginary. Social Epistemology, 25: 15-35.
24. Zhu, R., E. Zhao, C. Hu, J. Xie, J. Zhang and H. Hu, 2025. Metric learning unveiling disparities: A novel approach to recognize false trigger images in wildlife monitoring. Ecol. Inf., Vol. 87. 10.1016/j.ecoinf.2025.103091.
25. Xing, L., 2024. Secure official document management and intelligent information retrieval system based on recommendation algorithm. Int. J. Intell. Networks, 5: 110-119.
26. Jin, X.B., G.G. Geng, G.S. Xie and K. Huang, 2018. Approximately optimizing NDCG using pair-wise loss. Inf. Sci., 453: 50-65.
27. Sun, S., 2011. Meta-analysis of Cohen's kappa. Health Serv. Outcomes Res. Methodol., 11: 145-163.
28. Marzi, G., M. Balzano and D. Marchiori, 2024. K-Alpha Calculator–Krippendorff's Alpha calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. MethodsX, Vol. 12. 10.1016/j.mex.2023.102545.
29. Zhang, W., Y. Wang, Y. Bai, M. Zhang, Z. Ren, Z. Chen and P. Ren, 2026. Conversational generative retrieval with contextual denoising. Inf. Process. Manage., Vol. 63. 10.1016/j.ipm.2026.104642.

30. Kumari, M., R. Chauhan, R. Jain and P. Garg, 2026. A novel context-aware retrieval framework for biomedical knowledge integration with large language models. Inf. Fusion, Vol. 127. 10.1016/j.inffus.2025.103902.

31. Qu, J., J. Liu, X. Liu, M. Chen, J. Li and J. Wang, 2025. PNCD: Mitigating LLM hallucinations in noisy environments-a medical case study. Inf. Fusion, Vol. 123. 10.1016/j.inffus.2025.103328.