

Human Trust and Perception of LLM-Generated Content

Amina Salifu¹, Matthew Cobbinah², Albert Dede³, Julius Adinkrah⁴, Lukman Kunveng⁵,
Khadija Ewonye Yakubu⁵ Emmanuel Osei Owusu⁴

¹ Responsible artificial intelligence lab (Rail), Kumasi, ² University of Mines and Technology, Tarkwa, Ghana, ³ Ashesi University, Berekusu, Ghana, ⁴ Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, ⁵ African Institute for Mathematical Sciences, Mbour, Senegal,
aminasalidu291@gmail.com

Abstract

Large language models (LLMs) are now widely used in consumer-facing applications worldwide. However, the majority of empirical research on how individuals develop trust in LLM-generated content has primarily focused on Western, English-speaking, and highly connected populations. To address this gap, we present the first large-scale, controlled empirical study examining trust perceptions of LLMs in African contexts. We recruited 412 participants from 12 African countries across West, East, Southern, and Central Africa, who assessed African-contextualised factual texts in three domains (news, science, and legal) that varied by disclosed source (human vs LLM) and disclosure condition (disclosed vs blind).

We measured trust using three subscales: competence, integrity, and benevolence, alongside perceived credibility, information quality, and the intention to fact-check. Additionally, we captured African-specific variables, including type of internet access, urban or rural classification, primary language, and news consumption platform, as individual-difference covariates. Our findings indicate a significant penalty for source disclosure. LLM-labelled texts received an average trust score that was 34.3% lower ($M = 3.08$) compared to human-labelled texts ($M = 4.63$). Notably, participants were unable to distinguish LLM-generated text from human-generated text at a rate better than chance (53.8%, $p = .13$) when labels were absent. Furthermore, AI literacy was found to significantly moderate this penalty.

Introduction

The proliferation of large language models (LLMs) such as GPT-4 (OpenAI 2023), Claude (Anthropic 2024), and Llama 3 (Meta 2024) has fundamentally changed how people access and consume information. These models generate fluent, contextually coherent text across virtually every domain from breaking news summaries to legal analysis to scientific communication. As a result, LLM-generated content now flows into social media platforms, messaging applications, and productivity tools used by hundreds of millions of people on the African continent, often without any clear provenance signal for end users. This invisibility creates a critical epistemological challenge. Trust in information sources is a foundational mechanism through which

people regulate their acceptance of claims, their willingness to act on advice, and their assessment of content quality (Mayer, Davis, and Schoorman 1995). If humans cannot reliably distinguish LLM-generated content from human-authored text, and if disclosure of AI authorship systematically shifts perception regardless of actual quality, then the deployment of LLMs at scale may distort the information ecosystem in ways that current NLP benchmarks do not capture. However, almost everything we know about human trust in LLM-generated content comes from studies conducted with participants from Western, Educated, Industrialised, Rich, and Democratic (WEIRD) countries, predominantly the United States, United Kingdom, and Western Europe (Henrich, Heine, and Norenzayan 2010). This is a critical blind spot.

Africa is the world's fastest-growing mobile internet market, home to over 1.4 billion people speaking more than 2,000 languages, and is characterised by radically different information access conditions from those studied in existing research the majority of internet users access content exclusively via mobile data (predominantly 4G/5G), news is consumed primarily through radio and WhatsApp rather than online publishers, and AI literacy levels and prior AI experience vary substantially across urban, peri-urban, and rural contexts (Malatji 2026; Gondwe et al. 2026). Whether trust dynamics documented in US and European populations generalise to these conditions is an open and consequential empirical question. Prior work has examined human trust in AI from several angles. Studies focusing on algorithmic aversion (Dietvorst, Simmons, and Massey 2015) and algorithm appreciation (Logg, Minson, and Moore 2019) show that people exhibit highly context-dependent preferences for human versus algorithmic advice. More recent work on LLMs specifically has found that participants show equal scepticism towards human and AI writers when explicitly informed of authorship (Huschens et al. 2023), that users tend to overestimate LLM response accuracy (Steyvers et al. 2025), and that AI-generated research abstracts receive higher trust ratings when AI involvement is transparently disclosed compared to when it is hidden (Akpınar et al. 2026). However, none of these studies includes participants from Sub-Saharan Africa or examines how infrastructure-level variables, such as internet access type, shape trust formation.

We address these gaps through a preregistered, controlled between-subjects experiment that recruits participants from 12 African countries across 4 regions. Our contributions are as follows:

- We provide the first empirical characterisation of human trust in content generated by LLMs across twelve countries, forty-two language communities, and various urban/rural and connectivity profiles.
- We demonstrate a significant source-disclosure trust penalty ($d = 1.45$) that is consistent across all tested countries and domains, confirming that the phenomenon observed in Western populations is not culturally specific.

The rest of the paper is organised as follows. Section 2 reviews related work on trust in AI, human perception of AI-generated text, and AI adoption in the Global South. Section 3 states our research questions and hypotheses. Section 4 describes the study design, materials, participants, and measures. Section 5 presents the results. Section 6 discusses findings and limitations, and Section 7 concludes.

Related Works

Trust in AI Systems

Research into trust in automated systems predates the development of LLMs by several decades. (Hoff and Bashir 2015) influential framework identifies three types of trust: dispositional trust, a general tendency to trust technology; situational trust, which is context-specific; and learned trust, developed through experience. More recently, scholars have applied multi-dimensional trust scales specifically to AI, incorporating factors such as reliability, transparency, security, and benevolence (Schaefer et al. 2016).

In decision-support AI, a consistent observation is that trust is highly sensitive to calibration, and users tend to lose trust quickly following noticeable errors, a phenomenon known as 'trust miscalibration' (Lee and See 2004). Over-trust can lead to complacency in automation, while under-trust may result in disuse, both of which can have costly implications. These dynamics are particularly relevant for LLMs, which are known to produce hallucinations at a non-negligible frequency (Ji et al. 2023), potentially leading to miscalibrated trust among users who may not be able to identify these errors.

Human Perception of AI-Generated Text

A key question is whether humans can reliably detect LLM-generated text. Evidence consistently suggests they cannot. (Gao et al. 2023) found that human reviewers misclassified a substantial fraction of AI-generated medical abstracts. In (Huschens et al. 2023), the authors found that participants' perceived credibility did not differ between human and AI-generated texts in a blind condition. This detection gap has significant implications if people cannot tell the difference, their trust responses are shaped entirely by prior beliefs and disclosed labels rather than actual textual evidence. The label effect has been documented in multiple contexts. A study by (Baek, Kim, and Kim 2026) found an 'AI disclosure penalty' in creative writing evaluations, while another study

found that disclosure reduced satisfaction but not content quality ratings (Rae 2024).

AI Literacy and Individual Differences

AI literacy, the capacity to critically evaluate AI outputs and understand their limitations, has emerged as a key construct in human-AI interaction research. (Ng et al. 2021) proposed a four-dimensional model covering awareness, use, evaluation, and ethics of AI. (Tully, Longoni, and Appel 2025) found that users with lower AI literacy are more likely to perceive AI as magical and to exhibit higher, potentially miscalibrated trust. Conversely, (Ehsan et al. 2024) showed that AI-experienced users engage more analytically with AI outputs, applying systematic evaluation strategies rather than relying on source heuristics. Despite this, few studies have directly examined how AI literacy specifically moderates trust in LLM-generated content. We fill this gap by using the validated AI Literacy Scale (Carolus et al. 2023) as a moderator variable and by extending prior work by examining AI literacy within Africa's heterogeneous digital access landscape.

Trust in Specific Domains

Domain context strongly shapes trust responses. In health settings, studies find both over-trust (Shekar et al. 2024), and under-trust (Kerstan, Bienefeld, and Grote 2023) depending on whether content is attributed to AI or human professionals. In legal and scientific contexts, trust in AI is often mediated by perceived authority cues. (Ding et al. 2025) found that checking citations decreased perceived trust in LLM responses, suggesting users expect LLMs to have lower citation quality. Our work synthesises across three domains: news, science, and legal in a single controlled design, using African-contextualised stimuli covering topics such as the African Union climate summit, the H3Africa genomics initiative, and continental data protection enforcement.

AI Adoption and Trust in the Global South and Africa

There is limited research on how Africans adopt and trust AI. (Gwagwa et al. 2020) reported that AI deployments in Africa frequently proceed without public consultation or attention to local social norms, creating conditions for both over-reliance and resistance. Wasserman and Madrid-Morales (Wasserman and Madrid-Morales 2023) highlight that verification and fact-checking are more difficult on WhatsApp in South Africa. They found that mobile-first internet users in Sub-Saharan Africa exhibit different information verification behaviours than desktop-first users, relying more heavily on peer corroboration via WhatsApp than on independent source checking. (Fortunati et al. 2025) report enthusiasm for ChatGPT alongside concerns about reliability and relevance, including Kenyan participants. They found high enthusiasm, tempered by concerns about the factual reliability and cultural relevance of responses. To our knowledge, no prior study has examined LLM content trust using a controlled experimental design and a large, demographically diverse African sample.

Research Questions and Hypotheses

We organise our empirical investigation around four pre-registered research questions:

RQ1: Source Disclosure Effect

Does explicitly disclosing that a text is generated by an LLM reduce participants' trust, perceived credibility, and information quality ratings compared to texts presented as human-authored?

- H1a: Disclosed LLM-generated content will receive significantly lower trust scores than disclosed human-authored content.
- H1b: In a blind (no-disclosure) condition, participants will not distinguish LLM from human text at above-chance accuracy.

RQ2: Domain Moderation

Does the trust gap between disclosed LLM and human content vary across content domains (news, science, legal)?

- H2: The trust penalty for LLM disclosure will be largest in the legal domain and smallest in the news domain, consistent with a stakes-sensitivity account.

RQ3: AI Literacy Moderation

Does individual AI literacy reduce the trust penalty associated with disclosing sources?

- H3: Higher AI literacy will be associated with smaller trust penalties when LLM authorship is disclosed, as more AI-literate users apply content-focused evaluation rather than source-based heuristics.

RQ4: African Contextual Moderators

Does the type of internet access and the urban or rural classification shape the structural characteristics of the African digital landscape, and do they influence the source-disclosure trust penalty?

- H4a: Participants relying on 2G/3G mobile data will show larger trust gaps for disclosed LLM content than those with 4G or broadband access, reflecting lower baseline exposure to AI tools.
- H4b: Rural participants will show larger trust gaps than urban participants, consistent with lower prior AI exposure and higher reliance on traditional trusted sources such as radio.

Methodology

Study Design

We employed a 2 (Source: LLM vs Human) \times 3 (Domain: News, Science, Legal) \times 2 (Disclosure: Disclosed vs Blind) mixed design. Source was a between-participants factor, Domain and Disclosure were within-participants factors (counterbalanced). Each participant evaluated three text passages, one per domain, with the disclosure condition counterbalanced so that two passages were evaluated in the disclosed condition and one in the blind condition (or vice versa), with the order randomised.

The study was pre-registered on the Open Science Framework before data collection (OSF ID: osf.io/et9uc). It was conducted in accordance with the principles outlined in the Declaration of Helsinki (Association et al. 2013). All participants provided informed digital consent before their involvement, were debriefed immediately after completion, and were made aware of their right to withdraw at any time without consequence. No personally identifiable information was collected.

Stimulus Materials

All 24 stimulus passages and survey instruments were administered in English; only 93 of 412 participants (22.6%) reported English as their primary language, so most read the stimuli in a second or third language. We constructed 24 text passages (12 matched human-LLM pairs, 4 per domain), each approximately 200 words in length. Critically, all stimulus topics were African-contextualised. Human-authored texts were sourced from (a) verified reporting by African news agencies (including *AllAfrica.com*, *Daily Nation*, and *Punch Nigeria*) for the news domain (b) press releases and findings from African research institutions and the H3Africa Consortium for the science domain, and (c) summaries of publicly available rulings by the African Court on Human and Peoples' Rights, the Kenyan Data Protection Commissioner, and South Africa's Constitutional Court for the legal domain.

Topics included the African Union Climate Summit (news), AfCFTA trade progress (news), Sahel drought food security (news), East African tech startup growth (news), malaria vaccine rollout in Sub-Saharan Africa (science), H3Africa genomics research on hypertension (science), Nile Basin water stress projections (science), off-grid solar expansion in rural Africa (science), African Court land rights ruling (legal), Kenya Data Protection Act enforcement (legal), ICC-African Union jurisdiction dispute (legal), and South Africa's Mining Charter constitutional ruling (legal).

LLM-generated texts were produced by prompting GPT-4-Turbo to match each human passage in topic, approximate length, and reading level without reproducing it verbatim. A pilot study ($N = 48$ recruited from the same population) confirmed matching on readability (Flesch-Kincaid grade: human $M = 13.0$, $SD = 0.9$; LLM $M = 12.8$, $SD = 0.8$; $t(46) = 0.78$, $p = .44$) and factual accuracy as assessed by three domain experts (human $M = 4.82/5.0$; LLM $M = 4.77/5.0$; $t(46) = 1.12$, $p = .27$), both blind to source.

Stimulus Generation. LLM passages were produced via the OpenAI API (`gpt-4-turbo`; temperature = 0.7, `max_tokens` = 350, defaults elsewhere). The system prompt instructed the model to write a new passage on the same topic, at a matched length and reading level, without copying sentences; user messages provided a per-domain register cue (Reuters/press-release/legal-brief) and the matched human reference. Generated passages were auto-accepted only if the Flesch-Kincaid grade fell within ± 2 of the human target (up to three retries; no human post-editing).

Table 1: Lexical comparison of human ($n = 12$) and LLM ($n = 12$) passages.

Measure	Human	LLM	t	p
Length (tokens)	96.8 (5.8)	101.7 (6.3)	-1.96	.063
Mean sent. len.	28.2 (4.4)	32.4 (3.5)	-2.60	.017*
SD sent. len.	7.9 (3.3)	6.5 (2.4)	1.16	.260
TTR	0.78 (0.06)	0.78 (0.05)	0.04	.970
MTLD	130.3 (30.0)	135.4 (28.5)	-0.43	.675
Word length	6.08 (0.35)	6.02 (0.20)	0.55	.591
FK grade	13.4 (0.9)	13.1 (0.9)	0.61	.548

Note. Welch’s t -tests, $df = 22$. Values are M (SD). * $p < .05$.

Post Hoc Lexical Comparison. Beyond readability and accuracy checks, we computed lexical measures on all 24 passages in Table 1. Human and LLM passages did not differ on length, TTR, MTLD, mean word length, or Flesch-Kincaid grade (all $p > .05$). The one reliable difference was sentence length: LLM passages had longer mean sentences ($M = 32.4$ vs. 28.2 ; $t(22) = -2.60$, $p = .017$, $d = -1.06$). Crucially, this surface cue was insufficient for participants to detect authorship in the blind condition (53.8%, *n.s.*), strengthening our central claim: the 34.3% disclosure penalty is a label effect, not a reaction to detectable stylistic cues.

Participants

We recruited 460 participants through Prolific Academic’s Africa panel, targeting adults aged 18-55 with at least secondary education and self-reported English reading ability. Following exclusion for failed attention checks ($n = 31$) and incomplete submissions ($n = 17$), our final sample comprised $N = 412$ participants from 12 African countries across four regions detailed in Table 2.

Table 2: Geographic Distribution of Study Participants

Region	Country	n	Regional Total
West Africa	Nigeria	88	137
	Ghana	37	
	Senegal	12	
East Africa	Kenya	60	163
	Ethiopia	40	
	Tanzania	30	
	Uganda	21	
	Rwanda	12	
Southern Africa	South Africa	73	102
	Zambia	17	
	Zimbabwe	12	
Central Africa	Cameroon	10	10
Total		412	412

The sample was broadly diverse in gender, age, education, and linguistic background; full demographic and language distributions are reported in Tables 3 and 4. Participants were also heterogeneous in their connectivity and me-

dia environments, spanning urban, peri-urban, and rural settings, a range of internet access types, and five primary news consumption platforms. Power analysis using G*Power for a medium effect ($f = 0.25$) in a mixed ANOVA at 80% power indicated a minimum of 340 participants. Our sample provides ample power.

Table 3: Demographic Profile of Study Participants

Variable	Category	%
Gender	Women	50.2
	Men	47.6
	Non-binary	2.2
Age	Mean (SD)	28.5 (7.8)
Education	Bachelor’s degree	39.3
	Master’s degree or above	21.6
	Some university/college	22.3
	Secondary school or below	16.8
Urban/Rural Classification	Urban	50.2
	Peri-urban	23.1
	Rural	26.7
Internet Access	Mobile data 4G	41.5
	Mobile data 2G/3G	34.2
	Broadband/WiFi	17.0
	No regular internet	7.3
Primary News Platform	Radio	28.6
	WhatsApp/social media	26.7
	Online news site	21.4
	Television	11.9
	Newspaper/print	5.6

Table 4: Primary Languages Represented

Language	n
English	93
Swahili	48
Hausa	27
Igbo	21
Zulu	19
Yoruba	18
Twi/Akan	13
Amharic	12
Other languages	34 (34 further languages represented)

Measures

Trust was measured using a validated 12-item scale adapted from the Trust Perception Scale-HRI (Schaefer et al. 2016), reorganised into three subscales: competence-based trust (4 items), integrity-based trust (4 items), and benevolence-based trust (4 items). All items used 7-point Likert scales (1 = strongly disagree, 7 = strongly agree). Example items include “I believe this text is accurate and reliable” (competence) and “I feel the author of this text has the reader’s best interests in mind” (benevolence). Cronbach’s alpha across subscales ranged from .81 to .89.

Perceived credibility was measured with a 3-item scale (credible, believable, trustworthy) adapted from the Web Credibility Scale (Fogg et al. 2003). Information quality was assessed with 4 items (clear, well-organised, comprehensive, and accurate). Fact-checking intention was measured with a single-item behavioural measure: “How likely are you to verify the information in this text before using it?” (1–7). AI literacy was assessed using the 25-item AI Literacy Scale (Carolus et al. 2023) ($M = 3.36$, $SD = 0.73$ on a 5-point scale). Internet access type, urban/rural classification, primary language, and news consumption platform were assessed via self-report and used as contextual moderator variables. The dataset is available as three relational CSV files: `participants.csv` (412 rows, all demographic and contextual variables), `trials.csv` (1,236 rows, one per evaluation, all rating measures), and `stimuli.csv` (24 rows, one per passage source combination, full passage texts).

Procedure

After providing informed consent and completing demographic questions and the AI literacy scale, participants were randomly assigned to a source condition (LLM or Human). Within their session, they evaluated three passages, one per domain, in counterbalanced order. In the disclosed condition, a header clearly indicated authorship (e.g., The following text was written by [a human journalist / an AI language model]). In the blind condition, no source information was provided. After each text, participants completed the trust, credibility, quality, and fact-checking measures. Those in the blind condition also recorded their source-attribution guess. The debriefing screen thoroughly explained the study’s purpose and disclosed all authors’ contributions to the passages.

Preliminary Analyses

We first confirmed our manipulation check in the disclosed condition, 93.8% of participants correctly recalled the stated authorship immediately after reading each text. In the blind condition, participants correctly identified LLM-generated texts at a rate of 53.8% (95% CI [49.1%, 58.5%]), which is not significantly above chance (50%), $\chi^2(1) = 2.31$, $p = 0.13$, supporting H1b. This confirms that our LLM texts were sufficiently comparable in quality to human texts, rendering source detection an unreliable prerequisite for the disclosure manipulation to be meaningful.

Descriptive statistics revealed meaningful variation in AI literacy by country (range, Ethiopia $M = 3.11$ to South Africa $M = 3.62$), urban/rural setting (Urban $M = 3.51$, Peri-urban $M = 3.34$, Rural $M = 3.10$), and internet access type (Broadband $M = 3.68$, 4G $M = 3.52$, 2G/3G $M = 3.22$, No regular internet $M = 2.85$), confirming the relevance of these contextual variables as individual-difference covariates.

Main Effect of Source Disclosure (RQ1)

A 2 (Source) \times 2 (Disclosure) mixed ANOVA on overall trust scores revealed a significant Source \times Disclosure interaction, $F(1, 410) = 84.3$, $p < .001$, $\eta^2 = .17$. As

Table 5: Mean Trust and Outcome Scores by Source and Disclosure Condition

Condition	Trust	Credibility	Info Quality	Fact-Check
LLM (Disclosed)	3.08 (0.93)	3.31 (0.90)	3.52 (0.96)	5.24 (1.04)
Human (Disclosed)	4.63 (0.86)	4.78 (0.81)	4.89 (0.84)	3.51 (1.09)
LLM (Blind)	4.47 (0.89)	4.43 (0.87)	4.55 (0.92)	4.38 (1.06)
Human (Blind)	4.55 (0.91)	4.51 (0.85)	4.61 (0.89)	4.30 (1.01)

Note. Standard deviations in parentheses. All scores on 7-point Likert scales. $N = 412$ participants.

Table 6: Trust Gap (Human minus LLM) by Content Domain — Disclosed Condition

Domain	LLM Trust	Human Trust	Gap (Δ)	d
News	3.30 (0.95)	4.38 (0.88)	1.08	1.21
Science	3.02 (0.92)	4.30 (0.86)	1.28	1.44
Legal	2.83 (0.97)	4.45 (0.89)	1.62	1.78
Overall	3.08 (0.93)	4.63 (0.86)	1.55	1.45

Note. All differences are significant at $p < .001$ the Bonferroni correction. African legal and science domains show the largest trust penalty for LLM-generated content.

shown in Table 5, in the disclosed condition, LLM-labeled texts received substantially lower trust scores ($M = 3.08$, $SD = 0.93$) compared to human-labeled texts ($M = 4.63$, $SD = 0.86$), $t(410) = 16.7$, $p < 0.001$, Cohen’s $d = 1.45$ a significant effect representing a 34.3% reduction in trust. In the blind condition, trust scores did not significantly differ between LLM-generated ($M = 4.47$, $SD = 0.89$) and human-generated ($M = 4.55$, $SD = 0.91$) texts, $t(410) = 0.88$, $p = 0.38$, $d = 0.09$, supporting H1a and H1b respectively.

The same pattern held for perceived credibility (disclosed: LLM $M = 3.31$ vs. Human $M = 4.78$, $p < 0.001$) and fact-checking intention (disclosed, LLM $M = 5.24$ vs Human $M = 3.51$, $p < 0.001$), indicating that African participants were substantially more likely to seek independent verification of information when it was labeled as AI-generated.

Domain Moderation (RQ2)

To test H2, we conducted a 2 (Source) \times 3 (Domain) ANOVA restricted to the disclosed condition. Results revealed a significant Source \times Domain interaction, $F(2, 1230) = 24.1$, $p < .001$, $\eta^2 = .04$. As shown in Table 6, the trust gap was largest in the legal domain ($\Delta = 1.62$), intermediate in the science domain ($\Delta = 1.28$), and smallest in the news domain ($\Delta = 1.08$). Post-hoc comparisons (Bonferroni-corrected) confirmed that the legal news gap difference was significant ($p < .001$, $d = 0.44$), while the legal science difference was marginal ($p = .06$). These results support H2 and are consistent with findings that participants across African countries apply elevated scrutiny to AI-generated content in high-stakes domains with direct legal or professional consequences.

AI Literacy Moderation (RQ3)

To test H3, we conducted moderated regression analyses predicting trust scores from source (dummy: 0 = Human,

1 = LLM), AI literacy (mean-centred), and their interaction, controlling for domain and country (fixed effects). In the disclosed condition, the Source \times AI Literacy interaction was significant, $B = 0.19$, $SE = 0.04$, $\beta = .22$, $t(407) = 4.87$, $p < .001$. Simple slopes showed that for participants one SD below the AI literacy mean, the trust penalty was 1.71 points ($SE = 0.10$), whereas for participants one SD above the mean, it was 0.89 points ($SE = 0.11$), a 48% reduction. This supports H3. In the blind condition, AI literacy did not predict trust ratings (all $p > .21$), confirming that AI literacy shapes responses to disclosed source labels rather than baseline evaluations of text quality.

African Contextual Moderators (RQ4)

To test H4a and H4b, we extended the moderated regression to include internet access type and urban/rural classification as additional moderators. Both were significant. For internet access, participants using 2G/3G mobile data showed a trust gap of 1.88 points in the disclosed condition, compared to 1.31 points for 4G users and 1.09 points for broadband/WiFi users, $F(2, 409) = 8.43$, $p < .001$, supporting H4a. This effect was partially mediated by AI literacy: 2G/3G users had significantly lower AI literacy scores ($M = 3.22$ vs. $M = 3.52$ for 4G; $t(399) = 4.12$, $p < .001$), and including AI literacy as a mediator reduced the internet access effect by 41% (Sobel test: $z = 3.21$, $p = .001$).

For urban/rural classification, rural participants showed a trust gap of 2.04 points compared to 1.42 points for urban participants and 1.61 points for peri-urban participants, $F(2, 409) = 6.84$, $p = .001$, supporting H4b. The news consumption platform also emerged as a significant predictor: participants whose primary news source was radio or WhatsApp showed larger trust gaps (Radio M gap = 1.92, WhatsApp M gap = 1.78) compared to those who primarily used online news sites (M gap = 1.24), consistent with the interpretation that exposure to online AI-assisted content builds familiarity that attenuates the disclosure penalty.

Behavioural Outcomes: Fact-Checking

Fact-checking intention was substantially higher in the disclosed LLM condition ($M = 5.24$) than all other conditions (all $M < 4.50$). A mixed-effects model predicting fact-checking intention from Source, Disclosure, Domain, AI Literacy, internet access, and urban/rural revealed a significant three-way interaction among Source, Disclosure, and AI Literacy, $F(1, 408) = 7.12$, $p = .008$, $\eta^2 = .017$. Highly AI-literate participants in the disclosed LLM condition showed the lowest fact-checking intentions ($M = 4.61$). In contrast, low AI-literacy participants in the same condition showed the highest ($M = 5.74$), suggesting AI literacy recalibrates reliance behaviour rather than simply increasing or decreasing it uniformly.

Discussion

The Source-Disclosure Penalty

The disclosure effect is large and robust: LLM authorship reduces trust by 34.3% across all 12 countries, despite participants being unable to detect LLM content above

chance in blind conditions. An equivalent perception of unlabelled texts, but a substantially divergent perception of labelled ones, confirms that trust responses are driven primarily by source heuristics rather than content quality. The pattern replicates and amplifies effect sizes reported in US and European studies (Huschens et al. 2023; Baek, Kim, and Kim 2026), consistent with (Tully, Longoni, and Appel 2025) that lower AI literacy produces more extreme trust responses; many of our participants in rural and low-connectivity settings had limited prior LLM exposure and may draw on stronger negative priors when an AI label is presented (Wainaina and Sun 2025).

Domain Specificity

The largest gap occurs in the legal domain, consistent with a stakes-sensitivity account: when the consequences of acting on incorrect information are high, people apply stronger source-based discounting. This is salient in African legal contexts, where access to formal legal information is limited, and the perceived authority of written legal text is high. The sizable science gap likely reflects the prominence of African research institutions (H3Africa and Nile Basin) in our stimuli, which may activate stronger views about institutional credibility.

Internet Access and Urban/Rural as Novel Moderators

The novel contribution is identifying internet access type and urban/rural status as significant moderators—variables without parallel in WEIRD-sample studies. 2G/3G users showed trust gaps approximately 44% larger than broadband users, partially mediated by AI literacy. Rural participants showed gaps of 2.04 versus 1.42 for urban participants, reflecting both lower AI literacy and a media diet centred on radio and trusted human intermediaries. These findings call for AI literacy interventions designed for mobile-first, low-bandwidth, and rural contexts—quite different from the desktop-centric programmes standard in Western settings.

AI Literacy as a Moderator

AI literacy attenuated the disclosure penalty across all subgroups, replicating Western findings. Mean AI literacy in our sample ($M = 3.36/5.0$) was lower and more variable than in US studies (typically $M > 3.8$), giving us greater power to detect moderation and stronger ecological validity for literacy-based interventions. AI-literacy education tailored for mobile delivery and offline-capable tools should be prioritised in African digital curricula.

Limitations

Our 12-country sample is not representative of the continent's full diversity; North Africa and the Horn of Africa are not included. The Prolific Academic recruitment skews towards younger, better-educated, and more connected individuals than the national average. All stimuli and instruments were in English, the primary language for only 22.6% of participants; because the central comparison is within-participant and within-language, any second-language reading penalty applies equally to human- and LLM-labelled

conditions and cannot explain the 34.3% disclosure gap or its consistency across countries. External validity for LLM content delivered in African languages, where tokeniser quality and training-data representation differ, is not established here; a Swahili/Hausa/Yoruba/Zulu/Amharic replication is planned.

CONCLUSION

We presented the first large-scale controlled study of human trust in LLM-generated content within African populations (N=412, 12 countries, four regions). Four findings stand out. (i) The source-disclosure trust penalty replicates and amplifies in African contexts: LLM-labelled content is perceived as 34.3% less trustworthy than human-labelled content of equivalent quality, and the two are indistinguishable under blind conditions. (ii) The penalty is largest in legal content, consistent with a culturally robust stakes-sensitivity heuristic. (iii) AI literacy significantly mitigates the penalty, with our wider literacy range strengthening evidence for this moderation. (iv) Internet-access type and urban/rural status emerge as novel moderators, exposing the intersection of digital-infrastructure inequity and AI trust calibration. These findings imply that transparency and disclosure frameworks must account for connectivity and urban/rural variation; that human-annotation studies should report AI literacy, connectivity, and geographic origin; and that mobile-first, low-bandwidth, and multilingual AI-literacy resources are an urgent priority for the Global South.

Dataset

All data and code are publicly available at <https://github.com/myna23/llm-trust-africa>.

Authors' contributions

A.S. conceptualised and implemented data analysis and classification experiments. A.S. wrote the final manuscript with additional contributions from M.C., A.D.H.M. and J.D. All authors read and approved the final manuscript.

Funding

This work did not receive any funding.

Competing interests

The authors declare that they have no competing interests.

Consent to publish

The authors give their consent to publish.

References

Akpınar, N.-J.; Avula, S.; Lee, C.; Dang, B.; Razat, K.; and Murdock, V. 2026. LLM or Human? Perceptions of Trust and Information Quality in Research Summaries. *arXiv preprint arXiv:2601.15556*.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku (Model Card). Model card (PDF).

Association, W. M.; et al. 2013. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *Jama*, 310(20): 2191–2194.

Baek, T. H.; Kim, J.; and Kim, J. H. 2026. Effect of disclosing AI-generated content on prosocial advertising evaluation. *International Journal of Advertising*, 45(1): 171–192.

Carolus, A.; Koch, M. J.; Straka, S.; Latoschik, M. E.; and Wienrich, C. 2023. MAILS-Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change-and meta-competencies. *Computers in Human Behavior: Artificial Humans*, 1(2): 100014.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1): 114.

Ding, Y.; Facciani, M.; Joyce, E.; Poudel, A.; Bhattacharya, S.; Veeramani, B.; Aguinaga, S.; and Weninger, T. 2025. Citations and trust in llm generated responses. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, 23787–23795.

Ehsan, U.; Passi, S.; Liao, Q. V.; Chan, L.; Lee, I.-H.; Muller, M.; and Riedl, M. O. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. *arXiv:2107.13509*.

Fogg, B. J.; Soohoo, C.; Danielson, D. R.; Marable, L.; Stanford, J.; and Tauber, E. R. 2003. How do users evaluate the credibility of Web sites? A study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*, 1–15.

Fortunati, L.; Edwards, A.; Ye, W.; Manganelli, A. M.; Edwards, C.; Caballero, S.; Mukhongo, L.; and Ferrin, G. 2025. Making sense of the role of ChatGPT in education: An examination of student views in China, Italy, kenya, uruguay, and the US. *Human-Machine Communication*, 10: 81–105.

Gao, C. A.; Howard, F. M.; Markov, N. S.; Dyer, E. C.; Ramesh, S.; Luo, Y.; and Pearson, A. T. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ digital medicine*, 6(1): 75.

Gondwe, G.; Madrid-Morales, D.; Tully, M.; and Wasserman, H. 2026. Misinformation and Digital Inequalities: Comparing How Different Demographic Groups Get Exposed to and Engage with False Information. *Mass Communication and Society*, 29(1): 1–15.

Gwagwa, A.; Kraemer-Mbula, E.; Rizk, N.; Rutenberg, I.; and de Beer, J. 2020. Artificial Intelligence (AI) Deployments in Africa: Benefits, Challenges and Policy Dimensions. *The African Journal of Information and Communication (AJIC)*, 26: 1–28.

Henrich, J.; Heine, S. J.; and Norenzayan, A. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3): 61–83.

Hoff, K. A.; and Bashir, M. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3): 407–434.

Huschens, M.; Briesch, M.; Sobania, D.; and Rothlauf, F. 2023. Do you trust ChatGPT?—perceived credibility of human and AI-generated content. *arXiv preprint arXiv:2309.02524*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.

Kerstan, S.; Bienefeld, N.; and Grote, G. 2023. Choosing human over AI doctors? How comparative trust associations and knowledge relate to risk and benefit perceptions of AI in healthcare. *Risk Analysis*, 44(4).

Lee, J. D.; and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1): 50–80.

Logg, J. M.; Minson, J. A.; and Moore, D. A. 2019. AI-algorithm appreciation: People prefer algorithmic to human judgment. *Organizational behavior and human decision processes*, 151: 90–103.

Malatji, M. 2026. Bridging the AI divide in sub-Saharan Africa: Challenges and opportunities for inclusivity.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An Integrative Model of Organizational Trust. *Academy of Management Review*, 20(3): 709–734.

Meta. 2024. Llama 3 Model Card. Model card/release documentation.

Ng, D. T. K.; Leung, J. K. L.; Chu, S. K. W.; and Qiao, M. S. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2: 100041.

OpenAI. 2023. GPT-4 Technical Report.

Rae, I. 2024. The Effects of Perceived AI Use On Content Perceptions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, 1–14. Honolulu, HI, USA: Association for Computing Machinery.

Schaefer, K. E.; Chen, J. Y.; Szalma, J. L.; and Hancock, P. A. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, 58(3): 377–400.

Shekar, S.; Pataranutaporn, P.; Sarabu, C.; Cecchi, G. A.; and Maes, P. 2024. People overtrust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy. arXiv:2408.15266.

Steyvers, M.; Tejada, H.; Kumar, A.; Belem, C.; Karny, S.; Hu, X.; Mayer, L. W.; and Smyth, P. 2025. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2): 221–231.

Tully, S. M.; Longoni, C.; and Appel, G. 2025. Lower Artificial Intelligence Literacy Predicts Greater AI Receptivity. *Journal of Marketing*, 89(5).

Wainaina, P. K.; and Sun, Y. 2025. Educators' perceptions and willingness to integrate Generative Artificial Intelligence in teaching and research: evidence from Kenyan higher education. *Discover Education*, 4(1): 347.

Wasserman, H.; and Madrid-Morales, D. 2023. Engaging and disengaging with political disinformation on WhatsApp: A study of young adults in South Africa. Hate and Disinformation on WhatsApp: Global Perspectives (presentation/report PDF).