JOURNAL OF ACADEMIC RESEARCH FOR HUMANITIES
ISSN 2663-192x

ORCID of JAR: https://orcid.org/0009-0000-0723-9485
CrossRef DOI Number: 10.67244/jar.bwo-researches.v6i3.a623
Link to the Paper: https://jar.bwo-researches.com/index.php/jarh/article/view/623
Edition Link: Journal of Academic Research for Humanities JARH, 6(3) Jul-Sep 2026
HJRS Link: Journal of Academic Research for Humanities JARH (HEC-Recognised for 2026-2027)
YouTube Link for the Paper Video Overview: https://youtu.be/h05wsXF-_fI

Evaluating Cultural Balance and Representation in ChatGPT and DeepSeek-Generated EFL Materials

Corresponding &Author 1:	**Idrees**, Master's Student, Applied Linguistics, College of Foreign Languages and Cultures, Chengdu University of Technology, Chengdu, Sichuan, China, idreesafridi887@gmail.com, https://orcid.org/0009-0000-2972-2769
Author 2:	**Liu Yongzhi**, Dean of College of Foreign Languages and Cultures, Chengdu University of Technology, Chengdu, Sichuan, China, lyzhi@cdut.edu.cn, https://orcid.org/0009-0009-9298-7335

Paper Information

Citation of the paper:
(JARH)Idrees., Yongzhi, L., (2026). Evaluating Cultural Balance and Representation in ChatGPT and DeepSeek-Generated EFL Materials. In *Journal of Academic Research for Humanities, 6(3), 36–47.*

QR Code for the Paper:

Subject Areas for JARH:
1 Cultural Representation
2 AI Model
3 Social Sciences & Humanities

Timeline of the Paper at JARH:
Received on: 23-07-2026.
Reviews Completed on: 30-07-2026.
Accepted on: 04-08-2026.
Online on: 05-08-2026.

License:
Creative Commons Attribution-Share Alike 4.0 International License

Recognised for BWO-R:
HJRS HEC Journal Recognition System

Jointly Published by BWO Research INTL:
BWO RESEARCHES (PVT) LTD INTERNATIONAL

CrossRef DOI Image of the paper:
Crossref

Abstract

Generative artificial intelligence is increasingly used to create English as a Foreign Language (EFL) reading materials, yet fluent output may still simplify cultural groups. This study compared 24 classroom-oriented passages generated by ChatGPT GPT-5.6 and DeepSeek-V4 from 12 identical prompts submitted on 22 July 2026. Each prompt requested a 220–250-word intermediate-level passage representing Chinese, Pakistani, and mainstream English-speaking Western perspectives fairly. Directed qualitative content analysis examined inclusion, balance, intragroup variation, specificity, essentialization, the collectivist-individualist binary, evaluative hierarchy, stereotyping risk, intercultural sensitivity, and prompt compliance. ChatGPT produced 2,763 words and met the requested range in all 12 passages. DeepSeek produced 3,095 words and met the requested range in five passages. Both systems included the three perspectives and promoted respectful communication. ChatGPT used more qualifications and acknowledged internal variation more consistently. DeepSeek supplied more named cultural detail but relied more often on broad contrasts that framed Chinese and Pakistani contexts as collective and Western contexts as individualistic. The findings show that cultural inclusion does not ensure balanced representation. The proposed review criteria help teachers and curriculum developers evaluate variation, specificity, stereotyping risk, and classroom suitability before using AI-generated materials.

Keywords: culture; representation; ChatGPT; DeepSeek; EFL

1. Introduction

Generative AI now supports lesson planning, explanations, exercises, feedback, and reading-text production in English language education. Systems such as ChatGPT and DeepSeek respond quickly to instructions about topic, length, level, and activity type. Their value, though, depends on more than grammatical accuracy. Reading materials also communicate assumptions about family, education, politeness, work, gender, identity, and acceptable behaviour. AI-generated EFL passages therefore operate as cultural representations as well as language resources.

International guidance stresses human oversight, cultural diversity, teacher capacity, and a clear pedagogical purpose (UNESCO, 2023; OECD, 2026). Equity, access, preparation, and governance also shape responsible educational use (Varsik & Vosberg, 2024). These concerns matter because culture is internally diverse. Generation, class, religion, gender, education, location, migration, institution, and personal choice all affect practice. A passage may mention several societies and still reduce them to fixed contrasts such as collective versus individual, hierarchical versus egalitarian, or indirect versus direct (Adilazuarda et al., 2024).

Model research supports close cultural review. Cultural prompting improves alignment but does not remove model defaults (Tao et al., 2024). Alignment also changes with prompt language and training-data composition (AlKhamissi et al., 2024), while English-cultural dominance may appear when other cultural knowledge is expected (Wang et al., 2024). In ELT, Global Englishes challenges native-speaker-centred norms and calls for broader linguistic and cultural models (Rose et al., 2021). Automated systems may reproduce narrow pronunciation norms (Jeon et al., 2024), so teacher agency, careful prompting, and contextual review remain necessary (Lee et al., 2025; Moorhouse, 2024).

Recent educational studies reinforce the need to inspect complete AI outputs rather than rely only on general claims about usefulness. AI-supported comprehensible input does not always provide the most suitable level of challenge (Lashari et al., 2026). Studies also report dependency and reliability concerns in educational use, different ideological patterns across generative systems, and the need to preserve human creativity and learner autonomy (Shah et al., 2024; Minhas & Salim, 2025; Shahid, 2025). These findings support direct material-level analysis and continued teacher review.

This study compares complete classroom-ready passages generated by ChatGPT and DeepSeek under identical culturally explicit prompts. The systems were selected because both were accessible general-purpose platforms developed in different national and organizational contexts. The prompts requested fair representation of Chinese, Pakistani, and mainstream English-speaking Western perspectives and prohibited stereotypes. The study examines prompt compliance alongside how similarities and differences are constructed. These labels are analytical categories, not claims that the named societies share one uniform culture.

1.1 Problem Statement

Fluent, classroom-ready passages can conceal cultural simplification. Repeated descriptions of China and Pakistan as collective, formal, hierarchical, and indirect, alongside Western societies as independent, egalitarian, and direct, may turn complex practices into national types. Teachers therefore need evidence about how current systems represent culture even when prompts request fairness and stereotype avoidance.

1.2 Research Objectives

1. Examine how both systems represent the three requested perspectives.
2. Identify balance, specificity, intragroup variation, essentialization, and stereotyping risk.
3. Compare prompt compliance and observable classroom-readiness features.
4. Develop practical review criteria for teachers and curriculum developers.

1.3 Research Questions

1. How do ChatGPT and DeepSeek represent the

requested cultural perspectives?
2. What forms of generalization, balance, and intragroup variation appear?
3. How do the systems differ in specificity, sensitivity, fairness, compliance, and material-level pedagogical effectiveness?
4. What are the implications for classroom practice, curriculum design, professional development, and responsible AI policy?

1.4 Significance of the Study

The study distinguishes cultural inclusion from cultural balance and provides a matched comparison of two widely accessible systems. It also separates cultural specificity, which may support learning, from essentialization, which presents variable practices as fixed group traits. The framework offers teachers, curriculum developers, and educational decision-makers practical criteria for reviewing internal diversity, evaluative hierarchy, stereotyping risk, and classroom suitability. For learners, it supports critical comparison and contextual interpretation rather than memorized national traits.

2. Literature Review

Language materials select whose experiences, institutions, and practices appear worthy of study. Source, target, and international cultures have long shaped EFL materials (Cortazzi & Jin, 1999), while intercultural competence emphasizes curiosity, interpretation, interaction, and critical awareness rather than fixed facts (Byram, 1997). Learners should therefore examine how a text positions communities, whose perspective is treated as normal, and whether difference is explained or merely listed. Recent reviews still report imbalanced representation, surface-level treatment, and limited critical engagement (Zhang et al., 2024). Similar patterns appear in undergraduate materials, where visible cultural content can receive more attention than attitudes, interpretation, interaction, and critical knowledge (Balouch et al., 2025). These concerns remain relevant when teachers supplement or replace textbooks with generated passages.

Generative AI extends these concerns. Large language models reproduce patterns in training data without human understanding (Bender et al., 2021). In language education, they provide useful support but require pedagogical control, adaptation, and distributed teacher-learner agency (Kohnke et al., 2023; Godwin-Jones, 2024; Moorhouse & Wong, 2025). Global Englishes-oriented material development also faces authenticity and implementation challenges (Lo, 2025). Recent educational studies identify uneven challenges, dependency concerns, differing ideological patterns, and the continuing value of human creativity (Lashari et al., 2026; Shah et al., 2024; Minhas & Salim, 2025; Shahid, 2025).

2.1 Fairness, Transparency, Explainability, and Responsible AI Governance

Responsible use requires traceable human decisions. UNESCO (2023) emphasizes oversight and cultural diversity; explainability must be meaningful to educational stakeholders (Khosravi et al., 2022). Accountability is also needed when authority shifts among staff, students, platforms, and corporations (Bearman et al., 2023). Governance should translate principles into procedures for oversight, fairness, privacy, documentation, and correction (Papagiannidis et al., 2025). Operational review should consider accountability, human oversight, transparency and explainability, fairness and non-discrimination, and privacy and data governance. For EFL materials, reviewers should record the model, date, prompt, retention rule, substantive revisions, and reasons for accepting or rejecting culturally sensitive content. Such records make later checking, correction, and withdrawal possible when model behaviour or institutional requirements change.

2.2 Prompt Design, Multilingual Prompting, and Culturally Adaptive Generation

Prompt design affects cultural representation before a topic is developed. Anthropological and culture-specific prompting can improve alignment, although gains remain partial (AlKhamissi et al., 2024; Tao et al., 2024). Culture-aware prompts reduce but do not remove English-cultural dominance (Wang et

al., 2024). Culturally focused data and community knowledge bases improve coverage of internal variation (Li et al., 2024; Shi et al., 2024). Effective prompts should identify the language and learning context, require variation within groups, request evidence or source attribution where appropriate, and state the limits of national labels. The present prompts named three perspectives, requested fair representation, and prohibited stereotypes, but they did not request source attribution or use languages other than English. This limitation is important because prompt language itself may alter alignment.

2.3 AI Literacy Frameworks for Language Teachers

Teacher AI literacy combines technological competence, pedagogical knowledge, ethical use, prompt design, response evaluation, and contextual adaptation (UNESCO, 2024; Ma et al., 2024; Pan & Wang, 2025; Lee et al., 2025). These competencies are needed because polished output can still contain cultural omissions, defaults, or hierarchy.

2.4 Human Evaluation of AI-Generated Instructional Materials

Cross-cultural benchmarks report stronger performance for highly represented cultures and weaker adaptability in many Global South settings (Myung et al., 2024; Rao et al., 2025). Western-entity preference and stereotyping also appear in Arabic outputs, with part of the imbalance linked to pre-training representation, lexical ambiguity, and tokenization (Naous et al., 2024; Naous & Xu, 2025). National labels can conceal local, dynamic, and intragroup differences (Zhou et al., 2025). Benchmark scores alone therefore cannot establish instructional quality because a model may perform well on selected questions while producing reductive classroom passages. Teachers and students apply partly different criteria to AI-generated explanations, with teachers giving greater weight to pedagogical effectiveness and students emphasizing clarity and relatability (Lee & Song, 2024). Human review should examine linguistic accessibility, cultural quality, pedagogical fit, learner positioning, critical-discussion potential, and stereotyping risk. Sufficient review must therefore involve people who understand the subject, the target learners, and the represented communities.

2.5 Theoretical and Analytical Framework

The framework combines Global Englishes with intercultural competence. It examines how texts distribute linguistic authority and cultural value while distinguishing inclusion, balance, specificity, internal variation, and essentialization within one analytical process.

2.6 Research Gap

Prior studies often use surveys, benchmarks, isolated questions, pronunciation recognition, or textbook analysis. Fewer compare complete EFL passages generated by Western- and Chinese-developed systems under identical cultural prompts. This study addresses that gap through 24 matched passages and connects cultural analysis with prompt compliance, teacher review, and responsible governance.

3. Research Location

The data were generated and organized in Chengdu, Sichuan, China, on 22 July 2026. No participants, classrooms, student records, or institutional documents were involved.

4. Research Methodology

The study used comparative directed qualitative content analysis supported by descriptive numerical analysis. Directed analysis applies concepts established in prior theory and refines them during interpretation (Hsieh & Shannon, 2005). Purposive sampling was suitable because the study selected culturally salient EFL topics for comparison rather than statistical generalization.

4.1 Data Collection

Twelve identical English prompts were submitted in fresh conversations to ChatGPT GPT-5.6 through a paid web account and DeepSeek-V4 through a free web account. Only the first response was retained, with no regeneration or clarification. Topics were family relationships, university life, teacher-student communication, hospitality, workplace behaviour, gender roles, festivals, food practices, politeness, respect for older people, individual and collective values, and intercultural misunderstanding. Each prompt

requested a 220–250-word intermediate-level passage, fair representation of all three perspectives, stereotype avoidance, a short title, and three comprehension questions.

The topics were selected because they recur in university EFL materials and invite comparison of everyday practices, values, and communication. Using the same prompt wording for both systems created matched pairs and reduced variation caused by topic choice or instruction length. Fresh conversations limited carry-over from earlier exchanges, while retaining only the first response avoided selective regeneration. The model labels and account types were recorded from the interfaces on the collection date so that later studies can repeat the procedure under new versions.

Table 1. Matched Topics and Passage Codes

No.	Codes	Topic
1	C1 / D1	Family relationships
2	C2 / D2	University life
3	C3 / D3	Teacher-student communication
4	C4 / D4	Hospitality
5	C5 / D5	Workplace behaviour
6	C6 / D6	Gender roles
7	C7 / D7	Festivals
8	C8 / D8	Food practices
9	C9 / D9	Politeness
10	C10 / D10	Respect for older people
11	C11 / D11	Individual and collective values
12	C12 / D12	Intercultural misunderstanding

Source: Researcher-generated dataset collected on 22 July 2026.

4.2 Units of Analysis

The complete passage was the main unit; sentences and clauses supported analysis of hedging, generalization, evaluation, and comparison. Comprehension questions were reviewed for compliance and pedagogical focus but were not coded as cultural descriptions. The passages were labelled C1-C12 for ChatGPT and D1-D12 for DeepSeek. Material-level pedagogical effectiveness refers only to observable design features, including length control, required format, cultural inclusion, and classroom-oriented questions. It does not refer to measured learning gains, learner motivation, or classroom achievement because no students participated.

4.3 Coding Framework

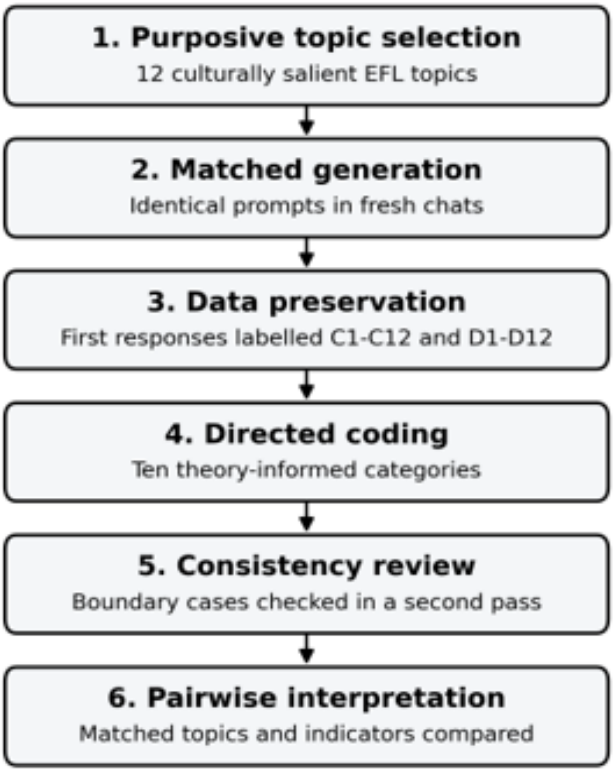
Ten categories drew on intercultural competence, Global Englishes, and critical AI literacy (Byram, 1997; Rose et al., 2021; Zhang et al., 2024; Jeon et al., 2024; Ma et al., 2024; Pan & Wang, 2025). Cultural inclusion recorded whether all three requested perspectives appeared. Representational balance examined comparable space, detail, and evaluative treatment. Intragroup variation captured regional, generational, institutional, religious, class, or personal differences. Specificity recorded named practices, events, objects, or concepts. Essentialization identified uniform or predictable claims about a group. The collectivist-individualist binary recorded repeated placement of China and Pakistan against Western societies. Evaluative hierarchy identified implied superiority, while stereotyping risk covered nationality-behaviour links without adequate qualification. Intercultural sensitivity recorded clarification, dialogue, respect, and avoidance of quick judgment. Prompt compliance covered length, level, inclusion, title, and comprehension questions.

4.4 Analytical Procedure and Consistency

The corresponding author coded all 24 outputs in two passes, so no intercoder statistic was calculated. The first pass applied the written categories; the second checked boundary cases, especially specificity versus essentialization. Topic pairs were then reviewed side by side, and word counts and compliance indicators were recorded. C1 was coded for variation because it named age, education, religion, location, income, and choice. D7 was coded for specificity because it named festivals and practices; such detail counted as essentializing only when framed as a general rule. D10 was coded for essentialization and stereotyping risk because "elders are never isolated" used absolute wording. C9 was coded for sensitivity because it treated direct and indirect requests as context-dependent and encouraged clarification. Frequencies are passage-level descriptive summaries, not inferential estimates. The

coding assessed model representations of the prompt categories, not the actual cultures of China, Pakistan, or Western societies.

Figure 1. Analytical Workflow



Note: The corresponding author completed two coding passes before pairwise interpretation.

4.5 Ethical Considerations

The research involved no participants or identifiable personal data. Prompts were researcher-written, outputs were retained unchanged, and model labels, access types, date, prompt structure, and first-response retention were documented. Findings apply to the stated interfaces and collection date because model behaviour changes over time.

5. Results

Both systems met the surface requirement of including Chinese, Pakistani, and mainstream English-speaking Western perspectives and ended every passage with three comprehension questions. Differences appeared in length control, qualification, cultural detail, internal variation, and the strength of national contrasts.

5.1 Prompt Compliance and Passage Length

ChatGPT produced 2,763 passage words across 12 texts, ranging from 223 to 239 words, with a mean of 230.3. All 12 met the requested range. DeepSeek produced 3,095 words, ranging from 236 to 297, with a mean of 257.9. Five passages, or 41.7%, met the range. Explicit intragroup variation appeared in 11 ChatGPT passages and six DeepSeek passages.

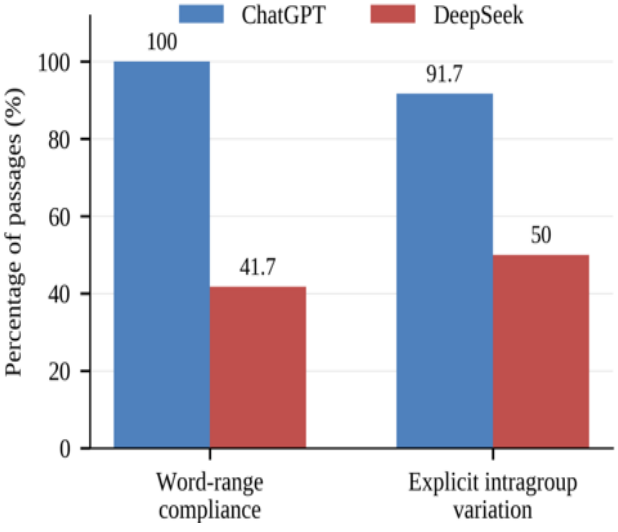
Length compliance matters because the prompts were designed for intermediate classroom use and requested comparable reading demands. DeepSeek exceeded the range in seven passages, which reduced direct comparability even though the texts remained broadly usable. Both systems supplied titles and three comprehension questions in every case. Surface compliance was therefore high, but the difference in length control shows that meeting visible format requirements should be checked separately from cultural quality.

Table 2. Comparative Descriptive Indicators

Indicator	ChatGPT	DeepSeek
Total words	2,763	3,095
Mean words	230.3	257.9
Word range	223-239	236-297
Within 220–250 words	12 (100%)	5 (41.7%)
Intragroup variation	11 (91.7%)	6 (50.0%)

Source: Researcher-generated dataset collected on 22 July 2026.

Figure 2. Word-Range Compliance and Intragroup Variation



Note: Values show the percentage of passages meeting each indicator.

5.2 Balanced Inclusion and Internal Variation

ChatGPT often began with a shared activity and then qualified examples. C1, for instance,

linked family differences to age, education, religion, location, income, and personal choice. This structure placed variation before interpretation. DeepSeek more often used three cultural blocks, with separate paragraphs for China, Pakistan, and Western societies. The structure gave each category visible space but also made national differences appear stable. D3 associated Chinese and Pakistani classrooms with authority, formality, and cautious disagreement, while Western classrooms were described as egalitarian, open, and informal. A respectful ending did not fully remove the hierarchy created by the main comparison. Explicit intragroup variation appeared across almost all ChatGPT topics but was concentrated in selected DeepSeek passages on gender, festivals, food, older people, and individual values.

5.3 Cultural Caution, Certainty, and Specificity

ChatGPT frequently used "some," "many," "may," "often," "depends," and "varies." C9 explained that direct and indirect requests could both be polite in context, and C7 linked festival participation to belief, family background, and personal choice. DeepSeek used stronger declarations, including "hospitality is considered a sacred duty," "seniority commands respect," "refusals are rarely direct," and "elders are never isolated." Such statements were memorable but left little space for migration, conflict, institutional care, changing practice, or inequality. DeepSeek also supplied more named detail, including filial piety, red envelopes, Eid, halal practices, naan, roti, Thanksgiving, and Confucian philosophy. This detail supported vocabulary and curiosity but became risky when contextual practices were presented as general rules. ChatGPT usually paired examples with advice to observe, ask, and avoid quick judgments. Its passages were therefore safer but sometimes less culturally substantive.

The food and festival passages illustrate this distinction. DeepSeek named more practices and objects, which made the texts vivid, but some statements lacked a region, setting, generation, or source. ChatGPT named fewer concepts and more often followed them with reminders that participation varies by belief, family background, and personal choice. In C8, learners were encouraged to observe and ask about utensils and dietary needs. The contrast suggests that specificity is educationally useful only when it is presented as an example rather than a predictable script for every member of a group.

5.4 The Collectivist-Individualist Binary

Both datasets contrasted China and Pakistan with Western societies. The first two were associated with extended family, consultation, hierarchy, shared meals, indirectness, and respect for authority; Western settings were linked to independence, first names, direct speech, personal space, and individual choice. DeepSeek used this binary more consistently, especially in the family and values passages, where collective duty and elder authority were placed against self-sufficiency and individual choice. ChatGPT also reproduced the contrast in several topics, but C11 directly stated that no society is fully individual or collective and added examples of people combining personal choice with family responsibility. Qualification therefore weakened, but did not eliminate, the binary.

5.5 Intercultural Message and Pedagogical Tension

Most passages ended with clarification, respectful questions, patience, or avoidance of quick judgment. ChatGPT commonly advised learners to ask about dietary needs, personal boundaries, or communication preferences. DeepSeek also promoted mutual respect, but positive conclusions sometimes followed several categorical descriptions and therefore did not fully revise the earlier framing. The fictional misunderstanding passages showed the tension clearly: national characters carried familiar expectations about punctuality, directness, silence, planning, mediation, or personal space, then resolved conflict through dialogue. The conflict-resolution message was useful, but the characters still carried recognizable national communication stereotypes. Comprehension questions mainly

tested factual recall, so they could reinforce the passage categories as correct answers rather than invite learners to question evidence, omission, or alternative experiences.

Positive endings were strongest when they changed how learners were asked to respond. C4 recommended asking about dietary needs and boundaries, C9 promoted clarification before judging directness, and C10 linked respect for older people with dignity and choice. DeepSeek passages on university life, workplaces, and politeness also encouraged exchange and mutual respect. Yet when a final sentence about harmony followed several broad national claims, it functioned more as a closing correction than as a change to the passage structure. Pedagogical value therefore depends on whether intercultural sensitivity shapes the whole text, not only its final sentence.

5.6 Summary of Findings

ChatGPT showed stronger length control, qualification, and recognition of internal variation. DeepSeek provided richer named detail but higher essentialization risk. Both achieved inclusion and positive tone, yet neither fully avoided familiar national binaries. Annexure C summarizes the coding patterns.

6. Discussion

The findings show that inclusion is not the same as balance. Both systems mentioned all requested perspectives, but their representational strategies differed. ChatGPT foregrounded hedging, shared values, and internal variation; DeepSeek foregrounded national comparison and concrete detail. Counting cultural mentions would therefore miss important differences in wording, organization, and evaluation. This supports earlier evidence that cultural prompting improves alignment without removing defaults (Tao et al., 2024; AlKhamissi et al., 2024; Wang et al., 2024). The anti-stereotype instruction improved tone, yet familiar cultural contrasts remained in the body of many passages.

DeepSeek's specificity offered educational value and risk. Learners may benefit from named festivals, foods, values, and practices because such references support vocabulary learning and cultural curiosity. Learning becomes reductive, though, when a context-bound example is presented as a national rule. ChatGPT's caution reduced stereotyping but sometimes produced less substantive cultural knowledge. The comparison therefore does not justify a simple claim that one system is superior. It reveals a trade-off between cultural richness and categorical simplification and shows why teachers should combine useful detail with clear qualification and verification.

The recurring association of Western settings with directness, independence, equality, and efficiency may position Western communication as an implicit classroom target, while formality and deference appear as practices learners should move beyond. This concern aligns with Global Englishes research on narrow automated norms (Jeon et al., 2024; Lee et al., 2025). Teacher judgment and contextual adaptation remain central (Godwin-Jones, 2024; Lo, 2025; Moorhouse & Wong, 2025). Human evaluation should therefore check not only correctness and readability but also who receives detail, authority, desirability, and internal diversity (Lee & Song, 2024). Reviewers should ask whether hedging is distributed equally, whether named examples are verified, and whether the text treats national categories as explanatory agents. Comprehension questions should ask learners to identify evidence, omissions, alternative experiences, and limits of national claims rather than reproduce the passage as unquestioned cultural knowledge.

These patterns resemble long-standing material-design problems rather than a completely new form of bias. Textbook research has also found imbalanced representation and limited critical intercultural engagement (Zhang et al., 2024; Balouch et al., 2025). Generative systems increase the speed and scale at which such materials can be produced, and their fluency can make simplification harder to notice. A complete passage with a title and questions may appear ready for use even when

its examples are unevenly qualified. The practical response is not to reject AI-generated content, but to apply a documented review process before classroom adoption.

7. Conclusion

The matched comparison found that ChatGPT met the requested word range in all passages and more consistently acknowledged internal variation. DeepSeek supplied more cultural detail but used broader national claims and a more stable collective-individual contrast. Both systems included all perspectives and promoted respectful interaction. The central finding is that cultural inclusion does not ensure cultural balance. A passage can mention every requested group yet still distribute caution, authority, detail, and desirability unevenly. Teachers should retain useful examples, qualify broad claims, add internal variation, verify culturally specific information, and redesign recall questions to support critical comparison. The review framework offers a transparent basis for repeating this evaluation across new models, prompt languages, and classroom contexts.

8. Recommendations

1. Review every AI-generated cultural passage before classroom use.
2. Require regional, generational, institutional, and individual variation in prompts.
3. Replace absolute wording with contextual descriptions and stated limits.
4. Add questions about evidence, omission, representation, and alternative experiences.
5. Combine AI output with locally produced texts, authentic media, and learner experience.
6. Include critical AI literacy and cultural review in teacher development and policy.
7. Document the prompt, language, model label, date, retention rule, and human revisions.
8. Assign accountable human reviewers and a process for correcting or withdrawing unsuitable content.

9. Limitations and Future Research

The study used a small purposive dataset collected on one date from two web interfaces. Hidden system instructions, sampling settings, and model updates may alter results. One researcher completed the coding, so independent reliability was not tested. English prompts and the broad "mainstream English-speaking Western" category also encouraged national comparison and grouped diverse societies. No teacher, learner, or cultural-specialist panel assessed the passages, classroom outcomes were not measured, and factual claims were not verified by representatives of the communities discussed. The study also did not compare passage-level judgments with an automated cultural benchmark. Future studies should compare prompt languages and model versions, involve diverse reviewers, test single-step and culturally adaptive prompting, verify factual claims, and combine qualitative coding with corpus and benchmark measures. Researchers should also report how different evaluator groups reach decisions and whether they apply the same criteria.

10. Innovation and Research Contribution

The study compares complete classroom-oriented passages rather than isolated benchmark items and separates inclusion from balance and specificity from essentialization. Its review criteria support replication across systems, languages, and educational settings.

11. Declarations

11.1 Funding

The study received no external funding.

11.2 Conflict of Interest

The authors declare no conflict of interest.

11.3 Data Availability

The prompts and 24 first-response passages may be provided as supplementary material or obtained from the corresponding author.

11.4 Generative AI Disclosure

ChatGPT and DeepSeek were the objects of analysis. Generative AI also supported language editing and formatting. The corresponding author reviewed the dataset, analysis, and references and accepts responsibility for the manuscript.

11.5 Author Contributions

Idrees: Conceptualization, data collection,

analysis, interpretation, writing, and revision. Liu Yongzhi: Supervision, conceptual guidance, methodology oversight, validation, and review and editing.

References

Adilazuarda, M. F., Mukherjee, S., Lavania, P., Singh, S., Aji, A. F., O'Neill, J., Modi, A., & Choudhury, M. (2024). Towards measuring and modeling "culture" in LLMs: A survey. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 15763–15784). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.882

AlKhamissi, B., ElNokrashy, M., Alkhamissi, M., & Diab, M. (2024). Investigating cultural alignment of large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 12404–12422). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.671

Balouch, A., Barich, S. N., & Khaskheli, I. H. (2025). Intercultural communication competence: A comprehensive review of Oxford English for Undergraduates. International Journal of Academic Research for Humanities, 5(1), 33–42. https://jar.bwo-researches.com/index.php/jarh/article/view/545

Bearman, M., Ryan, J., & Ajjawi, R. (2023). Discourses of artificial intelligence in higher education: A critical literature review. Higher Education, 86(2), 369–385. https://doi.org/10.1007/s10734-022-00937-2

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Byram, M. (1997). Teaching and assessing intercultural communicative competence. Multilingual Matters.

Cortazzi, M., & Jin, L. (1999). Cultural mirrors: Materials and methods in the EFL classroom. In E. Hinkel (Ed.), Culture in second language teaching and learning (pp. 196–219). Cambridge University Press.

Godwin-Jones, R. (2024). Distributed agency in second language learning and teaching through generative AI. Language Learning & Technology, 28(2), 5–30. https://doi.org/10.64152/10125/73570

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative Health Research, 15(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Jeon, J., Lee, S., & Coronel-Molina, S. M. (2024). Rethinking AI: Bias in speech-recognition chatbots for ELT. ELT Journal, 78(4), 435–445. https://doi.org/10.1093/elt/ccae035

Khosravi, H., Buckingham Shum, S., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. RELC Journal, 54(2), 537–550. https://doi.org/10.1177/00336882231162868

Lashari, N. F., Guo, F., & Maganhar, A. A. (2026). Artificial intelligence and comprehensible input in second language acquisition: Evaluating AI language learning tools through Krashen's Input Hypothesis. International Journal of Academic Research for Humanities, 6(3), 13–25. https://doi.org/10.67244/jar.bwo-researches.v6i3.a609

Lee, S., Jeon, J., McKinley, J., & Rose, H. (2025). Generative AI and English language teaching: A Global Englishes perspective. Annual Review of Applied Linguistics, 45, 85–108. https://doi.org/10.1017/S0267190525100184

Lee, S., & Song, K.-S. (2024). Teachers' and students' perceptions of AI-generated concept explanations: Implications for integrating generative AI in computer science education. Computers and Education: Artificial Intelligence, 7, 100283. https://doi.org/10.1016/j.caeai.2024.100283

Li, C., Chen, M., Wang, J., Sitaram, S., & Xie, X. (2024). CultureLLM: Incorporating cultural differences into large language models. Advances in Neural Information Processing Systems, 37, 84799–84838. https://doi.org/10.52202/079017-2693

Lo, A. W. T. (2025). The educational affordances and challenges of generative AI in Global Englishes-oriented materials development and implementation: A critical ecological perspective. System, 130, 103610. https://doi.org/10.1016/j.system.2025.103610

Ma, Q., Crosthwaite, P., Sun, D., & Zou, D. (2024). Exploring ChatGPT literacy in language education: A global perspective and comprehensive approach. Computers and Education: Artificial Intelligence, 7, 100278. https://doi.org/10.1016/j.caeai.2024.100278

Minhas, H., & Salim, M. (2025). Constructing nature through artificial intelligence: An eco-semantic discourse study of Gemini and DeepSeek. International Journal of Academic Research for Humanities, 5(4), 100–113. https://jar.bwo-researches.com/index.php/jarh/article/view/586

Moorhouse, B. L. (2024). Generative artificial intelligence and ELT. ELT Journal, 78(4), 378–392. https://doi.org/10.1093/elt/ccae032

Moorhouse, B. L., & Wong, K. M. (2025). Generative

artificial intelligence and language teaching. Cambridge University Press. https://doi.org/10.1017/9781009618823

Myung, J., Lee, N., Zhou, Y., Jin, J., Putri, R. A., Antypas, D., Borkakoty, H., Kim, E., Perez-Almendros, C., Ayele, A. A., Gutierrez-Basulto, V., Ibanez-Garcia, Y., Lee, H., Muhammad, S. H., Park, K., Rzayev, A. S., White, N., Yimam, S. M., Pilehvar, M. T., ... Oh, A. (2024). BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. Advances in Neural Information Processing Systems, 37, 78104–78146. https://doi.org/10.52202/079017-2483

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) (pp. 16366–16393). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.862

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) (pp. 6423–6443). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.326

OECD. (2026). OECD digital education outlook 2026: Exploring effective uses of generative AI in education. OECD Publishing. https://doi.org/10.1787/062a7394-en

Pan, Z., & Wang, Y. (2025). From technology-challenged teachers to empowered digitalized citizens: Exploring the profiles and antecedents of teacher AI literacy in the Chinese EFL context. European Journal of Education, 60(1), e70020. https://doi.org/10.1111/ejed.70020

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. The Journal of Strategic Information Systems, 34(2), 101885. https://doi.org/10.1016/j.jsis.2024.101885

Rao, A. S., Yerukola, A., Shah, V., Reinecke, K., & Sap, M. (2025). NormAd: A framework for measuring the cultural adaptability of large language models. 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) (pp. 2373–2403). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.120

Rose, H., McKinley, J., & Galloway, N. (2021). Global Englishes and language teaching: A review of pedagogical research. Language Teaching, 54(2), 157–189. https://doi.org/10.1017/S0261444820000518

Shah, M. H. A., Ali, Z., & Shah, A. (2024). Challenging factors towards the effective use of ChatGPT in education in Province Sindh, Pakistan: Application of TAM model. International Journal of Academic Research for Humanities, 4(2), 86–94. https://jar.bwo-researches.com/index.php/jarh/article/view/450

Shahid, M. M. (2025). AI matters: Debunking the myth that AI hinders the creative process. International Journal of Academic Research for Humanities, 5(2), 18–30. https://jar.bwo-researches.com/index.php/jarh/article/view/550

Shi, W., Li, R., Zhang, Y., Ziems, C., Yu, S., Horesh, R., De Paula, R. A., & Yang, D. (2024). CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 4996–5025). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.288

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. PNAS Nexus, 3(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

UNESCO. (2023). Guidance for generative AI in education and research. https://unesdoc.unesco.org/ark:/48223/pf0000386693

UNESCO. (2024). AI competency framework for teachers. https://unesdoc.unesco.org/ark:/48223/pf0000391104

Varsik, S., & Vosberg, L. (2024). The potential impact of artificial intelligence on equity and inclusion in education (OECD Artificial Intelligence Papers No. 23). OECD Publishing. https://doi.org/10.1787/15df715b-en

Wang, W., Jiao, W., Huang, J., Dai, R., Huang, J.-T., 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) (pp. 6349–6384). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.345

Zhang, H., Li, R., Chen, X., & Yan, F. (2024). Cultural representation in foreign language textbooks: A scoping review from 2012 to 2022. Linguistics and Education, 83, 101331. https://doi.org/10.1016/j.linged.2024.101331

Zhou, N., Bamman, D., & Bleaman, I. L. (2025). Culture is not trivia: Sociocultural theory for cultural NLP. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 25869–25886). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1256

Annexure A

Standard Prompt Structure

Each prompt requested an intermediate-level EFL passage of 220–250 words on one topic. It required fair representation of Chinese, Pakistani, and mainstream English-speaking Western perspectives, avoidance of stereotypes and unsupported generalizations, a short title, and three comprehension questions.

Standard form: Create a 220–250-word intermediate-level EFL reading passage about [TOPIC]. Represent Chinese, Pakistani, and mainstream English-speaking Western cultural perspectives fairly and respectfully. Avoid stereotypes, fixed claims, and unsupported generalizations. Use clear vocabulary suitable for university EFL learners. Give the passage a short title and add three comprehension questions. Do not explain your writing process and do not mention being an AI.

Annexure B

Data-Collection Record

1. Date: 22 July 2026
2. ChatGPT interface: GPT-5.6, paid web account
3. DeepSeek interface: DeepSeek-V4, free web account
4. Fresh conversation used for each prompt
5. Only the first response retained
6. Dataset: 24 passages, 12 from each system
7. Original wording and comprehension questions preserved

Annexure C

Comparative Coding Summary

Theme	ChatGPT Pattern	DeepSeek Pattern
Cultural inclusion	All perspectives appeared in every passage.	All perspectives appeared in every passage.
Intragroup variation	Frequent qualification and explicit internal diversity.	Present in selected passages, less consistently.
Cultural specificity	Moderate detail combined with cautions.	Rich named detail and concrete practices.
Essentialization risk	Lower because claims were often hedged.	Higher because several claims were broad or absolute.
Collective-individual binary	Present but often challenged by hybrid examples.	More stable contrast across cultural blocks.
Intercultural message	Clarification, boundaries, and avoidance of quick judgment.	Positive resolution often followed categorical descriptions.