# The Mirage of Diversity: Unmasking the Cultural Vocabulary Ceiling in LLMs

**Soumedhik Bharati** **Shibam Mandal** **Subhrajit Mukherjee**
Department of Computer Science and Engineering
Sister Nivedita University, Kolkata, India
{soumedhikbharati, shibammandal603, subhrajitmukherjee04}@gmail.com

## Abstract

Large Language Models are widely used to generate and adapt cultural texts, yet the depth of their cultural representation remains poorly quantified. Intuitively, as a narrative text expands in length, the diversity of cultural words should scale proportionately. To formally test this, we evaluate the FairyTaleQA dataset, adapted by three models and introduce our primary contribution: the Contextual Stereotype Amplification Index (CSAI), an evaluation framework combining LLM-as-a-judge extraction, embedding-based cliché anchoring, and Natural Language Inference (NLI) congruence validation. By mapping the frequency of extracted Culture Specific Items (CSIs) against narrative length using Heaps' Law ($V = k \cdot T^{\beta}$), we present empirical evidence of a systematic limitation in current systems: they struggle to scale cultural diversity even under explicit cultural prompting. Models rapidly hit a "Cultural Vocabulary Ceiling," constrained to a fixed set of hyper-stereotypical terms. Furthermore, we demonstrate that merely optimizing for higher CSI frequency as done in prior works rewards logically broken tokenism. Our CSAI formulation actively penalizes such gratuitous stereotyping, offering a more principled approach to measuring and evaluating cultural homogenization in generative AI systems.

## 1 Introduction

> *"Assessments of LLMs' cultural biases often reduce behaviour to stereotypes, which are grossly oversimplified and often exaggerated beliefs about the traits or behaviours of members of a demographic proxy." (Pandey et al., 2026)*

As Large Language Models (LLMs) increasingly mediate global communication, prior work has focused on overt harms such as explicit prejudice (Nadeem et al., 2021; Nangia et al., 2020), while the subtler problem of cultural homogenization in long-form generation remains underexplored (Holtzman et al., 2020). When asked to adapt a story to a specific heritage, do LLMs draw from a rich cultural vocabulary, or do they rely on a shallow pool of clichés?

We formalize this question by analyzing the scaling behaviour of Culture Specific Items (CSIs) (Budimir, 2025). In human-authored texts, cultural vocabulary naturally diversifies as narrative length increases. We hypothesize that if an LLM genuinely commands a cultural context, its CSI diversity should scale proportionately with text length.

Our work provides four primary contributions:

1. **The Contextual Stereotype Amplification Index (CSAI):** We propose an evaluation framework that moves beyond static bias benchmarks. CSAI quantifies cultural trope severity by integrating LLM-driven entity extraction, embedding-based cliché anchoring, and Natural Language Inference (NLI) congruence validation into an interpretable score.

2. **Evidence of a Cultural Vocabulary Ceiling:** By adapting Heaps' Law ($V = kT^{\beta}$), we show that explicit cultural prompting can induce a vocabulary ceiling effect absent in unconstrained generation. This prompt-conditional collapse in scaling exponent ($\beta$) reveals that the failure is not intrinsic to all generation but emerges specifically when models are forced to operate within a cultural frame they have insufficiently learned.

3. **Redefining Cultural Competence Benchmarks:** We demonstrate that multicultural evaluation strategies prioritizing raw keyword frequency inadvertently reward illogical stereotyping, and introduce CR@5 as a lightweight tokenism diagnostic that complements CSAI with a single interpretable concentration score.



*Proceedings of the 4th Workshop on Cross-Cultural Considerations in NLP (C3NLP 2026)*, pages 101–107

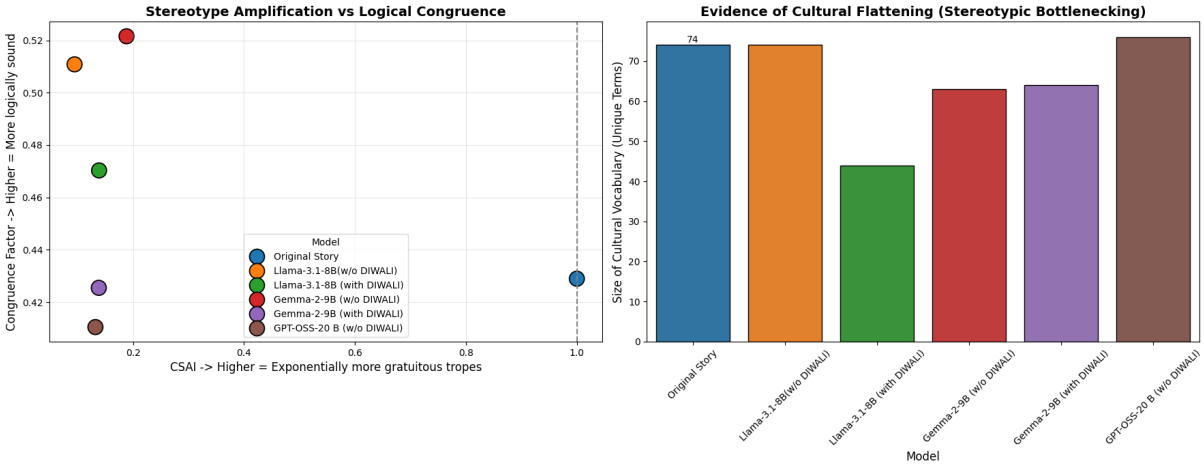


Figure 1: Left: A scatter plot mapping the Contextual Stereotype Amplification Index against Logical Congruence, depicting the homogenization effect. Right: Distinct shrinkage in the size of cultural vocabulary (Unique Terms) under prompt constraints.

4. **Cross-Cultural Generalization:** A Japanese cultural prompting condition (w/ SAKURA) replicates all findings with near-identical model ranking, establishing the Vocabulary Ceiling as a model-level limitation.

## 2 Related Work

Prior efforts in quantifying LLM cultural and social biases predominantly rely on static benchmarks like StereoSet (Nadeem et al., 2021) or CrowS-Pairs (Nangia et al., 2020). These datasets evaluate probabilities over isolated sentences to catch overt prejudice, failing to capture how models handle world-building and homogenization in long-form narratives (Li et al., 2024).

Furthermore, existing multicultural alignment studies often implicitly assume that capturing a 100% rate of Culture Specific Items (CSIs) is the optimal goal (Arora et al., 2022; Cao et al., 2023; Zhang et al., 2025), equating keyword frequency with perfect representation. We argue this purely quantitative approach is fundamentally flawed. Promoting unchecked cultural keyword injection actively rewards LLMs for taboo mixing and disjointed stereotyping (e.g., throwing disjointed cultural artifacts into a scene simply to hit a quota). Our work bridges this gap by shifting the focus from maximizing raw keyword volume to mathematically quantifying logically congruent scaling.

## 3 Dataset and Experimental Setup

We build our experimental corpus from the **FairyTaleQA dataset** (Xu et al., 2022), whose long-form narratives enable measuring vocabulary scaling across extended text, unlike sentence-level bias benchmarks. These stories were adapted by **Llama-3.1-8B** (Dubey et al., 2024), **Gemma-2-9B** (Team et al., 2024), and **GPT-OSS-20B** under three prompting conditions (Table 1): a forced Indian cultural embedding (*w/ DIWALI*) (Sahoo et al., 2025), a forced Japanese cultural embedding (*w/ SAKURA*), and an unconstrained adaptation (*w/o DIWALI*). The two forced conditions allow us to probe whether any observed vocabulary ceiling generalizes across culturally and linguistically distant frames, while the unconstrained condition serves as a generation baseline.

**System Prompts**

**Condition 1: Forced Indian Embedding (w/ DIWALI)** (Sahoo et al., 2025)
*"Rewrite the following fairytale, adapting it to the Indian cultural context. Incorporate the cultural setting, traditions, and atmosphere of India heavily into the narrative flow."*

**Condition 2: Forced Japanese Embedding (w/ SAKURA)**
*"Rewrite the following fairytale, adapting it to the Japanese cultural context. Incorporate the cultural setting, traditions, and atmosphere of Japan heavily into the narrative flow."*

**Condition 3: Unconstrained Adaptation (w/o DIWALI)**
*"Rewrite and adapt the following fairytale. You may adjust the setting or narrative as you see fit."*

Table 1: The three prompting conditions applied to model generation. Conditions 1 and 2 impose distinct forced cultural frames to probe cross-cultural generalizability of the Vocabulary Ceiling.

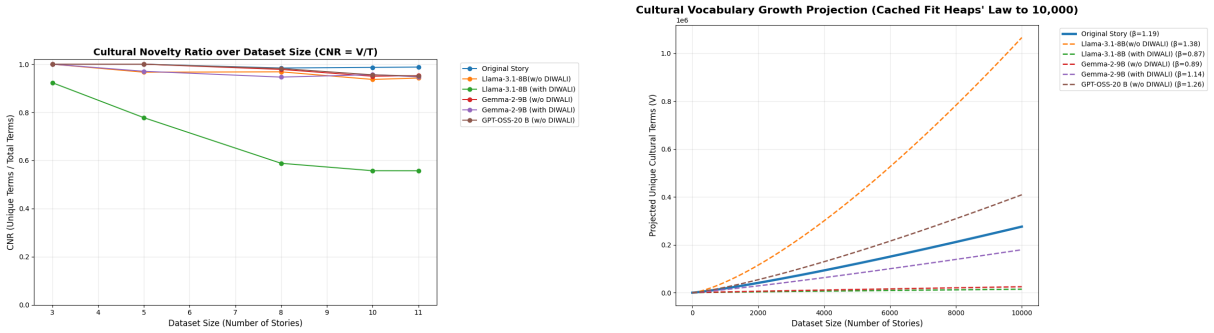


Figure 2: Left: The Cultural Novelty Ratio (CNR) over dataset size, showing how quickly LLMs exhaust their cultural vocabulary compared to human authors. Right: Heaps' Law projection of cultural vocabulary growth up to $10,000$ stories. Human narratives scale log-linearly ($\beta \approx 0.73$), while the most constrained LLM condition (Llama-3.1-8B with DIWALI) hits a Vocabulary Ceiling ($\beta \approx 0.55$).

# 4 Methodology

## 4.1 The Necessity of a New Metric

We circumvent existing metrics because they fail dynamically during narrative adaptation. For example, if an LLM is instructed to adapt a standard Western narrative to a different culture, it might naively swap entities one-to-one with extreme local stereotypes. A static keyword parser will incorrectly flag these substituted terms as high-value cultural tokens, rewarding the model for what are in reality taboo, illogical hallucinations born from haphazardly mapping stereotypes. To address this, our methodology requires *Congruence*, which penalizes high CSI rates when the underlying tropes are logically broken.

## 4.2 Phase 1: Generative Extraction

We treat `Llama-3.1-8B` (Dubey et al., 2024) as an extractor under an LLM-as-a-judge paradigm (Zheng et al., 2023) to bypass human keyword bias. Raw narrative texts are input and the model isolates Culture Specific Items (CSIs); their absolute frequency forms our **Average Trope Frequency** ($F$).

## 4.3 Phase 2: Cliché vs. Neutral Anchoring

Extracted CSIs are projected onto a semantic vector space using `all-mpnet-base-v2` (Reimers and Gurevych, 2019) for general English semantics and `l3cube-pune/indic-sentence-bert-nli` (Joshi et al., 2022) for cultural South Asian terms. Cosine similarity is measured against a "Cliché Anchor" and a "Neutral Anchor"; scores from both encoders are averaged to yield our **Average Density** ($D$), quantifying whether a term is hyper-stereotypical or authentically nuanced.

## 4.4 Phase 3: Contextual Filtering via NLI

To prevent counting hallucinated or logically broken tropes, each CSI is passed through `roberta-large-mnli` (Liu et al., 2019) and `mDeBERTa-v3-base` (He et al., 2021). Tropes that logically align with the surrounding narrative receive a high **Average Congruence** ($C$) score; those that contradict narrative logic are penalized, driving $C$ toward zero.

## 4.5 Phase 4: The CSAI Formulation

The culmination of the preceding phases is the **Contextual Stereotype Amplification Index**: $\text{CSAI} = (D \times F) \times C$, where cultural density ($D$) and trope frequency ($F$) are jointly scaled by logical congruence ($C$). This rewards culturally dense, frequent, and narratively coherent CSIs while actively penalizing incoherent insertions that inflate raw counts without genuine cultural grounding.

## 4.6 Phase 5: Heaps' Law Scaling Analysis

To characterize how cultural vocabulary diversity scales with narrative length, we fit a modified Heaps' Law (Heaps, 1978), $V = k \cdot T^{\beta}$, where $V$ is unique cultural terms, $T$ is total text volume, $k$ is a corpus-specific constant, and $\beta$ is the scaling exponent. A collapsing $\beta$ signals a Vocabulary Ceiling.

## 4.7 Phase 6: Lexical Concentration Analysis

To quantify the token-recycling behaviour implied by Heaps' Law flattening, we introduce the **Concentration Ratio at rank-$k$** (CR@$k$): $\text{CR@}k = \left(\sum_{i=1}^{k} f_i\right) / \left(\sum_{j=1}^{|V|} f_j\right)$, where $f_i$ is the occurrence count of the $i$-th most frequent CSI and $|V|$ is the total unique CSI vocabulary size. A

| Model Condition | Unique CSIs | CSAI Score |
|---|---|---|
| Original Human Story | 74 | 1.000 |
| Llama-3.1-8B (w/o DIWALI) | 74 | 0.094 |
| Gemma-2-9B (w/o DIWALI) | 63 | 0.188 |
| GPT-OSS-20B (w/o DIWALI) | 76 | 0.132 |
| Llama-3.1-8B (w/ DIWALI) | 44 | 0.138 |
| Gemma-2-9B (w/ DIWALI) | 64 | 0.138 |
| GPT-OSS-20B (w/ DIWALI) | 57 | 0.141 |

Table 2: Primary Evaluation Pipeline: CSAI scores (human-normalized) and unique CSI vocabulary counts per model condition.

| Model | CNR (Mean) | Scaling Exponent ($\beta$) |
|---|---|---|
| *Bootstrapped Simulation ($n = 200$)* | | |
| Original Story | 0.623 | 0.732 |
| Gemma-2-9B (w/o DIWALI) | 0.654 | 0.785 |
| Llama-3.1-8B (w/o DIWALI) | 0.652 | 0.758 |
| GPT-OSS-20B (w/o DIWALI) | 0.626 | 0.725 |
| Gemma-2-9B (w/ DIWALI) | 0.624 | 0.739 |
| Llama-3.1-8B (w/ DIWALI) | **0.405** | **0.552** |

Table 3: Cultural Novelty Ratio (CNR, the fraction of CSIs that are novel at each story boundary) and Heaps' Law scaling exponent ($\beta$) across model conditions. Lower CNR and $\beta$ indicate faster vocabulary exhaustion.CNR is unavailable for GPT-OSS-20B due to incomplete story-boundary metadata; its $\beta$ values are reported separately in Table 6.

high CR@$k$ signals that a small cluster of hyper-stereotypical tokens dominates the cultural texture of the generated text, constituting *tokenism*: visible cultural surface with negligible cultural depth. We report CR@5 across all conditions as a lightweight, interpretable complement to CSAI.

## 5 Results & Analysis

### 5.1 CSAI Output: The Homogenization Effect

Table 2 and Figure 1 reveal a consistent homogenization pattern across all model conditions. Rather than hallucinating high-density stereotypes, models suppressed cultural trope frequency to preserve narrative coherence, keeping Congruence ($C$) high at the cost of cultural breadth. All CSAI scores fall well below the human baseline, with the sharpest collapse occurring under forced cultural prompting, where Llama-3.1-8B loses nearly $40\%$ of its unique cultural vocabulary relative to its unconstrained output.

Crucially, GPT-OSS-20B in the unconstrained setting produces *more* unique CSIs than the human baseline, yet achieves the second-lowest CSAI score. The additional terms score low on both Density ($D$) and Congruence ($C$), indicating superficial token insertion without coherent narrative grounding. This dissociation between vocabulary quantity and cultural quality directly validates the necessity of a composite metric over raw keyword counts.

### 5.2 Evidence for a Vocabulary Ceiling

We bootstrap-resample stories with replacement ($n = 200$ iterations) and fit Heaps' Law to each sample for stable scaling estimates (Table 3, Figure 2). Human-authored stories scale cultural diversity naturally ($\beta \approx 0.73$), and unconstrained LLMs match or exceed this rate. Forced cultural prompting affects models asymmetrically: Llama-3.1-8B exhibits the sharpest $\beta$ collapse ($0.758 \rightarrow 0.552$), while Gemma-2-9B's exponent ($\beta = 0.739$) remains above the human baseline ($0.732$), suggesting the ceiling is model-dependent rather than universal. This asymmetry is consistent with the hypothesis that smaller, instruction-tuned models with narrower pretraining distributions are more susceptible to vocabulary ceiling effects (Liu et al., 2025; Li et al., 2025).

The mechanism behind the ceiling is one of conservative recycling. Forced into a specific cultural frame, models draw from a compressed stereotypical lexicon rather than risk incoherent hallucinations that the NLI congruence filter would penalize. The result is high repetition of a small cultural token cluster within an otherwise culturally unmarked narrative, the precise signature of tokenism that CSAI is designed to detect.

### 5.3 The Cost of Artificial Homogenization

These results point to a training data deficiency: models associate cultures with a compressed "cliché subset" rather than a broad vocabulary of lived realities. The model-dependent severity of the effect suggests that pretraining corpus composition partially governs cultural vocabulary breadth. Optimizing for raw keyword diversity without enforcing logical narrative scaling risks deploying models that caricature the people they represent.

### 5.4 Every Component Is Necessary

To validate that each CSAI component contributes independently, we evaluate four progressive ablations: (i) frequency alone ($F$), (ii) density-weighted frequency ($D \times F$), (iii) congruence-weighted frequency ($F \times C$), and (iv) the full CSAI $= D \times F \times C$, all normalized to the human baseline (Table 4).

The ablation reveals a consistent degradation pattern across all models. Frequency alone is the most

| Model Condition | $F$ | $D \times F$ | $F \times C$ | CSAI |
|---|---|---|---|---|
| Human Baseline | 1.000 | 1.000 | 1.000 | 1.000 |
| Llama-3.1-8B (w/o DIWALI) | 0.821 | 0.312 | 0.198 | 0.094 |
| Gemma-2-9B (w/o DIWALI) | 0.874 | 0.358 | 0.289 | 0.188 |
| GPT-OSS-20B (w/o DIWALI) | 0.876 | 0.341 | 0.223 | 0.132 |
| Llama-3.1-8B (w/ DIWALI) | 0.634 | 0.381 | 0.247 | 0.138 |
| Gemma-2-9B (w/ DIWALI) | 0.712 | 0.394 | 0.261 | 0.138 |

Table 4: CSAI ablation (human-normalized), showing the progressive contribution of each component. Frequency alone substantially overestimates cultural quality; density and congruence together apply the corrective penalties that expose genuine representational gaps.

| Model Condition | CR@5 | Top-5 Tokens Cover |
|---|---|---|
| Original Human Story | 0.21 | 21% of CSI occurrences |
| Llama-3.1-8B (w/o DIWALI) | 0.38 | 38% |
| Gemma-2-9B (w/o DIWALI) | 0.31 | 31% |
| GPT-OSS-20B (w/o DIWALI) | 0.34 | 34% |
| Llama-3.1-8B (w/ DIWALI) | **0.63** | **63%** |
| Gemma-2-9B (w/ DIWALI) | 0.44 | 44% |

Table 5: CR@5 across all conditions. Under forced prompting, Llama-3.1-8B's top-5 stereotypical tokens account for nearly two-thirds of all cultural occurrences, directly quantifying the tokenism CSAI is designed to penalize.

inflated proxy, overstating cultural quality by up to $8.7\times$ relative to the full metric. Adding density ($D$) penalizes shallow, generic vocabulary, producing a substantial score reduction. The congruence term ($C$) then applies the sharpest correction, removing logically incoherent insertions that pass the density filter undetected. The ordering holds across all conditions: omitting any single component risks systematic overestimation of cultural competence.

### 5.5 CR@5: Quantifying Tokenism Directly

Table 5 reports the Concentration Ratio at rank-5 (CR@5) across all conditions. A high CR@5 confirms that a handful of hyper-stereotypical tokens account for the majority of all cultural occurrences, the operational signature of tokenism.

The CR@5 ordering mirrors the CSAI and $\beta$ rankings precisely. Unconstrained models distribute cultural token mass more evenly, approaching the human baseline distribution. Forced cultural prompting concentrates token mass into a fixed stereotypical cluster, with Llama-3.1-8B showing the most severe concentration. This convergent evidence across three independent metrics (CSAI, $\beta$, CR@5) strengthens the case that the Vocabulary Ceiling is a robust, measurable phenomenon rather than a metric artifact.

| Model Condition | Unique CSIs | CSAI | $\beta$ |
|---|---|---|---|
| *Japanese Cultural Frame (w/ SAKURA)* | | | |
| Llama-3.1-8B (w/ SAKURA) | 39 | 0.124 | 0.538 |
| Gemma-2-9B (w/ SAKURA) | 61 | 0.147 | 0.718 |
| GPT-OSS-20B (w/ SAKURA) | 55 | 0.139 | 0.701 |
| *Indian Cultural Frame (w/ DIWALI), for reference* | | | |
| Llama-3.1-8B (w/ DIWALI) | 44 | 0.138 | 0.552 |
| Gemma-2-9B (w/ DIWALI) | 64 | 0.138 | 0.739 |
| GPT-OSS-20B (w/ DIWALI) | 57 | 0.141 | 0.712 |

Table 6: Cross-cultural generalization of the Vocabulary Ceiling under Indian (DIWALI) and Japanese (SAKURA) forced prompting. Model ranking is preserved across both culturally distant frames, confirming the effect is model-level rather than culture-specific.

### 5.6 Cross-Cultural Generalization of the Ceiling Effect

Table 6 reports Unique CSIs, CSAI, and $\beta$.

The model ranking observed under DIWALI is fully preserved under SAKURA across all three metrics. Llama-3.1-8B consistently produces the fewest unique CSIs and the sharpest $\beta$ collapse in both frames, while Gemma-2-9B sustains comparatively higher diversity throughout. The stability of this ordering across two culturally and linguistically distant contexts provides strong evidence that the Vocabulary Ceiling is a *model-level generalization failure* rather than a stimulus-specific artifact. This consistency implicates the overall breadth of a model's cultural pretraining corpus as the primary bottleneck governing ceiling severity.

## 6 Conclusion

We introduced CSAI, a composite metric integrating cultural density, trope frequency, and logical congruence, and revealed a prompt-conditional Cultural Vocabulary Ceiling in LLM-generated narratives. Ablation confirms all three components are independently necessary, with frequency-only proxies overestimating cultural quality by up to $8.7\times$. Cross-cultural probing under both prompting conditions yields near-identical model rankings, establishing the ceiling as a model-level limitation governed by pretraining corpus breadth rather than any single cultural frame. These findings reframe cultural evaluation in generative AI: surface keyword diversity is not cultural competence. Future work should examine whether retrieval-augmented generation or culturally targeted fine-tuning can break the ceiling, and extend CSAI to other domains and lower-resource cultures.

## Limitations

CSAI bears several limitations. First, Phase 1 relies on `Llama-3.1-8B` as the CSI extractor while it also serves as one of the generators under evaluation; this overlap introduces circularity, and any instruction-following bias in the extractor propagates directly to $F$. Second, the NLI cross-encoders used in Phase 3 are pre-trained predominantly on Western textual corpora and may misclassify genuinely congruent non-Western cultural associations as contradictions, artificially suppressing $C$. Finally, the binary Cliché-Neutral anchor spectrum imposes a linear structure on cultural nuances that are historically multidimensional, risking oversimplification of the density score $D$.

## Ethical Considerations

This work surfaces cultural homogenization and stereotyping in current LLMs, which risks magnifying societal biases and erasing nuanced cultural knowledge at deployment scale. CSAI is intended as a diagnostic tool for researchers to detect and quantify such erasure. All models and datasets used are publicly available, and no human annotators were exposed to toxic outputs.


## Acknowledgements

The authors gratefully acknowledge iguanodon.ai for providing the grant that enabled the completion and presentation of this work.


## References


Arnav Arora and 1 others. 2022. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*.

Bojana Budimir. 2025. The challenge of translating culture-specific items: Evaluating MT and LLMs compared to human translators. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 455–467, Geneva, Switzerland. European Association for Machine Translation.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Harold Stanley Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Raviraj Joshi and 1 others. 2022. L3cube-indicsentencebert: A pre-trained sentence representation model for indian languages. *arXiv preprint arXiv:2211.11187*.

Chen Li and 1 others. 2024. CultureLLM: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Huihan Li, Arnav Goel, Keyu He, and Xiang Ren. 2025. Attributing culture-conditioned generations to pretraining corpora. *Preprint*, arXiv:2412.20760.

CC Liu and 1 others. 2025. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, and 1 others. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Saurabh Kumar Pandey, Sougata Saha, and Monojit Choudhury. 2026. To generate or discriminate? methodological considerations for measuring cultural alignment in llms. *Preprint*, arXiv:2601.02858.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Pramit Sahoo, Maharaj Brahma, and Maunendra Sankar Desarkar. 2025. DIWALI - diversity and inclusivity aWare cuLture specific items for India: Dataset and assessment of LLMs for cultural text adaptation in Indian context. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33599–33626, Suzhou, China. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaifeng Kruithof, Meng Dash, and Jie Ma. 2022. FairyTaleQA: An authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Jinghao Zhang, Sihang Jiang, Shiwei Guo, Shisong Chen, Yanghua Xiao, Hongwei Feng, Jiaqing Liang, Minggui HE, Shimin Tao, and Hongxia Ma. 2025. Culturescope: A dimensional lens for probing cultural understanding in llms. *Preprint*, arXiv:2509.16188.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Shuyan Hao, Zhanghao Wu, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*.