

What is “Cultural Representation” in AI Benchmarks?

A Multidimensional Evaluation Framework

Carolyn Holtermann, Josefine Birkmann, Anne Lauscher,
University of Hamburg, Trustworthy AI Lab

Abstract

Cultural representation is a prerequisite for the comprehensive evaluation of cultural competence in AI systems. Yet there is no shared, structured way to assess what *cultural representation* in AI benchmarks involves or how it can be measured. We address this gap by introducing a six-dimensional framework that assesses cultural representation through a structured, dimension-specific scoring instrument, enabling systematic comparison across benchmarks. Applying our framework to 50 benchmarks explicitly targeting culture, we find that none achieves broad coverage across all dimensions. While procedural dimensions such as transparency and bias awareness are comparatively well addressed, benchmarks consistently score lower in linguistic and geographic representation, thematic breadth, and the conceptual grounding of culture. Strikingly, only 32% of benchmarks explicitly define culture, and annotator demographic information is absent in 82% of cases. By making these properties measurable and directly comparable, our framework provides a practical foundation for designing and evaluating benchmarks that more validly assess cultural competence in AI systems.

1 Introduction

Evaluating the cultural competence of AI systems is becoming increasingly important as these systems are deployed across diverse social and cultural contexts. Cultural differences shape values, communication styles, and social norms, all of which influence whether AI outputs are perceived as appropriate or useful (Pawar et al., 2024). When these differences are not adequately reflected, AI systems risk producing harmful or irrelevant outputs for large portions of their users (Tao et al., 2024). Despite growing awareness of these challenges, current AI systems continue to predominantly reflect Western perspectives (Hershcovich et al., 2022; Masoud et al., 2025; Myung et al., 2025; Shi et al.,

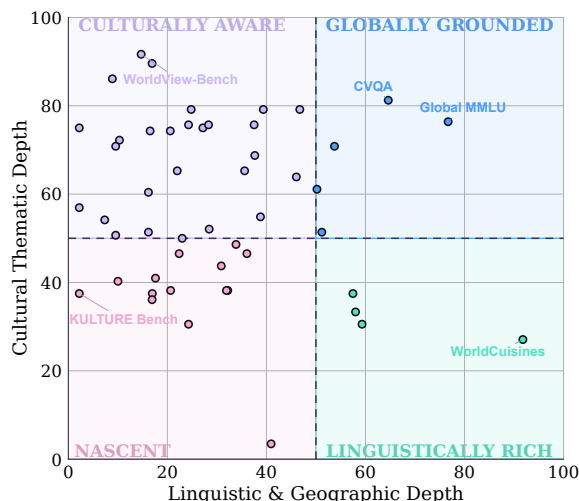


Figure 1: Evaluated benchmarks plotted by D1 (Linguistic & Geographic Representation, x-axis) and D2 (Cultural Content & Thematic Representation, y-axis), as percentages of maximum score. Dashed 50% midpoints are illustrative groupings, not formal cutoffs, defining quadrants: *Nascent*, *Linguistically Rich* (high D1, low D2), *Culturally Aware* (low D1, high D2), and *Globally Grounded*. Representative benchmarks are labelled.

2024; Romero et al., 2024). For example, Joshi et al. (2020) show that around 90% of NLP research focuses on fewer than ten languages, while large language models consistently score lower in non-Western contexts and often reproduce Western liberal values in moral decision-making tasks (Talat et al., 2022; Romero et al., 2024).

Evaluation benchmarks are a critical site of intervention in this dynamic. They define which competencies AI models must demonstrate and, by extension, which cultural perspectives are considered worth evaluating at all (Eriksson et al., 2025). Flawed or too narrow benchmarks thus not only fail to reflect inequalities, they actively entrench them by shaping model development incentives and determining which cultural gaps are visible on leaderboards. A growing number of benchmarks

now explicitly claim to address cultural representation (Chiu et al., 2025; Myung et al., 2025), yet it often remains unclear which conceptualisations of culture they adopt, which dimensions they cover, and whether their representations are methodologically consistent. Adilazuarda et al. (2024) identify the absence of a shared definition of culture as a root cause of incomparability across the field, while Zhou et al. (2025b) propose sociocultural linguistics as a theoretical foundation for cultural AI evaluation. These contributions establish a solid conceptual basis, but neither translates their theoretical grounding into an operational framework for systematically measuring and comparing cultural representation across benchmarks.

We address this gap by introducing a six-dimensional framework for assessing cultural representation in AI benchmarks, which we apply to 50 benchmarks that explicitly address culture. To our knowledge, this is among the first large-scale, quantitatively scored comparative analyses of this kind. Our results suggest that no benchmark achieves comprehensive coverage, with recurring gaps in linguistic and geographic diversity, as visualized in Figure 1, thematic breadth, and explicit conceptualisation of culture. Our contributions are threefold:

- We define cultural representation for NLP evaluation, synthesizing concepts from sociolinguistics, anthropology, and AI fairness into a coherent, annotation-ready construct.
- We present a six-dimensional scoring framework that enables systematic and transparent comparison across benchmarks.
- Using our framework, we conduct a systematic, human-annotated content analysis of 50 NLP benchmarks, revealing persistent structural gaps in linguistic coverage, geographic representation, and annotator transparency.

2 Related Work

Cultural bias and diversity. A substantial body of work demonstrates that training and evaluation data are predominantly drawn from Western, English-speaking, and high-resource contexts, systematically underrepresenting other cultural perspectives (Herscovich et al., 2022; Tao et al., 2024; Masoud et al., 2025; Adilazuarda et al., 2024).

Diversity audits of existing benchmarks. A parallel strand retrospectively analyses benchmark design with respect to diversity and representation.

Joshi et al. (2020) and Ponti et al. (2020) show that NLP benchmarks are heavily skewed towards few language families and dominant languages. From a normative perspective, (Raji et al., 2021; Paullada et al., 2021; Bowman and Dahl, 2021) argue that benchmark design is never neutral but encodes political and cultural assumptions. While these studies provide important critical impulses, they tend to address individual dimensions such as language, bias, or fairness, instead of treating cultural representation as a multidimensional construct (Santurkar et al., 2023).

Theorising culture in AI evaluation. Survey work by Adilazuarda et al. (2024); Liu et al. (2025a) reveals a deeper problem: there is no shared understanding of how culture should be defined and measured in AI research. Instead, culture is operationalised through a wide variety of heterogeneous proxies, which severely limits the comparability of existing work. Zhou et al. (2025b) address this directly, proposing sociocultural linguistics as a theoretical foundation for cultural AI evaluation. (AlKhamissi et al., 2025) similarly critique existing benchmarks through an anthropological lens, applying a four-part taxonomy of cultural framing (knowledge, preference, dynamics, bias) qualitatively to 20 hand-selected benchmarks.

These contributions diagnose the problem clearly and, in the case of Zhou et al. (2025b) and AlKhamissi et al. (2025), offer rich conceptual and qualitative taxonomies of cultural framing. None, however, provides a reproducible, dimension-specific scoring instrument for measuring and comparing cultural representation systematically. The present work addresses this gap.

3 Framework Construction

We propose a hierarchical framework comprising six dimensions, shown in Figure 2, subdivided into thematic blocks and granular subcategories. Benchmark attributes are mapped onto these dimensions and scored individually. We report and compare the dimensions separately, because they capture different kinds of properties whose relevance depends on a benchmark’s specific purpose and design. This structure enables comprehensive analysis of cultural representation without reducing culture to a simplistic proxy or a single overall number.

We design scoring within each dimension to be transparent and reproducible. For each criterion, we derive its point value directly from its answer

format: we distinguish binary criteria, which record whether a specific practice or disclosure is present, from ordinal criteria, which capture the degree to which a benchmark engages with a given aspect of culture, and from continuous coverage criteria, which measure the share of a defined reference space a benchmark covers, such as the world’s languages or countries. We assign more points to formats that convey more graded or continuous information than to simple binary disclosures, since these formats let a benchmark differentiate more finely between minimal and extensive engagement, not because we consider them more important. We detail the resulting point values for each criterion in Appendix A. To facilitate community adoption, we release all code and data along with an interactive Hugging Face¹ application containing the complete questionnaire² enabling researchers to score and compare benchmarks systematically along each dimension.

D1: Linguistic & Geographic Representation

This dimension evaluates linguistic and geographic representation in AI benchmarks. Although language and geographic location are commonly used to operationalise culture in NLP research, they indicate *where* culture is located instead of *what* it consists of; surface-level coverage alone does not guarantee that a benchmark captures underlying cultural meanings (Hershcovich et al., 2022).

We assess three aspects: typological language diversity, geographic coverage, and the inclusion of underrepresented groups. For typological diversity, we draw on Grambank features³, which capture structural linguistic variation beyond simple language counts and mitigate the overrepresentation of closely related high-resource languages (Ploeger et al., 2024; Joshi et al., 2020). For geographic coverage, we apply the UN M49⁴ classification at the level of continents, regions, and countries, computing the proportion of geographic units each benchmark covers. We further score the inclusion of dialects, regional variants, and writing systems; for transparent and theoretically motivated language and region selection; and for benchmarks that explicitly distinguish between language or country and culture (Adilazuarda et al., 2024). **Maximum score: 18 points.**

¹<https://huggingface.co/>

²<https://github.com/TAI-HAMBURG/cultural-representation-framework>

³<https://grambank.cllid.org>

⁴<https://unstats.un.org/unsd/methodology/m49/>

D2: Cultural Content & Thematic Representation Beyond the structural focus of D1, this dimension evaluates substantive cultural coverage. Culture encompasses values, norms, practices, symbols, narratives, and collective meaning systems (Tylor, 1871; Liu et al., 2025a), yet many benchmarks treat it implicitly or superficially (Zhou et al., 2025b).

We examine cultural domains across three thematic clusters: (1) values, norms, and worldviews; (2) narratives, symbols, and social references; and (3) practices and everyday life. Each cluster comprises several concrete domains (listed in Appendix A), which we score individually on a 0–2 scale to reflect that cultural representation is not a binary concept: 0 if absent (i.e., not documented in the reviewed materials), 1 if addressed implicitly or marginally, and 2 if systematically integrated. These per-domain scores account for the majority of the dimension’s total. The remainder rewards benchmarks for the number of distinct thematic domains covered, for organising these domains transparently, and for grounding thematic choices in explicit theoretical reasoning. **Maximum score: 25 points.**

D3: Data Foundation This dimension evaluates the quality, diversity, and transparency of a benchmark’s data foundation. Choices in data selection, sourcing, and annotation ultimately dictate cultural representation and serve as a primary source of cultural bias (Gebu et al., 2021; Bender et al., 2021). Our approach to this dimension is close to the life-cycle perspective of Reuel et al. (2024), who assess general AI benchmark quality across 24 benchmarks, but focuses on practices most consequential for cultural representation.

We assess task format, data creation process, documentation, quality control, and annotation practices. Open-ended task formats score higher than purely closed ones, as they allow more nuanced cultural expression (Gehrmann et al., 2023; Balepur et al., 2025). We award credit for transparent documentation of data sources, creation methods, filtering criteria, and validation procedures (Sambasivan et al., 2021). We score annotator diversity as well as quality control measures on an ordinal scale reflecting the rigor of review applied, and recruitment transparency as a binary disclosure criterion. Because cultural meaning depends strongly on annotator perspective and interpretation (Geiger et al., 2020; Aroyo and Welty, 2015),

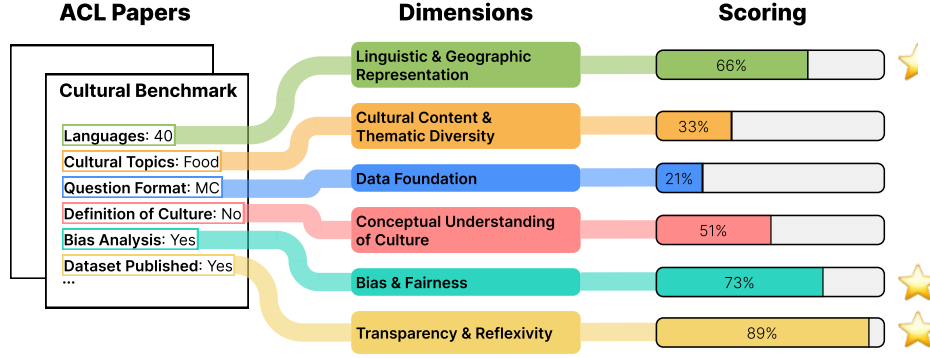


Figure 2: The evaluation framework applied to an example cultural benchmark. Benchmark attributes (left) are mapped onto six dimensions (center), each scored individually (right) relative to its own achievable points; scores are not combined into an overall value.

these criteria collectively capture whether a benchmark’s data foundation reflects a broad and well-documented range of cultural perspectives. **Maximum score: 22 points.**

D4: Conceptual Understanding of Culture To counter the widespread reliance on cultural proxies in AI research (Adilazuarda et al., 2024; Zhou et al., 2025b), this dimension evaluates theoretical engagement with the concept of culture. We do this to avoid privileging any single definition of culture. We look for evidence that authors treat culture as a theoretically informed construct and not as an unexamined assumption, including whether authors ground their treatment of culture in social science theory. **Maximum score: 4 points.**

D5: Bias & Fairness This dimension evaluates critical engagement with bias and fairness in benchmark design. Methodological choices carry implicit assumptions that can marginalise cultural groups and reinforce inequalities (Raji et al., 2021; Nadeem et al., 2021; Abid et al., 2021). We assess whether authors identify potential biases, analyse their data accordingly, report mitigation measures, and reflect on effects for marginalised communities. We award credits for explicit problem awareness and transparent documentation of remaining limitations. **Maximum score: 3 points.**

D6: Transparency & Reflexivity Because cultural representation is inherently selective and contextual (Bender et al., 2021), transparency is crucial for interpretability and reproducibility (Mitchell et al., 2019). This dimension assesses the clear documentation of data sources, selection criteria, and processing decisions, as well as explicit reflection on limitations regarding cultural coverage and

generalisability. **Maximum score: 2 points.**

Instead of a single aggregated score, we report dimensions separately as they capture distinct properties. D1 and D2 assess the cultural content a benchmark represents, D3 the practices that produced it, and D4 through D6 how the benchmark frames and documents its engagement with culture. For each dimension, we sum the points earned and divide by the dimension’s maximum, yielding a percentage score on a common scale. The differing maxima (D1: 18, D2: 25, D3: 22, D4: 4, D5: 3, D6: 2) reflect the number and granularity of the underlying criteria, not their relative importance. Because broad cultural coverage can coexist with weak documentation, and narrow coverage with strong transparency, we do not aggregate dimensions into a single overall score. The framework is therefore intended as a diagnostic instrument, with each dimension interpreted on its own terms. Full scoring details are provided in Appendix A.

4 Data Collection

To evaluate the applicability and comprehensiveness of our framework, we compile a curated sample of 50 benchmarks. We use this collection to demonstrate the practical application of our dimensions and to critically analyze current cultural representation within the field.

4.1 Benchmark Selection

We collect a set of benchmarks explicitly designed to assess cultural representation and related phenomena. To this end, we search the ACL Anthology and Google Scholar for papers containing the keywords cultural alongside benchmark, evaluation, or dataset in their title or abstract.

We restrict this search to publications from 2022 to 2025, reflecting the recent emergence of culturally oriented evaluation for modern large language models and multimodal systems.

From these results, we select a final sample of 50 papers for detailed human annotation (see Table 7 in Appendix B). We curate this sample carefully to ensure broad conceptual coverage across benchmark types, task formats, modalities, and cultural contexts. Consequently, the selected benchmarks span a wide range of evaluation paradigms, including question answering, classification, generation, and multimodal tasks, while covering diverse cultural domains and geographic regions.

4.2 Human Annotation

We evaluate the selected benchmarks using a structured questionnaire directly derived from our proposed framework. The questionnaire contains 58 questions with predefined answer options, complemented by open text fields where annotators document examples, edge cases, and specific justifications. This design operationalizes each framework dimension into concrete criteria, thereby enabling a systematic and reproducible assessment across all benchmarks.

One author manually annotated all 50 benchmarks using the information reported in the original papers and their documentation. During this step, the annotator identifies relevant evidence for each framework dimension and assigns scores according to the predefined criteria described in Section 3 and Appendix A. This ensures a consistent and comparable analysis of how existing benchmarks conceptualize and operationalize culture.

To verify reliability, a second annotator fluent in English and knowledgeable in the AI domain independently evaluates an overlapping set of 20 benchmarks. Because the questionnaire features diverse answer types, we compute annotator agreement separately for each format. We observe a Cohen’s κ of 0.86 for binary criteria, 91.00% agreement for ordinal criteria (0/1/2), and 91.88% agreement for categorical responses. Note that we report raw agreement here instead of a weighted Kappa variant because several sub-categories show skewed score distributions with limited variance, under which Kappa-based statistics can become unstable with only 20 double-annotated benchmarks. Taken together, these results indicate high and consistent reliability of the annotation framework.

Table 1 further reports agreement across the six

Dim.	Subcategories	Decisions	Agreement(%)
D1	10	200	91.88
D2	13	260	90.42
D3	17	340	93.82
D4	4	80	95.00
D5	3	60	95.00
D6	2	40	100.00

Table 1: Inter-annotator agreement across the six framework dimensions (D1–D6) based on the 20 double-annotated benchmarks.

framework dimensions (D1–D6), demonstrating that annotation reliability remains consistently high across all dimensions.

Where annotators diverged, discrepancies were discussed jointly to reach a final, agreed-upon score. This process also led us to refine a small number of ambiguously worded questionnaire items.

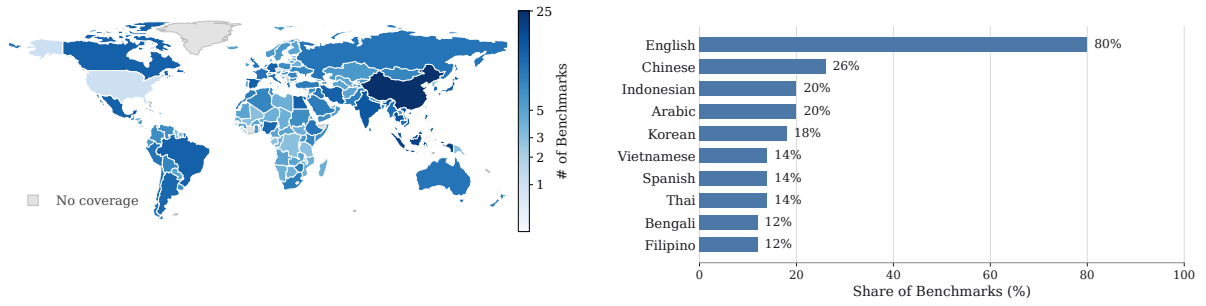
5 Results

5.1 Linguistic and Geographic Representation

Our analysis shows that the evaluated benchmarks provide limited linguistic diversity concentrated on only a small number of languages. Half of the evaluated benchmarks focus exclusively on English. More broadly, English is the most prevalent language overall, appearing in 80% of benchmarks, far ahead of the next most common language, Chinese (Figure 3b). Although language is widely used to operationalise cultural representation in NLP research (Pawar et al., 2024), this operationalisation rarely extends beyond a small set of high-resource languages.

The consideration of dialects is also limited. Only 24% of the benchmarks include dialectal variation, while explicit distinctions between locations are present in only 4%. Cultural differences within linguistic regions are therefore often not represented in a differentiated manner. Instead, culture is reduced to broader categories such as individual languages or countries, leaving regional and everyday linguistic differences largely unaddressed.

Moreover, we find that geographic coverage is equally skewed (Figure 3a). Representation clusters in parts of East and Southeast Asia and in Western industrialised countries, while Africa, Central Asia, and South America receive little or none. This imbalance is not merely a descriptive limitation, as cultural contexts absent from linguistic and geographic coverage are structurally excluded from evaluation. These representational gaps there-



(a) Number of evaluated benchmarks covering each country. Grey indicates no coverage. (b) Share of benchmarks covering the ten most represented languages. English appears in 80% of benchmarks; no other language exceeds 30%.

Figure 3: Geographic and linguistic coverage across evaluated benchmarks, revealing strong concentration in a small number of countries and an overwhelming dominance of English.

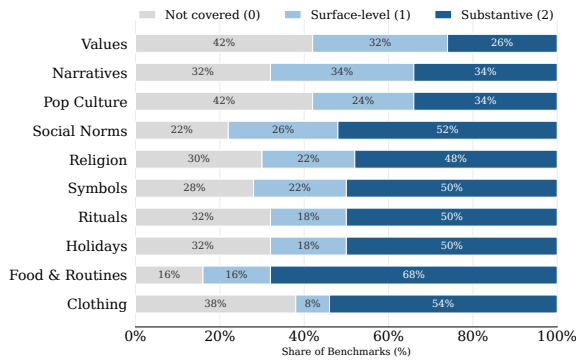


Figure 4: Coverage depth across cultural domains in evaluated benchmarks. Each bar shows the proportion of benchmarks scoring 0 (not covered), 1 (surface-level), or 2 (substantive) for a given domain.

fore constrain the evaluation space, leave certain cultural perspectives structurally invisible, risking the systematic marginalisation of underrepresented communities.

5.2 Cultural Content and Thematic Diversity

Cultural content is broadly present across the evaluated benchmarks, but depth and thematic balance vary considerably. As Figure 4 shows, observable and everyday domains dominate: Food & Routines is substantively covered in 68% of benchmarks, and Social Norms in 52%. We hypothesize that domains such as typical foods, everyday practices, and social conventions can be operationalised as explicit, relatively unambiguous knowledge and translated into closed evaluation formats without requiring interpretive judgment.

In contrast, dimensions characterized by interpretation, contextual dependence, or ambiguity, includ-

ing values and normative beliefs, are systematically underrepresented. Moral and social *values* reach substantive coverage in only 26% of benchmarks, the lowest of all domains, with a further 42% not covering them at all. Narratives and Pop Culture show moderate substantive rates (34% each) but also high rates of complete absence. These are precisely the dimensions that require contextual depth, cultural embeddedness, and interpretive sensitivity, which are capacities that closed and standardised formats are poorly equipped to assess. The result is a structural skew: benchmarks capture culture where it is easiest to measure, not necessarily where it matters most.

Our results further show that only a small number of benchmarks achieve both broad and in-depth coverage across multiple cultural domains. For instance, while 70% of benchmarks cover *Religion* in some form, only 48% do so substantively. In many cases, benchmarks emphasize only individual topics in depth. Thematic breadth measured by presence alone would therefore overstate the depth of cultural engagement across the field.

Figure 1 brings these two dimensions together. Linguistic and geographic representation and thematic breadth are the two properties most commonly used to operationalise cultural diversity in NLP, making this pairing a natural basis for comparison. The joint view surfaces complementary profiles among benchmarks: WorldView-Bench achieves strong thematic coverage with limited linguistic and geographic reach, while WorldCuisines shows the opposite pattern, and Global MMLU is among the few benchmarks combining relatively broad coverage on both dimensions. Most bench-

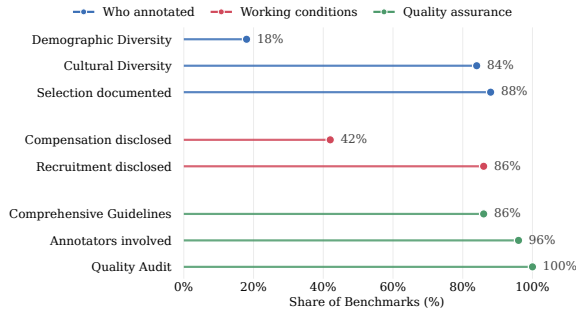


Figure 5: Annotation transparency across evaluated benchmarks grouped by annotator identity (blue), working conditions (red), and quality assurance (green). Quality assurance is consistently reported, whereas annotator demographics (18%) and compensation (42%) are rarely disclosed.

marks, however, remain limited, particularly in linguistic and geographic representation.

Some of this variation can be attributed to deliberate specialization. SEAVQA (Urailetpasert et al., 2024), for example, focuses on specific regional contexts, while FoodieQA (Li et al., 2024b) centers on food culture. Such benchmarks may therefore score lower in overall breadth while offering greater depth or specificity within their intended scope. These scores should consequently be interpreted as measures of cultural diversity coverage, not of overall benchmark quality, which also depends on properties such as transparency and documentation.

5.3 Data Foundation

The results indicate that formal and methodological standards for data creation are established across many benchmarks. The documentation of data creation processes, the use of annotation, and quality assurance measures are systematically addressed in a large proportion of the benchmarks.

However, information about annotators is frequently incomplete. In 82% of the benchmarks, information about the social composition of annotators, including age, gender, or education, is missing. Although many benchmarks report using annotators with appropriate cultural or regional expertise, for example through language proficiency or local affiliation, it often remains unclear how systematically annotators were selected and which dimensions of diversity were considered.

Closed and prestructured task formats dominate the analyzed benchmarks. Multiple-choice tasks account for 54% of the benchmarks, while binary

tasks account for 18%. These formats enable standardized evaluation, comparability, and efficient model ranking. However, they provide only limited opportunities to evaluate whether models can establish connections, provide explanations, or handle ambiguity. Their reliance on fixed answer options and unambiguous solutions restricts the assessment of complex and context-dependent cultural judgments. They may also encourage strategies based on elimination or statistical plausibility without requiring an understanding of the underlying cultural relationships (Clark et al., 2019).

Most of the analyzed benchmarks (74%) use partially automated data creation processes. Automation increases scalability and reduces costs, but it may reproduce dominant knowledge, smooth over cultural nuances, or reinforce stereotypical patterns (Bender et al., 2021). Automated methods can support tasks such as the preliminary structuring of data, while central steps involving content assessment and contextualization require human expertise. Transparent source documentation, review by knowledgeable individuals, and clearly documented decision rules can support the systematic assessment of cultural appropriateness.

Overall, the results reveal a tension between representing cultural complexity in a differentiated manner and enabling standardized and scalable evaluation.

5.4 Conceptualization of Culture

Only 32% of the benchmarks explicitly define what they mean by culture. A theoretical framing is provided in 30% of the benchmarks, while a conceptual framing is provided in 24%. Most of the analyzed benchmarks therefore do not rely on an explicitly developed or theoretically grounded understanding of culture.

Culture is often treated implicitly and operationalized through proxies such as language, nationality, or thematic domains, without clearly articulating the underlying assumptions. This makes it difficult to determine which aspects of culture are captured and how they are conceptualized. This lack of explicit grounding is particularly notable given that all selected benchmarks evaluate culture or related cultural phenomena.

Taken together, these results point to a consistent pattern across dimensions, which we examine in the following section.

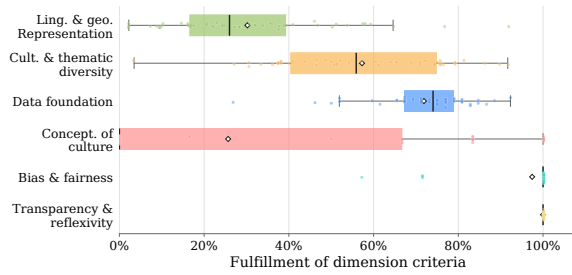


Figure 6: Distribution of dimension-level scores across evaluated benchmarks, expressed as percentage fulfillment of each dimension’s scoring criteria. Each boxplot shows the interquartile range of scores, with the solid vertical line indicating the median and the diamond indicating the mean.

6 Discussion

Our analysis of 50 curated benchmarks indicates a clear imbalance across the evaluated dimensions, as shown in Figure 6. We find that formal quality criteria such as transparency, bias documentation, and data practices are widely met, while the dimensions that require deeper engagement with culture, including linguistic and geographic coverage, thematic breadth, and conceptual grounding, lag behind. Although our sample is not exhaustive, it spans a broad range of benchmark types, task formats, modalities, and cultural contexts, and the findings are sufficiently consistent to identify structural priorities for the field.

Most benchmarks do not explicitly define what they mean by culture, despite this being their central object of evaluation. Without a shared or at least stated definition, it is difficult to determine what a benchmark actually measures, to compare results across benchmarks, or to identify which aspects of culture remain unaddressed.

Observable cultural domains such as food and social norms are well represented, while normative and symbolic dimensions such as values and belief systems receive considerably less attention. This likely reflects the relative ease of operationalising surface-level content in closed evaluation formats and may lead benchmarks to prioritize cultural knowledge that is easy to test while overlooking the dimensions through which culture most directly shapes thought, reasoning, and communication.

The prevalence of automated pipelines and closed response formats reflects real resource constraints in benchmark construction. However, these choices come with trade-offs: automated methods tend to draw on existing digital text, which skews

towards dominant languages and cultural contexts (Bender et al., 2021), and closed formats limit the space for culturally varied responses. The near-absence of annotator demographic reporting further limits the extent to which the cultural representativeness of annotation decisions can be assessed.

From these findings, five priorities emerge for future benchmark development.

First, linguistic and geographic coverage should be expanded and more systematically justified. This includes not only increasing the number of represented languages, but also accounting for variation within linguistic communities, such as dialects, scripts, and regional forms. Such choices should be documented transparently.

Second, benchmarks should provide more balanced thematic coverage by complementing the current emphasis on observable everyday practices with substantive engagement with normative and symbolic dimensions, including values, belief systems, and shared narratives.

Third, benchmark developers should explicitly define what they mean by culture or, at minimum, make their underlying assumptions transparent. A clearer conceptual foundation would improve both the interpretability of individual benchmarks and the comparability of results across the field, and the selection of cultural content should likewise be theoretically informed or explicitly justified.

Fourth, data collection and annotation should incorporate diverse cultural perspectives. The social composition, recruitment, and working conditions of annotators should be considered deliberately and reported transparently, and more participatory annotation processes may help ensure that cultural meanings are not filtered through a narrow set of perspectives.

Fifth, evaluation formats should move beyond predominantly closed and standardised tasks. Open-ended and context-rich approaches can better capture cultural depth, ambiguity, and context dependence, even if they complicate aggregation and model comparison.

7 Conclusion

We introduced a six-dimensional evaluation framework that operationalises cultural representation in AI benchmarks as a set of measurable, dimension-specific criteria, and applied it to 50 benchmarks explicitly targeting culture. To support adoption, we release an interactive scoring application that al-

lows the broader community to assess and compare benchmarks on a shared basis.

We show that cultural representation cannot be reduced to a single score: benchmarks consistently satisfy formal criteria such as transparency and bias reporting while falling short on more substantive dimensions, including linguistic and geographic coverage, thematic breadth, and conceptual grounding. Because translating culture into measurable criteria necessarily involves simplification, our framework should be treated as a structured approximation rather than a comprehensive account of culture. It nonetheless provides a practical basis for assessing existing benchmarks and for guiding more systematic, context-sensitive, and participatory approaches to future benchmark design.

Limitations

This study has several limitations. First, the analysis is restricted to publicly available benchmark documentation. Consequently, we could assess only those design choices and practices explicitly reported in publications and accompanying materials. Undocumented decisions or implicit practices may have been missed, and some benchmarks may account for cultural considerations more extensively than their documentation suggests.

Second, operationalizing cultural representation necessarily reduces complex theoretical concepts to measurable indicators. As a result, cultural dynamics, power relations, and context-dependent meanings can only be captured to a limited extent. The proposed framework should therefore not be interpreted as a comprehensive representation of cultural coverage.

Third, several parameters of the scoring schema, including the number of sub-criteria assigned to each dimension and the point values derived from answer format, were chosen for transparency and reproducibility. We did not validate them against an independent theoretical model or conduct a sensitivity analysis comparing these choices against alternatives, such as equal weighting across all items. Thus, dimensions with fewer criteria, such as Transparency & Reflexivity, may appear near ceiling partly for this reason.

Finally, despite the use of standardized criteria and double annotation, some assessments remain interpretative and may be influenced by subjective judgment. Future work could improve the robustness and reproducibility of the findings by involv-

ing a larger number of independent annotators and applying the framework to a broader and more diverse sample of benchmarks.

References

- Abubakar Abid, Maheen Farooqi, and James Zou. 2021. [Persistent anti-muslim bias in large language models](#). In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.
- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. [Towards measuring and modeling “culture” in LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.
- Mai AlKhamissi, Yunze Xiao, Badr AlKhamissi, and Mona Diab. 2025. [Hire your anthropologist! rethinking culture benchmarks through an anthropological lens](#). *Preprint*, arXiv:2510.05931.
- Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibae, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025. [Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.
- Lora Aroyo and Chris Welty. 2015. [Truth is a lie: Crowd truth and the seven myths of human annotation](#). *AI Mag.*, 36(1):15–24.
- Nishant Balepur, Rachel Rudinger, and Jordan Boyd-Graber. 2025. [Which of these best describes multiple choice evaluation with LLMs? a\) forced B\) flawed C\) fixable D\) all of the above](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

- Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, EunJeong Hwang, and Vered Shwartz. 2024. [From local concepts to universals: Evaluating the multi-cultural understanding of vision-language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782, Miami, Florida, USA. Association for Computational Linguistics.
- Samuel R. Bowman and George Dahl. 2021. [What will it take to fix benchmarking in natural language understanding?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.
- Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhan-syah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, Vicky Feliren, Bahrul Ilmi Nasution, Manuel Antonio Rufino, Genta Indra Winata, Rian Adam Rajagede, Carlos Rafael Catalan, and 73 others. 2025. [Crowd-source, crawl, or generate? creating SEA-VL, a multicultural vision-language dataset for Southeast Asia](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18685–18717, Vienna, Austria. Association for Computational Linguistics.
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. [CulturalBench: A robust, diverse and challenging benchmark for measuring LMs’ cultural knowledge through human-AI red-teaming](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.
- Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. [Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.
- Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. 2025. [Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation](#). *Preprint*, arXiv:2502.06559.
- Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. [Massively multi-cultural knowledge acquisition and lm benchmarking](#). *Preprint*, arXiv:2402.09369.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. [Datasheets for datasets](#). *Commun. ACM*, 64(12):86–92.
- Sebastian Gehrmann, Elizabeth Clark, and Thibault Selam. 2023. [Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text](#). *J. Artif. Int. Res.*, 77.
- R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. [Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?](#) In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* ’20*, page 325–336, New York, NY, USA. Association for Computing Machinery.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. [Challenges and strategies in cross-cultural NLP](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. [Beyond aesthetics: Cultural competence in text-to-image models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 13716–13747. Curran Associates, Inc.
- Amr Keleg and Walid Magdy. 2023. [DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.
- Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. [CLiCK: A benchmark dataset of cultural and linguistic intelligence in Korean](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.
- Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min

- An, James Thorne, and Alice Oh. 2025. [When tom eats kimchi: Evaluating cultural bias of multimodal large language models in cultural mixture contexts](#). *Preprint*, arXiv:2503.16826.
- Kyuhee Kim and Sangah Lee. 2025. [Nunchi-bench: Benchmarking language models on cultural reasoning with a focus on korean superstition](#). *Preprint*, arXiv:2507.04014.
- Huihan Li, Liwei Jiang, Jena D. Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024a. [Culture-gen: Revealing global cultural perception in language models through natural language prompting](#). *Preprint*, arXiv:2404.10199.
- Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Herscovich, and Desmond Elliott. 2024b. [FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA. Association for Computational Linguistics.
- Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. 2023a. [On the cultural gap in text-to-image generation](#). *Preprint*, arXiv:2307.02971.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025a. [Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art](#). *Transactions of the Association for Computational Linguistics*, 13:652–689.
- Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. [Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.
- Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025b. [CultureVLM: Characterizing and improving cultural understanding of vision-language models for over 100 countries](#). *Preprint*, arXiv:2501.01282.
- Zhixuan Liu, Youeun Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. 2023b. [Towards equitable representation in text-to-image synthesis models with the cross-cultural understanding benchmark \(ccub\) dataset](#). *Preprint*, arXiv:2301.12073.
- Ziyi Liu, Priyanka Dey, Jen tse Huang, Zhenyu Zhao, Bowen Jiang, Rahul Gupta, Yang Liu, Yao Du, and Jieyu Zhao. 2025c. [Can llms grasp implicit cultural values? benchmarking llms’ cultural intelligence with cq-bench](#). *Preprint*, arXiv:2504.01127.
- Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. [SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.
- Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka, and Sriparna Saha. 2025. [Sanskriti: A comprehensive benchmark for evaluating language models’ knowledge of indian culture](#). *Preprint*, arXiv:2506.15355.
- Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleven, and Miguel Rodrigues. 2025. [Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. [Model cards for model reporting](#). In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* ’19*, page 220–229, New York, NY, USA. Association for Computing Machinery.
- Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Ward Church, and Mohamed Elhoseiny. 2022. [Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture](#). *Preprint*, arXiv:2211.10780.
- Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Ward Church, and Mohamed Elhoseiny. 2024. [No culture left behind: Artelingo-28, a benchmark of wikiart with captions in 28 languages](#). *Preprint*, arXiv:2411.03769.
- Erfan Moosavi Monazzah, Vahid Rahimzadeh, Yadollah Yaghoobzadeh, Azadeh Shakery, and Mohammad Taher Pilehvar. 2025. [PerCul: A story-driven cultural evaluation of LLMs in Persian](#). 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 12670–12687, Albuquerque, New Mexico. Association for Computational Linguistics.
- Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arif Hasan, Maram Hasanain, Tameem Kabbani, Fahim

- Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024. [Aradice: Benchmarks for dialectal and cultural capabilities in llms](#). *Preprint*, arXiv:2409.11404.
- Abdullah Mushtaq, Imran Taj, Rafay Naeem, Ibrahim Ghaznavi, and Junaid Qadir. 2025. [Worldview-bench: A benchmark for evaluating global cultural perspectives in large language models](#). *Preprint*, arXiv:2505.09595.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2025. [Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages](#). *Preprint*, arXiv:2406.09948.
- 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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.
- Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 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.
- Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. [Benchmarking vision language models for cultural understanding](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.
- Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. 2025. [CROPE: Evaluating in-context adaptation of vision and language models to culture-specific concepts](#). 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 7917–7936, Albuquerque, New Mexico. Association for Computational Linguistics.
- Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. [Data and its \(dis\)contents: A survey of dataset development and use in machine learning research](#). *Patterns*, 2(11):100336.
- Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrana, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. [Survey of cultural awareness in language models: Text and beyond](#). *Preprint*, arXiv:2411.00860.
- Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. [What is “typological diversity” in NLP?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5681–5700, Miami, Florida, USA. Association for Computational Linguistics.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal common-sense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.
- Tushar Pranav, Eshan Pandey, Austria Lyka Diane Bala, Aman Chadha, Indriyati Atmosukarto, and Donny Soh Cheng Lock. 2025. [Rice-vl: Evaluating vision-language models for cultural understanding across asean countries](#). *Preprint*, arXiv:2512.01419.
- Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. [Ai and the everything in the whole wide world benchmark](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 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)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.
- Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. 2024. [Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices](#). *Preprint*, arXiv:2411.12990.
- David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. [Cvqa: Culturally-diverse multilingual visual question answering benchmark](#). *Preprint*, arXiv:2406.05967.
- 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.
- Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA. Association for Computing Machinery.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. *Whose opinions do language models reflect?* Preprint, arXiv:2303.17548.
- Burak Satar, Zhixin Ma, Patrick Amadeus Irawan, Wilfried Ariel Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. 2025. *Seeing culture: A benchmark for visual reasoning and grounding*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22227–22243, Suzhou, China. Association for Computational Linguistics.
- Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025. *GIMMICK: Globally inclusive multimodal multitask cultural knowledge benchmarking*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9605–9668, Vienna, Austria. Association for Computational Linguistics.
- Florian Schneider and Sunayana Sitaram. 2024. *M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks*. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4309–4345, Miami, Florida, USA. Association for Computational Linguistics.
- Weiyang Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rog rio Abreu De Paula, and Diyi Yang. 2024. *CultureBank: An online community-driven knowledge base towards culturally aware language technologies*. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.
- Shivalika Singh, Angelika Romanou, Cl mentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. *Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.
- Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. *SEA-HELM: Southeast Asian holistic evaluation of language models*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12308–12336, Vienna, Austria. Association for Computational Linguistics.
- Zeeraq Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. *On the machine learning of ethical judgments from natural language*. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States. Association for Computational Linguistics.
- Yan Tao, Olga Viberg, Ryan S Baker, and Ren  F Kizilcec. 2024. *Cultural bias and cultural alignment of large language models*. *PNAS Nexus*, 3(9).
- Edward Burnett Tylor. 1871. *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Language, Art, and Custom*, volume 1. John Murray, London. Available online via Darwin Online.
- Norawit Urailetprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. *SEA-VQA: Southeast Asian cultural context dataset for visual question answering*. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Bangkok, Thailand. Association for Computational Linguistics.
- Viacheslav Vasilev, Julia Agafonova, Nikolai Gerasimenko, Alexander Kapitanov, Polina Mikhailova, Evelina Mironova, and Denis Dimitrov. 2025. *Ruscode: Russian cultural code benchmark for text-to-image generation*. Preprint, arXiv:2502.07455.
- Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. *SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.
- Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. 2024b. *KULTURE bench: A benchmark for assessing language model in Korean cultural context*. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 914–927, Tokyo, Japan. Tokyo University of Foreign Studies.
- Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024c. *CDEval: A benchmark for measuring the cultural dimensions of large language models*. In *Proceedings of the*

- 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.
- Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. [WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines](#). 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 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.
- Wilson Wongso, David Samuel Setiawan, Steven Limcorn, and Ananto Joyoadikusumo. 2024. [Nusabert: Teaching indobert to be multilingual and multicultural](#). *Preprint*, arXiv:2403.01817.
- Jincenzi Wu, Jianxun Lian, Dingdong Wang, and Helen M. Meng. 2025. [SocialCC: Interactive evaluation for cultural competence in language agents](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33242–33271, Vienna, Austria. Association for Computational Linguistics.
- Pengju Xu, Yan Wang, Shuyuan Zhang, Xuan Zhou, Xin Li, Yue Yuan, Fengzhao Li, Shunyuan Zhou, Xingyu Wang, Yi Zhang, and Haiying Zhao. 2025. [Tcc-bench: Benchmarking the traditional chinese culture understanding capabilities of mllms](#). *Preprint*, arXiv:2505.11275.
- Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, and 2 others. 2025. [Can MLLMs understand the deep implication behind Chinese images?](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14402, Vienna, Austria. Association for Computational Linguistics.
- Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A. Hedderich. 2025. [MAKIEval: A multilingual automatic WiKidata-based framework for cultural awareness evaluation for LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23104–23136, Suzhou, China. Association for Computational Linguistics.
- Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. [World-ValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*
- Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.
- Li Zhou, Lutong Yu, Dongchu Xie, Shaohuan Cheng, Wenyan Li, and Haizhou Li. 2025a. [Hanfu-bench: A multimodal benchmark on cross-temporal cultural understanding and transcreation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24616–24638, Suzhou, China. Association for Computational Linguistics.
- Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025b. [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)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.

A Scoring Schema per Dimension

Scoring. Each questionnaire item is mapped to a point value, and the maximum points an item can contribute are determined by its answer format rather than by a subjective judgement of its importance. We use three tiers reflecting how finely an item discriminates. Binary (yes/no) items are worth 1 point. Ordinal items graded on a 0/1/2 scale (absent, marginal, systematic) are worth 2 points. Continuous coverage items are worth 3 points, awarded in proportion to the share of a defined reference space they cover (e.g., the full set of UN M49 countries or Grambank features). Count items (annotator diversity) award 1 point per attribute considered, up to a maximum of 3. Because every point value follows from the item’s format, the scheme is fully reproducible from the questionnaire alone. Coverage-based criteria receive more weight than binary disclosure items because they carry more information, not because we deem them more important.

Linguistic and geographic coverage are scored against their true global reference spaces: Grambank typological feature coverage for languages, and the UN M49 classification (6 continents, 22 subregions, 193 countries) for geography. We deliberately do not rescale these to empirically attainable levels. Because no benchmark approaches full global coverage, coverage scores are uniformly low, and this low score is itself a substantive finding about the state of the field rather than an artefact to be normalised away.

Each dimension’s score is the sum of the points its items earn, divided by the maximum obtainable for that dimension, and reported as a percentage. Dimensions are scored and reported independently. They are never combined into a composite or overall score, since each captures a distinct facet of cultural representation whose relevance depends on a benchmark’s purpose. The dimensions’ differing maxima (D1: 18, D2: 25, D3: 22, D4: 4, D5: 3, D6: 2) reflect the number and granularity of their sub-criteria, not their relative importance.

Some dependencies between criteria are inherent to the framework design. In D1, nested geographic units measure coverage at different spatial resolutions, and a small number of criteria capture related aspects across dimensions, such as data source disclosure in D3 and D6. However, since dimensions are scored and interpreted separately, these overlaps do not affect an overall composite score, but

they do mean that dimension scores should not be treated as fully independent.

Table 2 summarizes the distribution of scores across the six framework dimensions for the 50 evaluated benchmarks.

Dim.	Mean	Median	IQR	Range
D1 (18)	30.2	25.2	26.6	3.6–92.5
D2 (25)	59.5	60.4	32.0	4.0–92.0
D3 (22)	70.7	72.7	9.1	36.4–90.9
D4 (4)	25.0	0.0	75.0	0.0–100.0
D5 (3)	97.3	100.0	0.0	66.7–100.0
D6 (2)	100.0	100.0	0.0	100.0–100.0

Table 2: Summary statistics for D1–D6 across the 50 evaluated benchmarks. Scores are percentages of each dimension’s maximum; raw maximum points are shown in parentheses.

The following tables provide the scoring schemas across the dimensions.

Category	Subcategory	Scoring rule	Max. Points
Linguistic Coverage	Language Diversity	$3 \times$ Grambank typological cov	3
	Location Differentiation	Yes $\rightarrow$ 1	1
Geographic Coverage	Continents	$3 \times (\# \text{continents}/6)$	3
	Regions	$3 \times (\# \text{UN M49 subregions}/22)$	3
	Countries	$3 \times (\# \text{countries}/193)$	3
Varieties	Dialects	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
	Writing Systems	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
Repr. of Marginalised Groups	Underrepresented Groups	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
Theoretical Grounding	Theoretical Grounding	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
Culture as multidim. concept	Culture as Concept	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
Total			18

Table 3: Scoring schema for Dimension D1 (Linguistic & Geographic Representation).

Category	Subcategory	Scoring rule	Max. Points
Cultural Narratives & Symbolic Systems	Social Norms & Conventions	0/1/2	2
	Shared Narratives & Coll. Memory	0/1/2	2
	Popular Culture & Media	0/1/2	2
	Symbols & Semiotic Markers	0/1/2	2
Cultural Practices & Everyday Life	Rituals & Ceremonies	0/1/2	2
	Dress & Aesthetic Practices	0/1/2	2
	Food Practices & Culinary Trad.	0/1/2	2
	Festivals & Commemorations	0/1/2	2
Values, Beliefs & Worldviews	Values & Ethical Norms	0/1/2	2
	Religion & Belief Systems	0/1/2	2
Thematic Coverage & Structure	Thematic Breadth	$3 \times \min(\# \text{topics}, 10)/10$	3
	Reflexivity on Thematic Distribution	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
	Theory-based Categorisation	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
Total			25

Table 4: Scoring schema for Dimension D2 (Cultural Content & Thematic Representation).

Category	Subcategory	Scoring rule	Max. Points
Question Format	Format Type	open $\rightarrow$ 2; MC $\rightarrow$ 1; binary $\rightarrow$ 0	2
Data Constr. & Documentation	Transp. & Traceable Process Doc.	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
	Construction Method	manual $\rightarrow$ 2; semi $\rightarrow$ 1; auto $\rightarrow$ 0	2
Data Selection	Data Sources Disclosed	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
	Validation vs. External Ref. Sources	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
Data Filtering	Dataset Cleaned & Deduplicated	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
	Filtering for Cultural Relevance	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
Quality Control	Final Quality Review of Items	manual $\rightarrow$ 2; auto $\rightarrow$ 1; No $\rightarrow$ 0	2
	Answer Consistency Check	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
Annotator Involvem. & Documentation	Annotator Involvem.	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
	Docu. & Justification of Selection	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
	Cult. & Reg. Relevant Annotators	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
	Annotator Diversity	max 3	3
	Recruitment Procedure Disclosed	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
	Compensation Disclosed	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
	Compr. Annotation Guidelines	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
	Quality Control of Annotations	Yes $\rightarrow$ 1; No $\rightarrow$ 0	1
Total			22

Table 5: Scoring schema for Dimension D3 (Data Foundation). Dimension score = earned points / 22.

Category	Subcategory	Scoring rule	Max. Points
D4: Concept. Understanding of Culture	Culture Explicitly Defined	Yes → 1; No → 0	1
	Definition Elaborated & Justified	Yes → 1; No 0	1
	Definition Supported by Literature	Yes → 1; No → 0	1
	Embedded in Theoretical Framework	Yes → 1; No → 0	1
Total			4
D5: Bias & Fairness	Bias Awareness	Yes → 1; No → 0	1
	Bias Analysis & Mitigation	Yes → 1; No → 0	1
	Fairness & Inclusivity	Yes → 1; No → 0	1
Total			3
D6: Transparency & Reflexivity	Disclosure of Data Sources	Yes → 1; No → 0	1
	Critical Reflexivity	Yes → 1; No → 0	1
Total			2

Table 6: Scoring schema for Dimensions D4–D6.

B Overview of the Analyzed Benchmarks

Table 7 provides an overview of the 50 benchmarks included in our analysis together with their corresponding references.

Benchmark	Reference	Benchmark	Reference
ARADICE	(Mousi et al., 2024)	Hanfu-Bench	(Zhou et al., 2025a)
ArtELingo	(Mohamed et al., 2022)	KULTURE Bench	(Wang et al., 2024b)
ArtELingo-28	(Mohamed et al., 2024)	M5	(Schneider and Sitaram, 2024)
BLEnD	(Myung et al., 2025)	MAKIEVAL	(Zhao et al., 2025)
CAMeL	(Naous et al., 2024)	MAPS	(Liu et al., 2024)
CCUB	(Liu et al., 2023b)	MIXCUBE	(Kim et al., 2025)
CDEval	(Wang et al., 2024c)	NormAd	(Rao et al., 2025)
Challenging Cross-Cultural (C3)	(Liu et al., 2023a)	Nunchi-Bench	(Kim and Lee, 2025)
CII-Bench	(Zhang et al., 2025)	NusaBert	(Wongso et al., 2024)
ClicK	(Kim et al., 2024)	PALM	(Alwajih et al., 2025)
CQ-Bench	(Liu et al., 2025c)	PERCUL	(Moosavi Monazzah et al., 2025)
CROPE	(Nikandrou et al., 2025)	Rice-VL	(Pranav et al., 2025)
CUBE	(Kannen et al., 2024)	RusCode	(Vasilev et al., 2025)
CULTURALVQA	(Nayak et al., 2024)	SANSKRITI	(Maji et al., 2025)
CULTURE-GEN	(Li et al., 2024a)	SeaEval	(Wang et al., 2024a)
CultureAtlas	(Fung et al., 2024)	SEA-HELM	(Susanto et al., 2025)
CulturalBench	(Chiu et al., 2025)	SEA-VL	(Cahyawijaya et al., 2025)
CultureVLM	(Liu et al., 2025b)	SEA-VQA	(Urailetrprasert et al., 2024)
CVQA	(Romero et al., 2024)	SEACrowd	(Lovenia et al., 2024)
DIWALI	(Sahoo et al., 2025)	Seeing Culture (SCB)	(Satar et al., 2025)
DLAMA	(Keleg and Magdy, 2023)	SocialCC	(Wu et al., 2025)
FoodieQA	(Li et al., 2024b)	TCC-Bench	(Xu et al., 2025)
GIMMICK	(Schneider et al., 2025)	WorldCuisines	(Winata et al., 2025)
Global MMLU	(Singh et al., 2025)	WorldValuesBench	(Zhao et al., 2024)
GLOBALRG	(Bhatia et al., 2024)	WorldView-Bench	(Mushtaq et al., 2025)

Table 7: Overview of the 50 analyzed benchmarks.