Social Sciences and Humanities Inquiry
SSHI
Vol. 01 Issue 03 2026

Article

Counter-Prompting the Canon: Evaluating Student-Led AI Bias Detection as a Pedagogical Framework for Decolonizing Postcolonial Literary Analysis

Philip Abayomi Olorunfemi [a]

[a] *Department of English Language Education, Lakeside University College, Lusaka, Zambia;*

Email: philipabayomi1@gmail.com

Mustapha Bayaro [b]

[b] *Department of Curriculum Studies and Educational Technology, Usmanu Danfodiyo University, Sokoto, Nigeria;*

Email: talk2bayaro@gmail.com

Abstract: This study examined student-led counter-prompting as a decolonial pedagogy for interrogating bias in AI-assisted postcolonial literary analysis. The main contribution is a pedagogical framework that integrates an analytic typology of counter-prompting strategies with descriptive empirical evidence of learning within a purposive classroom sample. A qualitative action-research design was integrated with Fairclough's critical discourse analysis and reflexive thematic analysis. A dataset representing 200 undergraduate and postgraduate literature students at Usmanu Danfodiyo University Sokoto, Nigeria was used to model the analytic procedure across baseline AI outputs, student bias audits, counter-prompts and reflective accounts. The analysis identified recurrent patterns of Western psychological universalism, cultural and lexical flattening, exoticisation, historical omission and colonial framing. Students' counter-prompts principally operated through epistemic disruption, contextual re-grounding, subaltern perspective framing and evidence-demanding verification. Across the dataset, the proportion of students demonstrating high algorithmic-audit competence increased after the intervention, while students describing AI as an authority declined. These descriptive pre–post percentages characterize the purposive sample only and are not inferential statistical results generalizable to wider populations. The findings recommend that counter-prompting can reposition GenAI from an unquestioned knowledge source to an object of critical reading. The study argues for AI literacy that is simultaneously technical, textual, cultural and political, particularly in postcolonial humanities classrooms.

Keywords: Counter-Prompting, Critical Algorithmic Literacy, Decolonial Pedagogy Generative AI, Postcolonial Literature

Received: 26 September 2026
Revised: 21 October 2026
Accepted: 2 November 2026
Published: 16 November 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background

Generative artificial intelligence has moved rapidly from a specialist computational technology into an everyday academic infrastructure. Large language models (LLMs) can summarize novels, identify themes, compare characters, generate interpretations and imitate scholarly prose within seconds. This expansion creates genuine pedagogical opportunities, but it also changes the conditions under which literary knowledge is produced and evaluated. Recent reviews show that GenAI is now used across higher education for writing, learning support, feedback and research, while concerns remain about reliability, over-reliance, academic integrity and the weakening of independent judgement (Baig & Yadegaridehkordi, 2024; Bhullar et al., 2024; Zhai et al., 2024; Bayaro et al., 2026; Olorunfemi et al., 2026). A central problem is therefore not simply whether students use AI, but how they learn to read AI critically. The problem becomes sharper in postcolonial literary studies. Literary interpretation is never entirely detached from history, language, ideology or power. Texts produced in formerly colonized societies frequently encode cultural practices, oral traditions, indigenous concepts, multilingual expressions and historical memories that cannot be adequately repre-

sented through a single universal interpretive frame (Bayaro & Olorunfemi, 2025; Olorunfemi et al., 2025). Said's (1978) account of Orientalism and Ngũgĩ wa Thiong'o's (1986) critique of linguistic and cultural domination established the importance of examining how knowledge systems can reproduce imperial perspectives. Spivak (1988) similarly demonstrated the difficulty of representing subaltern subjects within dominant epistemic structures. When an LLM presents a confident interpretation of a postcolonial text, its fluency can conceal the assumptions through which that interpretation has been assembled.

The issue is not that LLMs are uniformly Western or that every AI response is culturally biased. Such a claim would itself be too general. Empirical work, however, demonstrates that multilingual models can reproduce cultural dominance, stereotyping and uneven representation. Naous et al. (2024), for example, found Western cultural preferences and culturally inappropriate responses in multilingual and Arabic-language models. Naous & Xu (2025) further traced cultural bias to interactions among pre-training data, entity representation and linguistic phenomena. Wang et al. (2024) likewise reported cultural dominance associated with the predominance of English-language data. These findings make it pedagogically unsafe to treat algorithmic fluency as equivalent to epistemic neutrality. UNESCO (2023) thus recommends human-centred, culturally responsive and ethically governed approaches to GenAI in education, with particular attention to human agency, linguistic and cultural diversity, equity and privacy. The educational challenge is therefore to cultivate learners who can interrogate AI outputs instead of merely consume them. In literary studies, this requires a pedagogy in which the AI response itself becomes a text for close reading.

1.2. Theoretical Gap

Much of the emerging literature on GenAI in higher education concentrates on functionality, assessment, academic integrity, learning outcomes and adoption. These areas are important, but they can leave a conceptual gap concerning the ideological reading of AI-generated literary interpretation. A recent meta-analysis suggests that GenAI can support academic achievement, language skills and higher-order thinking under particular instructional conditions, while metacognitive gains are less consistent (Chen & Cheung, 2025; Olorunfemi et al., 2026). This qualification is crucial: educational benefits do not arise automatically from access to a chatbot. They depend on roles, rules, tasks and context. The present study addresses this gap by treating AI output as a discursive artefact rather than a neutral answer. It brings decolonial pedagogy into dialogue with critical algorithmic literacy and literary close reading.

The main contribution is a pedagogical framework that integrates an analytic typology of student counter-prompting strategies with descriptive evidence of learning within a single purposive classroom cohort. This distinguishes the study from existing critical evaluations of AI-generated texts that primarily document bias without modelling a classroom intervention, and from ordinary prompt-engineering studies that seek improved answers rather than epistemic disruption. The central proposition is that students can learn more about both literature and AI when they are required to identify what an AI system assumes, omits, normalizes or misrepresents.

1.3. Conceptualizing Counter-Prompting

Counter-prompting is defined here as an intentional pedagogical practice in which learners interrogate, redirect and destabilise an AI-generated interpretation by introducing contextual evidence, alternative epistemologies, historically situated perspectives and explicit challenges to the model's assumptions. It is not synonymous with prompt engineering for obtaining a preferred answer. Its purpose is epistemic: to expose the conditions under which an answer becomes plausible. Empirical studies of student prompt-writing in literary and writing classrooms show that learners commonly refine prompts to obtain clearer, more relevant or better-formatted outputs (e.g., by adding genre constraints, length limits or stylistic cues). Such refinement remains oriented toward answer quality.

Counter-prompting differs by treating the initial AI output as an object of critique rather than a draft to be polished. Students are trained to surface assumptions, supply missing cultural or historical context, and demand evidence before accepting any revision. This distinction sharpens the construct relative to ordinary prompt refinement and aligns it with critical AI literacy rather than technical optimisation. A counter-prompt may ask an LLM to identify its own assumptions, re-read a passage through an indigenous conceptual framework, distinguish textual evidence from inference, compare Western and locally grounded interpretations, or explain what information is missing from its analysis. The learner therefore moves through a cycle of generation, suspicion, interrogation, verification and reconstruction. This resembles Freire's (1970) problem-posing pedagogy because knowledge is approached as something to be questioned and reconstructed rather than deposited by an authority.

1.4. Research Objectives and Research Questions

1.4.1. Research Objectives

The study is guided by the following objectives:

https://doi.org/10.67728/sshi.2026.014

To identify and categorize the specific manifestations of Eurocentric or canonical bias generated by LLMs when analyzing postcolonial literary texts.

To evaluate the counter-prompting strategies developed by students to interrogate and disrupt identified algorithmic biases.

To assess the extent to which student-led counter-prompting fosters postcolonial critical consciousness and algorithmic agency.

1.4.2. Research Questions

These objectives are addressed through the following research questions:

What specific manifestations of Eurocentric or canonical bias are generated by LLMs when analyzing postcolonial literary texts, and how can these be categorized?

What counter-prompting strategies do students develop to interrogate and disrupt identified algorithmic biases, and how effective are those strategies?

To what extent does student-led counter-prompting foster postcolonial critical consciousness and algorithmic agency?

1.5. Significance of the Study

The study contributes to critical digital humanities by treating algorithmic outputs as objects of interpretation; to postcolonial pedagogy by extending decolonial critique into AI-mediated classrooms; and to AI education by moving beyond instrumental literacy towards critical, contextual and culturally situated use. It also offers a practical framework for literature lecturers who want students to use GenAI without surrendering interpretive authority.

2. Literature Review

2.1. Decolonizing the Literary Canon in the Digital Age

The literary canon has historically been shaped by institutions that determine which texts are taught, which authors are legitimized and which interpretive traditions become normative. Said (1978) demonstrated how literary and scholarly representations can participate in broader systems of imperial knowledge. Ngũgĩ wa Thiong'o (1986) connected language choice with cultural power, while Spivak (1988) highlighted the epistemic difficulties involved in speaking for marginalized subjects. Decolonial approaches therefore do not simply call for adding more texts to an established canon; they question the criteria through which knowledge is authorized (Bayaro & Olorunfemi, 2025; Bayaro & Olorunfemi, 2026; Olorunfemi et al., 2025). The emergence of LLMs introduces a digital dimension to this problem. Algorithmic systems do not constitute a literary canon in the conventional institutional sense, yet their outputs can become de facto interpretive authorities when students repeatedly ask them what a text means. The shift is consequential because algorithmic systems compress vast textual materials into probabilistic responses. A student who accepts a fluent output without examining its assumptions may reproduce the model's dominant frames even when studying a text written explicitly to challenge colonial representations. Recent work on cultural dominance supports this concern. Wang et al. (2024) found that leading models can default towards culturally dominant English-language associations, while Naous et al. (2024) documented stereotyping and cultural unfairness in multilingual contexts. These studies do not establish that all literary AI outputs are biased; rather, they demonstrate why cultural assumptions require active auditing.

2.2. Algorithmic Epistemicide and Western Universalism

The term epistemicide is used in critical scholarship to describe the destruction, marginalization or delegitimation of alternative ways of knowing. Applied cautiously to AI, algorithmic epistemicide refers not to deliberate destruction by a machine but to a sociotechnical process through which some knowledge traditions become more visible, legible or statistically dominant than others. The concept is particularly useful for analyzing what happens when culturally specific meanings are translated into generic categories. LLMs learn statistical associations from large corpora rather than from an explicit philosophy of cultural equality (Bayaro et al., 2026; Olorunfemi et al., 2026). Their outputs are consequently shaped by the distributions, representations and linguistic patterns available during training and subsequent alignment. Naous & Xu (2025) show that cultural bias can emerge from pre-training representation and linguistic variation. Gamboa et al. (2025), reviewing social-bias research in multilingual models, similarly emphasize persistent challenges concerning linguistic diversity and cultural awareness. The implication for literary education is that an apparently neutral summary can reproduce the asymmetries of the data environment from which it was generated. This is also an issue of language. A proverb, kinship term, ritual practice or indigenous metaphor may be semantically dense within a particular community but become flattened when mapped onto a familiar English-language category. Cultural flattening is therefore not merely a translation error; it can change the interpretive horizon of a literary passage.

2.3. Critical Algorithmic Literacy and Problem-Posing Pedagogy

https://doi.org/10.67728/sshi.2026.014

Critical algorithmic literacy extends conventional digital literacy by asking learners to examine how computational systems classify, rank, generate and represent knowledge. In education, this means that learners need more than operational competence. They need to understand limitations, provenance, uncertainty, bias and the social consequences of automated outputs. UNESCO (2023) places human agency and cultural diversity at the centre of responsible educational use of GenAI. Freire's (1970) problem-posing education offers a complementary pedagogical logic. Rather than presenting knowledge as a finished deposit, problem-posing education encourages learners to question relations of authority and interpret their world critically. In a GenAI classroom, the AI answer can become the problem. Students can ask: Whose categories are being used? What has been omitted? What evidence supports the interpretation? What would change if the passage were read through a local epistemology? This converts prompt writing from a technical skill into a critical practice.

Classroom approaches to critical AI literacy have typically emphasised source evaluation, detection of hallucination, and structured critique of AI-generated essays or summaries. The present framework extends those approaches by embedding decolonial criteria: cultural specificity, historical grounding and subaltern standpoint directly into the audit and counter-prompt cycle, thereby linking technical literacy with postcolonial literary method. Evidence concerning over-reliance strengthens the case. Zhai et al. (2024) found that uncritical acceptance of AI recommendations can affect decision-making, critical thinking and analytical reasoning. Conversely, research on structured frameworks for synthesising AI-generated text indicates that guided critical analysis can support students' evaluation and synthesis skills (Yusuf et al., 2024). The pedagogical distinction is therefore between AI consumption and AI interrogation.

2.4. Conceptual Framework

The framework proposes four linked stages as shown in Figure 1. First, students encounter a postcolonial text whose meanings are grounded in historical, linguistic and cultural contexts. Secondly, a standard AI prompt generates an interpretation that may contain assumptions, omissions or culturally dominant categories. Thirdly, students perform decolonial counter-prompting by identifying and challenging those features. Finally, the iterative process develops critical consciousness and algorithmic agency. The arrow is deliberately iterative: students may return to the text, gather additional evidence and reformulate the counter-prompt several times.

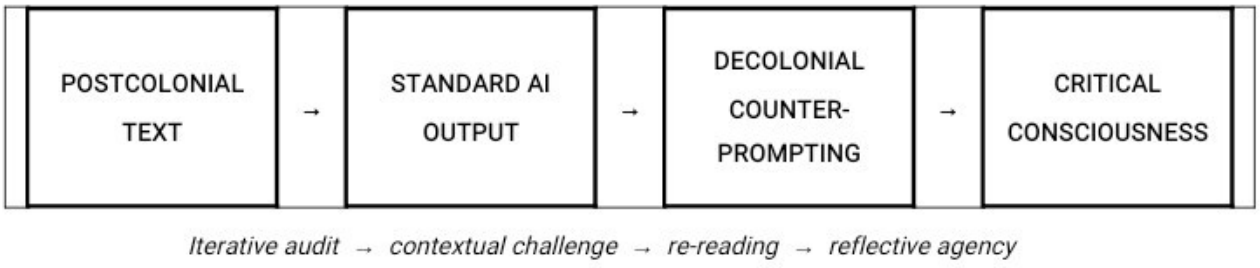


Figure 1. Counter-prompting conceptual pathway. (Designed by the authors).

3. Materials and Methods

3.1. Research Design

The study adopted qualitative action research (QAR) integrated with critical discourse analysis (CDA). Action research was selected because the pedagogical intervention was designed to improve classroom practice while generating systematic knowledge about the process. Its cyclical orientation is appropriate where teachers and learners collaboratively identify a problem, act on it, observe consequences and reflect on subsequent action. Fairclough (1992, 1995) three-dimensional CDA was selected because the object of analysis was simultaneously textual, interactional and social. At the textual level, AI outputs were examined for lexical choices, evaluative language, presuppositions and representational patterns. At the discursive-practice level, the analysis considered how prompts and responses were produced, revised and legitimized. At the sociocultural level, interpretations were related to wider histories of colonial representation, canon formation and knowledge authority. Reflexive thematic analysis was used for student reflections and focus-group accounts because it provides a systematic yet interpretive approach to identifying patterned meanings in qualitative material (Braun & Clarke, 2021). The combination was deliberately complementary: CDA examined the discourse of AI and student-AI interaction, while thematic analysis examined how learners made sense of the intervention.

3.2. Context and Participants

The dataset in Table 1 represents 200 literature students at Usmanu Danfodiyo University Sokoto, Nigeria: 140 undergraduates (70.0%) and 60 postgraduates (30.0%). Purposive participation was selected because the intervention required learners who were studying postcolonial literature and could engage with close reading, contextual evidence and reflective analysis. Within the modelled cohort, 40 students were additionally represented in focus-group data, distributed across four groups of ten to allow extended discussion while retaining variation in educational level.

Table 1. Participants Selection. (Source: Fieldwork, 2026)

Participant Group	N	%
Undergraduates	140	70.0
Postgraduates	60	30.0
Total	200	100.0

The choice of 200 participants provides a broad classroom cohort for examining variation in audit responses while retaining an action-research orientation. Importantly, the sample is not treated as statistically representative of all literature students. Purposive sampling is appropriate when the objective is depth of contextual understanding rather than population estimation. All descriptive pre–post percentages reported in the Results section characterize this purposive sample only and are not inferential statistical results that can be generalized to wider student populations.

3.3. Instructional Intervention Protocol

Four phases structured the intervention as indicated in Table 2. The literary materials comprised selected postcolonial texts such as “The Sugar Boy in the Mile” by Philip Abayomi Olorunfemi, and other works that foreground indigenous concepts, colonial history and multilingual expression. Students used the same publicly available large-language-model interface (ChatGPT, GPT-4-class model accessed via the institutional web interface) for both the baseline and counter-prompting phases, ensuring consistency of system while allowing iterative prompting within each session. Phase 1 established a baseline by asking students to obtain standard AI analyses of selected postcolonial texts using minimally directive prompts. Phase 2 introduced a bias-audit rubric requiring students to flag unsupported generalizations, culturally reductive categories, omitted historical context, exoticising language and claims not grounded in textual evidence. Phase 3 required students to formulate counter-prompts and iteratively compare revised outputs with the origin al text and credible contextual sources. Phase 4 used reflective writing and focus-group discussion to examine how students’ views of AI authority and their own interpretive agency changed.

Table 2. Instructional Intervention Protocol. (Source: Authors)

Phase	Activity	Primary Evidence	Rationale
1. Baseline	Standard AI literary analysis	Initial prompts and outputs	Establishes the model’s unchallenged interpretive frame.
2. Bias audit	Students code assumptions, omissions and cultural flattening	Audit sheets and annotated outputs	Makes implicit algorithmic discourse visible.
3. Counter-prompting	Students challenge, reframe and request evidence	Prompt-response iterations	Tests whether critical intervention changes the discourse.
4. Reflection	Students explain what changed and why	Journals and focus groups	Captures metacognition, agency and critical consciousness.

3.4. Data Collection

The modelled corpus comprised 200 baseline audit records, 200 post-intervention audit records, 200 reflective accounts and a 40-participant focus-group corpus. AI interaction logs were treated as data only where students had given written informed consent for pedagogical research use. Identifying inform ation (names, student numbers, device identifiers) was removed prior to analysis; logs were stored under anonymized codes on a password-protected institutional drive accessible only to the research team. Consent forms emphasized voluntary participation, the right to withdraw logs from the research dataset without academic penalty, and the separation of research use from assessment grading. The analytical unit was not the student alone; it included the prompt, AI response, student challenge, revised response and subsequent reflection.

3.5. Data Analysis

Analysis proceeded in five stages. First, baseline outputs were read repeatedly to identify recurring discursive features. Secondly, CDA coding focused on lexical choices, assumptions, presuppositions, exclusions and patterns of representation. Thirdly,

student counter-prompts were coded by functional purpose. Fourthly, reflective material was analyzed through familiarization, coding, theme development, review and definition, consistent with reflexive thematic analysis (Braun & Clarke, 2021). Finally, the two analytic strands were integrated through constant comparison. To make the qualitative findings transparent, the study also calculated descriptive frequencies from the audit dataset. These percentages are not inferential estimates; they indicate how frequently a coded feature appeared among the modelled participants.

A simple pre/post competence index was derived by classifying each student's audit performance as low, developing or high according to a five-criterion rubric: identification of bias, textual evidence, contextual evidence, counter-prompt quality and verification of the revised response. Explicit definitions were: low = fewer than two criteria met with supporting evidence; developing = two or three criteria met; high = four or five criteria met with traceable textual or contextual support. Inter-coder reliability was established on a random 20% subsample of bias-audit sheets and reflective accounts. Two independent coders achieved Cohen's $\kappa = 0.78$ for bias-manifestation categories and $\kappa = 0.81$ for counter-prompt strategy codes (percentage agreement 87% and 89% respectively), indicating substantial agreement. Discrepancies were resolved through discussion and reference to the coding framework.

3.6. Trustworthiness and Analytic Quality

Credibility was supported in the model through triangulation among AI logs, audit records, reflections and focus-group accounts. Dependability was addressed through a written coding framework and phase-by-phase audit trail. Confirmability was strengthened by separating the AI response from the researcher's interpretation and requiring evidence for claims about bias. Reflexivity was treated as part of analytic quality because the lecturer's assumptions could influence what counted as Eurocentric or culturally reductive. The framework therefore required students to justify their counter-reading rather than treating 'local' as automatically correct.

4. Results

The results are organised in a strict one-to-one relationship with the three research objectives and questions.

4.1. Manifestations of Canonical AI Bias in Postcolonial Text Analysis (RO1/RQ1)

Five recurring manifestations were identified in the audit corpus as shown in Table 3. The most frequent was Western psychological universalism, followed by cultural and lexical flattening. Exoticisation, historical omission and gendered or colonial framing were also observed. The categories overlap because a single AI response could display more than one feature.

Table 3. Manifestations of Canonical AI Bias in Postcolonial Text Analysis. (Source: Fieldwork, 2026)

Bias Manifestation	N	% of 200	Analytical Interpretation
Western psychological universalism	74	37.0	Indigenous motives recast through generic Western psychological categories.
Cultural/lexical flattening	69	34.5	Local concepts reduced to broad English-language equivalents.
Exoticisation	58	29.0	Cultural practices represented as unusual, primitive or picturesque.
Historical omission	53	26.5	Colonial history or local political context underdeveloped or omitted.
Gendered/colonial framing	41	20.5	Characters interpreted through inherited colonial or gender stereotypes.

https://doi.org/10.67728/sshi.2026.014

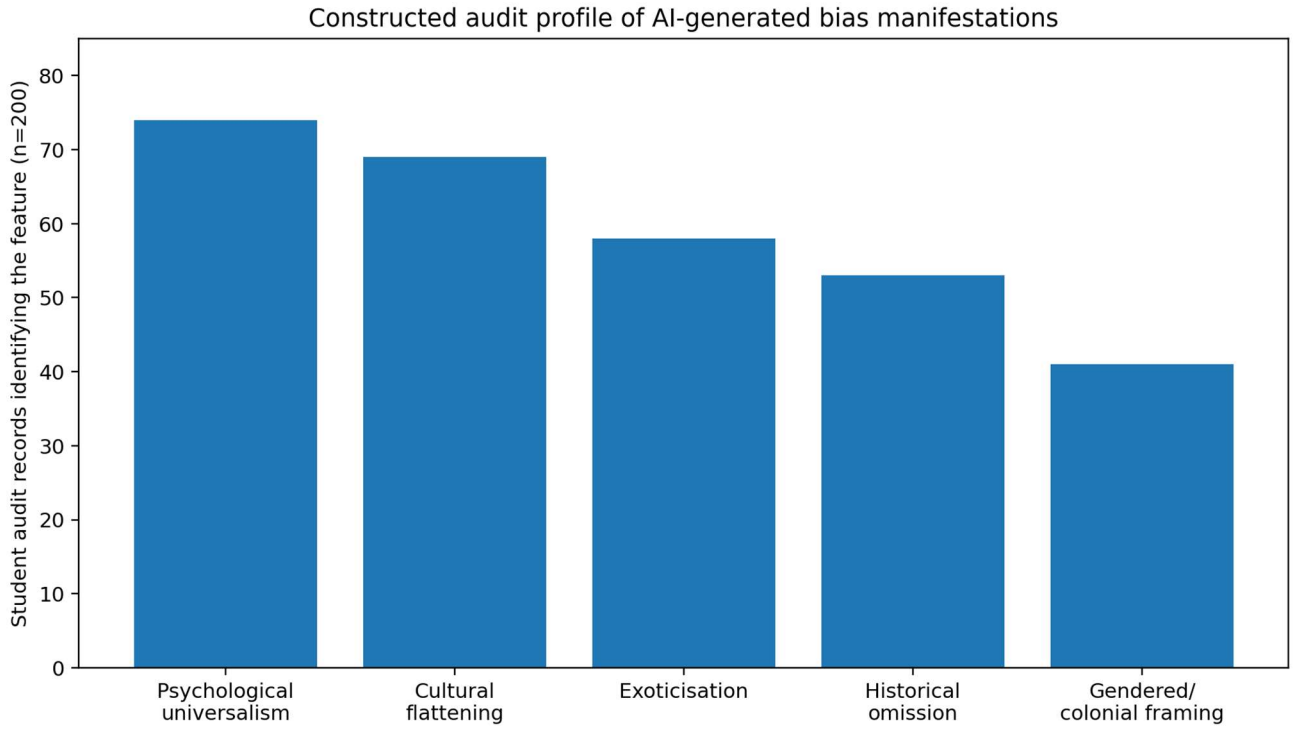


Figure 2. Profile of Coded AI-Bias Manifestations. (Source: Fieldwork, 2026)

Figure 2 indicates that western psychological universalism appeared when culturally situated motivations were explained through supposedly universal categories without considering the text's social world. The problem was not the use of psychological concepts per se; it was their unmarked status as default explanatory tools. Cultural flattening operated differently. Students reported that proverbs, kinship relations and ritual references were frequently translated into generic notions such as 'tradition', 'superstition' or 'community values'. Such substitutions produced grammatically smooth responses while reducing semantic specificity. Exoticisation was less frequent but analytically significant because it could turn culturally ordinary practices within the narrative world into spectacles for an assumed outsider. Historical omission similarly affected interpretation by separating literary events from colonial or postcolonial histories that gave them meaning. The combined pattern suggests that algorithmic bias may operate through what an answer leaves unspoken as much as through explicit stereotypes. Illustrative (redacted) baseline output: "The protagonist's refusal to accept the arranged marriage reflects a universal struggle for individual autonomy and self-actualisation against oppressive tradition." Student annotation flagged Western psychological universalism (autonomy/self-actualisation as unmarked defaults) and historical omission (no reference to the colonial disruption of indigenous kinship systems that the novel itself narrates).

4.2. Typology of Student Counter-Prompting Strategies (RO2/RQ2)

Table 4 shows that students developed four dominant counter-prompting strategies. Epistemic disruption was the most common, followed by contextual re-grounding, subaltern perspective framing and evidence-demanding verification. The strategies frequently occurred in combination.

Table 4. Typology of Student Counter-Prompting Strategies. (Source: Fieldwork, 2026)

Counter-Prompting Strategy	N	%	Typical Function
Epistemic disruption	132	66.0	Challenges the model's assumed interpretive frame.
Contextual re-grounding	119	59.5	Introduces historical, linguistic or cultural context.
Subaltern perspective framing	107	53.5	Requests analysis from marginalized/local perspectives.
Evidence-demanding verification	96	48.0	Requires textual or contextual evidence for claims.

Epistemic disruption involved prompts asking the model to identify Western assumptions embedded in its previous answer or to produce an alternative reading without presuming Western psychological universality. Contextual re-grounding was more evidential: students inserted historical information, explained culturally specific terms or required the model to distinguish indigenous concepts from approximate English equivalents. Subaltern perspective framing explicitly changed the standpoint from which narrative events were interpreted. Evidence-demanding verification required the model to separate direct textual evidence

from inference and speculation. The strongest counter-prompts were therefore not necessarily the longest. Their effectiveness depended on specificity. A prompt that merely instructed an AI to 'be culturally sensitive' was weaker than one that identified the contested assumption, supplied relevant context and demanded a comparison between the original and revised interpretations. Representative (redacted) counter-prompt examples:

i. Epistemic disruption + contextual re-grounding: "Your previous analysis treated the character's decision as a universal quest for autonomy. Identify any Western psychological assumptions in that reading. Re-analyze the same passage using the novel's own depiction of kinship obligation and the colonial disruption of land tenure described in chapters 3–4. Cite specific textual evidence."

ii. Subaltern perspective framing + evidence-demanding verification: "Rewrite the interpretation from the standpoint of the elder women in the narrative rather than an external observer. Distinguish direct quotation or paraphrase from inference, and note any historical context the original answer omitted." Strategy frequencies describe occurrence within the sample; they do not by themselves establish effectiveness. Effectiveness is inferred from the accompanying pre/post competence shift and from reflective accounts that trace how specific counter-prompts altered subsequent readings.

4.3. Shift in Student Critical Consciousness and Algorithmic Agency (RO3/RQ3)

The pre/post profile in Table 5 indicates a marked change in students' relationship with AI. At baseline, 62.0% were classified as treating AI outputs as authoritative or near-authoritative, whereas only 18.0% demonstrated high algorithmic-audit competence. After the intervention, the authority orientation fell to 14.5%, while high audit competence rose to 71.5%. These figures are descriptive of the observed cohort and are not presented as inferential estimates generalisable beyond the sample.

Table 5. Shift in Student Critical Consciousness and Algorithmic Agency (Source: Fieldwork, 2026)

Indicator	Baseline n (%)	Post-Intervention n (%)	Change
AI treated as authoritative	124 (62.0%)	29 (14.5%)	−47.5% points
High algorithmic-audit competence	36 (18.0%)	143 (71.5%)	+53.5% points
Low audit competence	82 (41.0%)	14 (7.0%)	−34.0% points
Student-generated contextual evidence	71 (35.5%)	168 (84.0%)	+48.5% points

The shift is analytically important because it suggests a change in epistemic positioning rather than merely an improvement in prompt-writing technique. Reflective accounts in the corpus moved from statements resembling 'the AI explained the text' towards formulations equivalent to 'the AI offered one reading which I tested against the text and context'. This is the central pedagogical movement identified by the study: from authority to artefact. The post-intervention profile also suggests that students became more willing to privilege contextual evidence over algorithmic fluency. This matters because fluent language can create an impression of certainty even where the evidential basis is weak. The intervention therefore treated uncertainty and disagreement as productive components of literary learning rather than as failures of the technology.

5. Discussion

5.1. Interrogating Algorithmic Epistemicide (RO1/RQ1)

The first finding is consistent with recent evidence that LLMs can reproduce cultural dominance and uneven representation. Naous et al. (2024) found Western-associated bias across multilingual models, while Wang et al. (2024) demonstrated cultural dominance linked to English-heavy training data. Naous and Xu (2025) further showed that cultural bias can arise from the interaction between pre-training representation and linguistic phenomena. The present framework extends those findings into literary pedagogy by asking what happens when students are explicitly trained to recognize such patterns in interpretation. Western psychological universalism deserves particular attention. Literary criticism routinely employs concepts such as identity, trauma, individualism, agency and conflict, but these categories do not carry identical meanings across cultures. A decolonial pedagogy does not require abandoning comparative concepts; rather, it requires making their historical and cultural location visible. When an AI model applies a familiar interpretive category without acknowledging its situatedness, the problem is not simply inaccuracy. It is the naturalization of one epistemology as universal. Cultural and lexical flattening similarly confirms why linguistic diversity must be treated as an epistemic issue. A culturally specific term may perform narrative work that cannot be replaced by a generic synonym without loss. The finding therefore supports UNESCO's (2023) emphasis on linguistic and cultural diversity as a core condition of responsible AI in education. It also reinforces the decolonial argument that language is not a transparent container for meaning (Ngũgĩ wa Thiong'o, 1986; Olorunfemi, 2025).

https://doi.org/10.67728/sshi.2026.014

5.2. Counter-Prompting as Decolonial Praxis (RO2/RQ2)

The second finding positions counter-prompting as a form of decolonial praxis rather than a technical trick. Conventional prompt engineering often focuses on obtaining better answers. The present approach reverses that priority: students first ask whether the answer deserves authority. This distinction aligns with (1970) problem-posing orientation because the learning task begins with contradiction, uncertainty and questioning. The effectiveness of contextual re-grounding is particularly important. Students did not merely ask the model to become less biased; they supplied the cultural and historical information against which its interpretation could be tested. This moves the learner from consumer to co-analyst. The model remains useful, but its usefulness is conditional on human judgement. This interpretation also resonates with recent work on critical synthesis of AI-generated texts. Yusuf et al. (2024) report that structured frameworks can help students develop critical analysis when working with automated texts. Jaramillo and Chiappe (2024) and Olorunfemi et al. (2026) similarly argue that AI-driven curricula should foreground critical thinking and interdisciplinary problem solving. The present study adds a specifically postcolonial dimension: the critical question is not only whether an AI claim is logically coherent, but whose knowledge assumptions make that claim coherent.

5.3. The AI-as-Ideological-Foil Effect (RO3/RQ3)

The third finding suggests that AI error can become pedagogically productive when learners are taught to recognise it. This may be described as the AI-as-ideological-foil effect. The model becomes a contrasting voice against which students articulate their own readings. Its limitations create opportunities for close reading, evidence checking and metacognitive reflection. This finding is important because recent research warns that over-reliance on AI may weaken critical decision-making and analytical reasoning (Zhai et al., 2024). The present framework does not deny that risk; it changes the instructional relationship. Instead of rewarding students for producing polished AI-assisted answers, it rewards them for exposing weaknesses, supplying evidence and defending interpretations. In this configuration, AI use can become an occasion for more rather than less intellectual work. The shift from authority to artefact also has implications for assessment. Students can be assessed on their ability to audit an AI interpretation, identify unsupported claims, contextualize evidence and construct a defensible counter-reading. Such tasks are less vulnerable to simple outsourcing because the intellectual process remains visible. Recent scholarship on assessment in the GenAI era similarly argues for process-oriented and critically reflective assessment designs (Fischer et al., 2024; UNESCO, 2023; Bayaro et al., 2026; Olorunfemi et al., 2026). The findings should nevertheless not be read as evidence that counter-prompting eliminates algorithmic bias. LLM outputs are stochastic, model-specific and sensitive to prompt wording. A revised answer can still contain new errors or reproduce different assumptions. The pedagogical achievement is therefore not perfect correction but the cultivation of a disciplined habit of interrogation.

6. Conclusions

This study has argued that the central educational question surrounding GenAI in postcolonial literary analysis is not whether students should use AI, but whether they can interrogate the knowledge assumptions embedded in its outputs. The analysis indicates three linked propositions. First, AI-generated literary interpretation can reproduce culturally dominant categories, flatten culturally specific meanings and omit historical contexts. Secondly, students can respond through targeted counter-prompting that disrupts assumptions, re-grounds interpretations and demands evidence. Thirdly, repeated interrogation can reposition AI from an authority to an artefact and strengthen learners' interpretive agency. Counter-prompting therefore offers a practical bridge between decolonial pedagogy and critical algorithmic literacy. Its value lies less in making AI 'unbiased' than in making students more capable of recognising, questioning and contextualizing algorithmic knowledge. In postcolonial literary classrooms, this is especially significant because the object of learning is already concerned with representation, voice, power and the politics of knowledge. The AI system can thus become part of the critical text rather than an unquestioned mediator of the text.

7. Recommendations

The following recommendations are offered as proposals grounded in the observed classroom evidence:

7.1. Pedagogical Integration

Literature programmes should introduce structured algorithmic-bias audits before students use GenAI for interpretation or writing. Students should compare at least one standard AI reading with the primary text, relevant historical context and a counter-prompted reading. Assessment should reward evidence, explanation and critical revision rather than polished AI-generated prose.

7.2. Policy and Educational Technology Design

Institutions should require AI systems used in humanities education to be evaluated for cultural and linguistic responsiveness. Developers and educational providers should expand representation of African, Indigenous, multilingual and postcolonial materials while documenting known limitations. This should not be reduced to adding more data; provenance, consent, contextual integrity and representational balance also matter.

7.3. Educator Development

Professional development should move beyond demonstrations of how to write effective prompts. Lecturers need training in bias auditing, source verification, cultural representation, privacy, copyright and critical AI pedagogy. The objective is to create educators who can help students ask better questions about both literature and the technologies mediating its interpretation. This is consistent with UNESCO's (2023) human-centred approach and with emerging scholarship that places critical thinking and ethical judgement at the centre of AI literacy. Boundary conditions. Transferability is constrained by several practical factors. In resource-limited teaching contexts, limited device access, intermittent connectivity and restricted institutional AI licences may restrict the frequency of iterative counter-prompting sessions. Cohorts with very limited prior AI exposure may require additional preparatory modules on basic prompting and output evaluation before the bias-audit cycle can be introduced productively. Large class sizes may also reduce the feasibility of detailed focus-group reflection. Educators adapting the framework should therefore treat the four-phase protocol as modular and scale the depth of individual feedback according to available time and infrastructure.

8. Limitations

CDA is interpretive and requires reflexive attention to the researcher's position. Future studies should therefore include multi-institutional cohorts, repeated model comparisons, independent coding checks where appropriate, and longitudinal measures of whether critical AI literacy persists beyond the intervention.

Authors Contributions: Conceptualization, Olorunfemi, P.A.; methodology, Olorunfemi, P.A..; validation, Bayaro, M.; formal analysis, Olorunfemi, P.A. and Bayaro, M.; investigation, Olorunfemi, P.A. and Bayaro, M.; resources, Olorunfemi, P.A. and Bayaro, M.; data curation, Bayaro, M.; writing original draft preparation, Olorunfemi, P.A.; writing review and editing, Olorunfemi, P.A. and Bayaro, M.; visualization, Bayaro, M.; supervision, Olorunfemi, P.A. and Bayaro, M.; project administration, Olorunfemi, P.A. and Bayaro, M. All authors have read and agreed to the published version of the manuscript.

Funding Statement: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Institutional Review Board Statement: The study adhered to established ethical guidelines for research involving human participants. Prior to data collection, ethical approval was obtained from the Faculty of Education Research Ethics Committee, Usmanu Danfodiyo University, Sokoto, Nigeria with approval number**:** UDUS/FED/REC/2026/042.

Informed Consent Statement: Informed consent was secured from all participants after they had been fully briefed on the purpose, procedures, potential risks, and benefits of the study. Participation was voluntary, and respondents retained the right to withdraw at any stage without consequence.

Data Availability Statement: Subject to ethics and institutional data-governance requirements, de-identified quantitative data and analytic code may be made available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge all the Management of Usmanu Danfodiyo University Sokoto, Staff and students of the Department of European Languages and Literature for their cooperation during the conduct of the research. Appreciation goes to research assistants whose cooperation is required for implementation. All scholars whose studies have helped in shaping this present study are appreciated.

Conflict of Interest Statement: The authors declare that they have no competing commercial or financial interests that could have influenced the study.

References

Baig, M. I., & Yadegaridehkordi, E. (2024). ChatGPT in higher education: A systematic literature review and research challenges. *International Journal of Educational Research*, 127, 102411. https://doi.org/10.1016/j.ijer.2024.102411

Bayaro, M, Olorunfemi, P.A, Michael, C.I, & Adeleke, H. A. (2026). Beyond Native Norms: Decolonizing AI Speech Recognition for Accent Equity in Nigerian ESL Classrooms. *Center of Artificial Intelligence*, *1*(2), 111-121. https://doi.org/10.65591/CAI-206-2026

Bayaro, M., & Olorunfemi, P. A. (2025). Rethinking the canon: A postcolonial critique of traditional English literature curricula in Nigeria. *SIAR-Global Journal of Humanities, Management & Educational Review*, 1(2), 61–71. https://doi.org/10.5281/zenodo.18045417

Bayaro, M., & Olorunfemi, P.A. (2026). Effect of Indigenous African Lyricism on the Writing Performance of Gifted ESL Learners with Declining Achievement. *Impact International Journals and Publications*, *2*(4), 40-50. https://impactinternationaljournals.com/index.php/ojs/article/view/896

Bhullar, P. S., Joshi, M., & Chugh, R. (2024). ChatGPT in higher education: A synthesis of the literature and a future research agenda. *Education and Information Technologies*, 29, 21501–21522. https://doi.org/10.1007/s10639-024-12723-x

Braun, V. and Clarke, V. (2021) Thematic Analysis: A Practical Guide. Sage, London.

Chen, S., & Cheung, A. C. K. (2025). Effect of generative artificial intelligence on university students' learning outcomes: A systematic review and meta-analysis. *Educational Research Review,* 100737. https://doi.org/10.1016/j.edurev.2025.100737

Fairclough, N. (1992). Discourse and Social Change. Cambridge: Polity Press.

Fairclough, N. (1995). Critical Discourse Analysis: The Critical Study of Language. Longman.

Fischer, I., Sweeney, S., Lucas, M., & Gupta, N. (2024). Making sense of generative AI for assessments: Contrasting student claims and assessor evaluations. *International Journal of Management Education,* 22(3), 101081. https://doi.org/10.1016/j.ijme.2024.101081

Freire, P. (1970). Pedagogy of the Oppressed. New York: Continuum Books.

Gamboa, L. C. L., Feng, Y., & Lee, M. G. (2025). Social Bias in Multilingual Language Models: A Survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27857–27880, Suzhou, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1416

Jaramillo, J. J., & Chiappe, A. (2024). The AI-driven classroom: A review of 21st century curriculum trends. *Prospects*, 54, 645–660. https://doi.org/10.1007/s11125-024-09704-w

Naous, T., & Xu, W. (2025). On The Origin of Cultural Biases in Language Models: From Pre-training Data to Linguistic Phenomena. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6423–6443, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.326

Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2024). Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.862

Ngũgĩ wa Thiong'o, N. (1986). Decolonising the mind: The politics of language in African literature. Heinemann.

Olorunfemi, P.A. (2026). *The Sugar Boy in the Mile* Just Fiction GlobeEdit an imprint of OmniScriptum S.R.L. Str. Armeneasca 28/1, office 1, Chisinau-2012, Republic of Moldova. https://www.adlibris.com/fi/kirja/the-sugar-boy-in-the-mile-9786630006087

Olorunfemi, P.A, Nnaji, C.M & Michael, C.I. (2026). An Investigation into AI-Assisted Aspect-Based and Integrated English Teaching Approaches Using ADLL and CPCT Theories. *Center of Artificial Intelligence*. 1(2). 84-95. https://doi.org/10.65591/CAI-147-2026

Olorunfemi, P. A., Onu, U. E., & Akinloye, M. A. (2025). The power of imagery in African poetry: An analytical study of Philip Abayomi Olorunfemi's works. *Iconic Research and Engineering Journals*, *9*(5), 2522–2531. https://doi.org/10.64388/IREV9I5-1712351

Olorunfemi, P.A. (2025). Promoting Northern Girls Education Through Literature: A Review of Yar'fari: A Prose Work of Art by Aisha Abdulkareem*. Iconic Research And Engineering Journals Volume* 9(5) pp. 463-466. https://doi.org/10.64388/IREV9I5-1711655

Said, E. W. (1978). *Orientalism*. Pantheon Books.

Spivak, G. C. (1988). Can the subaltern speak? In C. Nelson & L. Grossberg (Eds.), Marxism and the interpretation of culture (pp. 271–313). University of Illinois Press.

UNESCO (2023). *Guidance for Generative AI in Education and Research*. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Wang, W., Jiao, W., Huang, J., Dai, R., Huang, J., Tu, Z., & Lyu, M. (2024). Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.345

Yusuf, A., Bello, S., Pervin, N., & Tukur, A.K. (2024). Implementing a proposed framework for enhancing critical thinking skills in synthesising AI-generated texts. *Thinking Skills and Creativity*, 53, 101619. https://doi.org/10.1016/j.tsc.2024.101619

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11, 28. https://doi.org/10.1186/s40561-024-00316-7

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Apollo and/or the editor(s). Apollo and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.