# *Hire Your Anthropologist!*
# Rethinking Culture Benchmarks Through an Anthropological Lens

**Mai AlKhamissi**[1*]    **Yunze Xiao**[1*]    **Badr AlKhamissi**[2]    **Mona Diab**[1]

[1]Carnegie Mellon University    [2]EPFL
{malkhami,yunzex,mdiab}@andrew.cmu.edu    badr.alkhamissi@epfl.ch

## Abstract

Cultural evaluation of large language models is increasingly important, but many benchmarks still treat culture as static facts or uniform values, conflicting with anthropological views of culture as dynamic, situated, and enacted in practice. In this position paper, we qualitatively analyze 20 handpicked cultural benchmarks and introduce a four-part framework for how they frame culture: *knowledge*, *preference*, *performance*, and *bias*. We then identify six recurring methodological issues, including treating countries as cultures, flattening within-culture diversity, and over-relying on simplified survey formats. Drawing on anthropological methods, we propose concrete improvements such as scenario-based narratives, community-involved design and validation, and contextual evaluation of model behavior. Overall, we aim to push cultural benchmarking beyond recall-style tasks toward evaluations that better capture how models handle complex cultural situations.

## 1 Introduction

Large language models (LLMs) are now being implemented in translation systems (Zhu et al., 2024), educational tools (Sonkar et al., 2024), search engines (Ziems et al., 2023), and generative platforms that engage diverse public opinions across cultural, linguistic, and political contexts (Ziems et al., 2023). As these models become more central to knowledge production and everyday decision-making, it is crucial to assess their sensitivity to cultural nuance(Xiao et al., 2025b). In response, NLP researchers are constructing benchmarks that aim to measure *cultural competence*: the ability of models to respond appropriately to regionally specific norms, moral frameworks, idioms and sociopolitical identities.

However, in this growing body of work, the concept of *culture* is often treated as a background
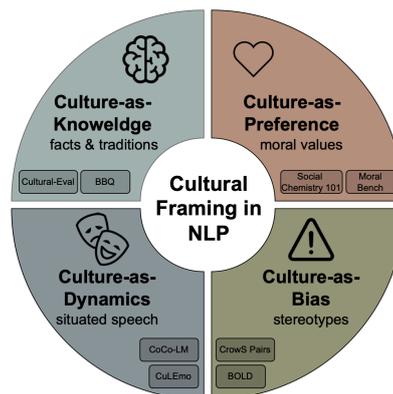
---

[*] Equal contribution.



Figure 1: **Cultural Framing in NLP.** Our taxonomy of how culture is framed in NLP evaluation. Each quadrant represents a distinct theoretical lens on culture: defining what it entails, illustrating how it is expressed, and providing two representative benchmarks for each framing.

variable rather than as a central analytical concern. Benchmark designers rarely engage with anthropology, a discipline that has developed rich tools to understand how culture is lived, contested, and transmitted. Anthropologists study culture not as a fixed set of values or national traits but as a dynamic and situated process shaped by history, power, and everyday meaning-making (Geertz, 1973).

To bridge the gap between NLP and anthropology, we analyze 20 cultural benchmarks used in NLP. We argue that while these benchmarks address important questions, such as moral diversity, stereotype reproduction, and regional variation—they often rely on reductive or decontextualized models of culture. Consequently, our contribution is as follows: (1) We examine how culture is *conceptualized* within benchmarks, developing a taxonomy of cultural framings used by NLP researchers; (2) We show how six common design choices oversimplify culture, reducing it to surface traits or stereotypes, and restricting which forms of cultural knowledge are recognized or ignored;

(3) We offer recommendations for building benchmarks that reflect culture as a lived experience.

## 2 Methods

The choice to analyze 20 benchmarks reflects a methodological commitment to qualitative, humanistic traditions rather than large-scale survey methods. We do not claim statistical representativeness; instead, our goal is to surface recurring conceptual and procedural patterns in how "culture" is operationalized in benchmark design, prioritizing conceptual clarity over statistical coverage. This approach foregrounds epistemologies, assumptions, and cultural politics often obscured in large-scale corpus analyses, emphasizing interpretive depth, reflexivity, and contextual understanding—aligning with scholarship that highlights the limits of scalable methods for capturing nuance, bias, and power dynamics. Scholars such as Blodgett et al. and voices in ACM interactions (Liao et al., 2024; Widder and Kneese, 2025) stress that meaningful cultural analysis demands slow, interdisciplinary, and critically reflexive engagement, a position this study adopts through its carefully curated sample. We sample for coverage across task formats and widely cited benchmark families, prioritizing resources frequently used to support claims about cultural competence, which enables interpretive depth while spanning multiple operationalizations (knowledge, preference, dynamics, bias).

## 3 Background and Related Work

### 3.1 What is *Culture*

Anthropology treats culture as dynamic, situated, and contested, rather than as a stable list of national traits. Across major traditions, culture is approached as historically contextual (e.g., cultural relativism), interpretive meaning-making, and interactional performance, with researchers explicitly reflecting on how accounts of culture are produced (see Appendix A). We draw on these perspectives to motivate four evaluation lenses that align with common NLP task designs: what is known (knowledge), what is valued or preferred (preference), how meaning is enacted in context (dynamics), and where identity-linked harms emerge (bias).

### 3.2 Culture Benchmarks in NLP

NLP research has developed benchmarks to evaluate the handling of cultural knowledge using language models, but these works often define *cul-ture* inconsistently and simplistically compared to anthropological scholarship. Although NLP researchers increasingly recognize that linguistic and social norms vary across cultures and are underrepresented in current resources, significant gaps remain between computational approaches and social science perspectives on culture. The purpose of this work is to foster interdisciplinary dialogue between these fields.

Recent work on cultural and bias evaluation has produced diverse datasets investigating the concept of culture. CultureAtlas (Fung et al., 2024) broadens multilingual evaluation, integrating everyday cultural knowledge across diverse global contexts, while SeaExam (Liu et al., 2025a) emphasizes reasoning and consistency specifically in Southeast Asian scenarios. Furthermore, CDEval (Wang et al., 2024b) systematically maps six cultural dimensions to seven domains, while LLM-GLOBE (Karinshak et al., 2024) adapts the well-established GLOBE cultural values framework to evaluate open-ended LLM outputs.

Datasets focused on normative and moral reasoning continue to expand, exemplified by Social Chemistry 101 (Forbes et al., 2020), which distills common-sense social norms from real-world interactions; and MoralBench (Ji et al., 2024) provides explicit ethical dilemmas for detailed evaluation. Moreover, NormAd (Rao et al., 2025) systematically assesses how models judge the acceptability of everyday etiquette scenarios in 75 countries.

Parallel efforts in bias evaluation are significantly diversified: BBQ (Parrish et al., 2022) identifies stereotypical biases in question-answering tasks; CrowS-Pairs (Nangia et al., 2020) targets implicit biases in masked language models; Social Bias Frames (Sap et al., 2020) annotates offensive implications in language model outputs; and BOLD (Dhamala et al., 2021) evaluates bias in open-ended biographical generation. The VL-BiasBench (Wang et al., 2024a) and multilingual M5 (Schneider and Sitaram, 2024a) suites extend bias assessment to vision language models, while M-Bias (Raza et al., 2024) offers data-driven interventions to reduce toxicity without sacrificing contextual coherence.

Finally, novel datasets are increasingly expanding into affective and application-specific areas: CuLEmo (Belay et al., 2025) benchmarks emotion recognition across cultures, CASA (Qiu et al., 2025) evaluates social and cultural awareness of LLM-powered web agents, and X-FACTR (Jiang

et al., 2020) assesses multilingual factual knowledge through a cloze-style fill-in-the-blanks question in 23 typologically diverse languages.

## 3.3 Critiques within NLP

Recent work in NLP has exposed fundamental weaknesses in the conceptualization of culture in cultural benchmarks. In the literature, there is a consensus that current evaluation frameworks reduce culture to survey-derived constructs that privilege western-centric assumptions while failing to capture dynamic, situated, and contested cultural meanings (Liu, 2024a). Hershcovich et al. (2022) identifies unequal cultural representation advocate for targeted sampling and Distributionally Robust Optimization to prevent cultural generalization from being disproportionately shaped by dominant groups. Zhou et al. (2025) extends these concerns by critiquing the over-reliance on nationality-based categorizations and rigid survey methods that flatten dynamic and intersectional cultural identities, proposing more context-sensitive and stereotype-aware datasets to reflect lived cultural realities.

Several influential papers have further articulated these shortcomings. For example, Chiu et al. (2024) highlight how existing benchmarks are often static, relying heavily on sources such as Wikipedia, and are constrained by narrow survey formats that limit the representation of cultural diversity. These practices risk amplifying urban and western-centric narratives while marginalizing underrepresented perspectives. To counteract this, they introduce CulturalBench, a benchmark designed through human-authored and verified questions that span multiple regions and cultural dimensions, with independent annotator validation to improve robustness and inclusivity. In the same vein, Culture is Not Trivia (Zhou et al., 2025) critiques the over-reliance on nationality and rigid survey methods, arguing that such approaches flatten dynamic and intersectional cultural identities. They propose more context-sensitive and stereotype-aware datasets to better reflect lived cultural realities and support inclusive benchmark design.

**Mitigation Attempts** Researchers have attempted to address these limitations through technical interventions, most notably cultural prompting and semantic augmentation. For example, AlKhamissi et al. (2024) proposed *anthropological prompting* to make LLMs better reflect the opinions of digitally underrepresented groups. Similarly, Tao et al. (2024) shows that explicitly instructing LLMs to adopt specific cultural perspectives can significantly reduce Western bias in most countries and territories, although this approach sometimes introduces new biases and rests on the problematic assumption that culture can be cleaned up by demographic targeting. Although these innovations show promise in controlled settings, they fundamentally maintain the flawed premise that culture can be distilled through survey data or prompt engineering, an assumption increasingly questioned within the research community.

## 3.4 Existing Frameworks for Cultural Evaluation

The framework presented by Vijay and Hershcovich (2024) provides an important foundation for cross-cultural NLP by mapping out four main cultural elements: linguistic form, common ground, aboutness, and values, to guide researchers through the conceptual complexities of working with cultural diversity. Their approach highlights strategic challenges and high-level pathways for ensuring more fair and representative NLP systems. In contrast, our work expands this foundation by focusing directly on how culture is operationalized in evaluation metrics and benchmark design.

Concurrent with our work, Adilazuarda et al. (2024) Adilazuarda et al. offer a high-level taxonomy that distinguishes between demographic and semantic proxies of culture, revealing the field's reliance on operational shortcuts rather than explicit theorizing cultural processes. Complementarily, Liu et al. (2023) (Liu, 2024b) propose a fine-grained taxonomy that organizes cultural elements into ideational, linguistic, and social dimensions, grounded in anthropological definitions and emphasizing the interplay between language and social practice.

Havaldar et al. (2025) and Pawar (2025) push cultural evaluation beyond static text benchmarks. Havaldar et al. introduce a sociocultural framework for the evaluation of conversational LLM, focusing on *linguistic style* as a key site where cultural differences manifest, and proposing new metrics such as *conversational structure, stylistic sensitivity,* and *subjective correctness* to assess how models adapt across cultural contexts. Pawar (2025) similarly broadens the field by offering a cross-modal taxonomy, connecting textual, visual, and multimodal forms of cultural understanding through di-

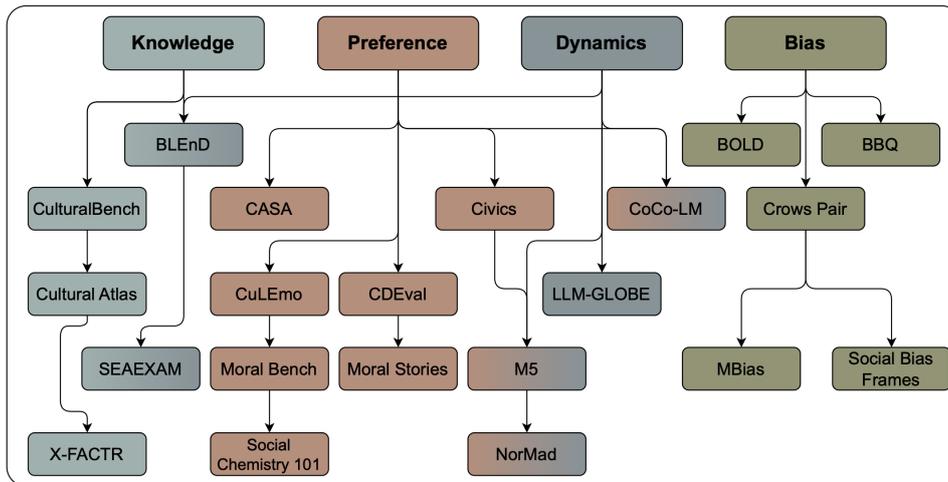**Mapping of 20 Cultural Benchmarks Onto Taxonomy**



Figure 2: **Mapping of 20 Cultural Benchmarks to the 4 Cultural Dimensions** We map the benchmarks analyzed in this work onto the taxonomy proposed in §4. Each benchmark inherits the dimension of its parent(s).

mensions of *representation, knowledge,* and *social sensibility*

Together, these frameworks have provided important structure for an emerging interdisciplinary space, but they also highlight the limits of static categorization. Building on this work, the taxonomy advanced here engages both perspectives: acknowledging the functional distinctions emphasized by Adilazuarda et al. (2024) and the theoretical breadth introduced by Liu et al. (2023). We also draw on more recent developments such as Havaldar et al. (2025), who extend cultural analysis into conversational and stylistic metrics by designing a new benchmark, and Pawar (2025), which integrates multimodal dimensions of cultural understanding. Our framework differs from these by centering the conceptual modeling of culture itself rather than its downstream manifestations in text, interaction, or modality. In our conception, culture is not treated as a set of characteristics, but as a shifting interpretive process situated in discourse, interaction, and power, enabling a richer analysis of how LLMs reproduce and transform cultural meaning across contexts.

## 4 Culture Benchmark Taxonomy

While technical mitigation strategies have advanced the capture of cultural complexity in NLP, our contribution is to systematically examine how culture itself is conceptualized within leading benchmarks. We begin by offering a taxonomy that clarifies the diverse definitions of culture used across these resources, helping reveal the range of cultural perspectives that inform benchmark design. Building on this taxonomy, we present recommendations in §6 to inform future benchmark development and encourage a more deliberate engagement with cultural nuance. We first introduce a four-part taxonomy for cultural evaluation in NLP: culture-as-knowledge, culture-as-preference, culture-as-dynamics, and culture-as-bias. The taxonomy is both descriptive and prescriptive. Descriptively, it distills how current benchmarks operationalize culture; prescriptively, it guides future work to combine lenses rather than treat them in isolation, encouraging multidimensional designs that better capture cultural nuance.

**Culture-as-Knowledge.** Culture consists of facts, traditions, and symbolic references that can be recalled or identified. This lens treats culture as a body of factual or object-level information tied to national or regional identities, for example, benchmarks that assess whether models recognize culturally specific foods, holidays, or practices. This view echoes early *trait list* approaches in anthropology (Boas, 1911).

**Culture-as-Preference.** This angle conceptualizes culture as a set of shared moral, social, or political attitudes, often derived from surveys or aggregate annotations. It reflects a more psychological or normative model of culture as shared values or aggregated moral intuitions, typically measured through survey instruments. Although widely used in cross-cultural psychology (Hofstede, 1980; Inglehart and Welzel, 2005), this framing has

been critiqued in anthropology for reducing culture to a "thin" notion of coherence and boundedness within groups, obscuring internal contestation, hybridity, and change (Ortner, 2006; Abu-Lughod, 1991). Nevertheless, it remains attractive because it is highly operationalizable: surveys and annotation schemes allow for scalable comparisons across populations and across models, even if this comes at the cost of nuance.

**Culture-as-Dynamics.** Culture is not treated as static, but as an entity that acts in specific contexts through language, narrative, and interaction. It draws on linguistic and symbolic anthropology to understand culture as enacted, improvised, and contextual, shaped by speaker positionality, audience, and occasion (Goffman, 1959; Hymes, 1974; Bauman and Briggs, 1990). This view understands culture as something that people actively *perform* in context, rather than an attribute they possess. It focuses on how meaning is shaped through interaction, making it better suited for capturing how models respond to different social situations. Benchmarks such as CoCo-LM, BLEND, Moral Stories, and CuLEmo exemplify this view by modeling how culture is performed in dialogue, story, or emotion expression.

**Culture-as-Bias.** This dimension particularly frames cultures through a series of stereotypes. The goal of this identification is harm reduction and the elimination of bias. It tests how the model outputs reflect and reinforce stereotypes. We introduce this category specifically for NLP evaluations: culture is inferred from patterns of stereotyping or discriminatory output, especially along dimensions of race, gender, or class. Although these benchmarks play an essential role in fairness research, they often rely on binary identity comparisons without interrogating the cultural production of these categories. We aim to introduce an intersectional lens through this category (Crenshaw, 1991; Alim, 2009). BBQ, Social Bias Frames, BOLD, M-Bias, and VL-Bias Bench all operate within this paradigm, focusing on identity-based harm and fairness violations.

## 4.1 Contribution

While Hershcovich et al. (2022) advocate for broad methodological reflection and community engagement, our taxonomy critically dissects actual evaluation practices, exposing how specific methodological choices shape, limit, or obscure cultural complexity. We provide prescriptive strategies for

recording ambiguity, contestation, and dissent as core data signals, and for institutionalizing plural ground truths within benchmarks. Therefore, while their framework maps the strategic landscape of the field, our work intervenes more directly at the level of metrics, offering concrete recommendations for embracing the messiness and multiplicity of lived culture in NLP evaluation.

In contrast, we interrogate how culture is framed in the evaluation metrics themselves, making visible methodological practices such as annotation, bias assessment, and consensus building that can flatten cultural complexity. By analyzing how culture is already encoded in evaluation and urging benchmarks to record zones of dispute and ambiguity, our taxonomy moves beyond cataloging toward a design roadmap for resources that reflect multiplicity, contestation, and lived experience. .

## 4.2 Mapping Benchmarks Onto the Taxonomy

We review 20 prominent cultural benchmarks in NLP and map them to the four cultural dimensions. Datasets such as *CultureAtlas* (Feng, 2024), *BLEnD* (Myung et al., 2025), and *CulturalBench* (Chiu et al., 2024) exemplify **Culture-as-Knowledge**, focusing on factual recall tied to national or regional identity. **Culture-as-Preference** is illustrated by benchmarks such as *Social Chemistry 101* (Forbes et al., 2020), *MoralBench* (Ji et al., 2024), and *NormAd* (Rao et al., 2025), where models are evaluated for alignment with aggregated moral or social judgments. **Culture-as-Dynamics** appears in datasets such as *CoCo-LM* (Meng et al., 2021), *BLEnD*, and *MoralStories*, foregrounding the context-sensitive nature of language and emotion. Finally, **Culture-as-Bias** is captured by benchmarks such as *BBQ* (Parrish et al., 2022), *Social Bias Frames* (Sap et al., 2020), and *BOLD* (Dhamala et al., 2021), which examine how models reproduce identity-based harms.

Figure 2 presents the category breakdown. Each of the four parts of our framework highlights a distinct facet of cultural life and can fit different evaluation goals. However, most of the prominent NLP benchmarks to date lean on two narrow frameworks: culture-as-knowledge and culture-as-preference, while treating culture as a set of static facts or simple surveyable attitudes. In contrast, anthropological work views culture as dynamic, situated, and relational. In our recommendations (§6), we will expand on how to use this taxonomy

in future work.

## 5 Methodological Limitations and Ways Forward

We identify six methodological limitations that flatten cultural complexity in NLP benchmarks. These essentialist frameworks undermine our taxonomy (knowledge, preference, dynamics, bias). Table 1 maps these constraints while highlighting benchmarks that creatively avoid them, emphasizing the need for more reflexive, context-sensitive evaluation

The limitations are: (a) Platform bias favoring Western, urban demographics; (b) Conflating nation states with cultural boundaries; (c) Treating individual annotators as cultural representatives; (d) Reducing moral reasoning to survey scales; (e) Assuming cultural consensus where disagreement exists; and (f) Stripping context from cultural scenarios.

### 5.1 Platform Bias and Demographic Skew

Many cultural benchmarks are drawn from Reddit, X, or Wikipedia, which mostly mirror Western, male, urban, and digitally literate users. This privileges certain demographics over others. Social Chemistry 101 inherits Reddit's community biases, and datasets like BBQ, BOLD, CulturalAtlas, and others import similar skews; the effect extends to CASA, CDEval, CulturalBench, LLM-GLOBE, SeaExam, and X-FACTR. Treating such homogeneous pools as globally representative risks erasing offline, indigenous, rural, or otherwise marginalized worldviews.

Digital anthropology demonstrates that digitally mediated expressions, such as posts and opinions on social networks, do not automatically reflect the full complexity of 'real social facts', as recognized in ethnographic research (Miller, 2018). Miller stresses that digital platforms produce distinctive forms of sociality and cultural interaction, deeply shaped by the technological frameworks and demographics of their users; Interpreting these online traces as direct representations of offline life risks missing the broader social, cultural, and material contexts in which meaning is formed (Horst and Miller, 2012; Miller, 2018). Anthropological analysis insists on holistic ethnography, situating digital worlds within the complexities of larger social relations, and cautioning against reducing the mediation of online experience to simple, universal truths

(Horst and Miller, 2012; Miller, 2018). Digital is not a transparent lens but an active mediator: sometimes intensifying difference, sometimes shaping whose voices become visible or invisible, and always requiring careful contextual understanding to avoid misrepresenting the realities of diverse communities (Horst and Miller, 2012; Miller, 2018).

Partial remedies exist. FLEAD (Federated Learning to Exploit Annotator Disagreements) (Rodríguez-Barroso et al., 2024) models each annotator's perspective to surface minority viewpoints that majority aggregation can suppress, yet its reach is limited by the sparse annotator metadata and the exclusion of communities without digital access.

### 5.2 Nation-State as a Proxy for Culture

Several NLP benchmarks adopt nationality as the primary cultural framework, assuming that country boundaries are mapable to homogeneous cultures. SeaExam and CulturalAtlas assign norms by country with statements such as 'In Chinese culture...', while BLEND, CDEval, LLM-GLOBE and X-FACTR rely on national groups that obscure internal ethnic, linguistic and religious differences. SeaExam aggregates Singapore, Malaysia, and Thailand under one label, and LLM-GLOBE conflates language with culture (for example, Arabic as "Middle Eastern culture"), overlooking diasporic contexts. In agreement with our work, Havaldar et al. also shows how, instead of defining cultures, LLMs instead use proxies using nation states.

Social science scholarship underscores why such assumptions are fundamentally problematic. Gupta and Ferguson argues that nation states are politically constructed entities, shaped by transnational flows, colonization histories, and shifting political economies, rather than natural or coherent cultural units. Treating national borders as cultural boundaries flattens complex identities and histories, ignoring the dynamic, overlapping, and contested nature of culture that ethnographic research reveals. This critique aligns with broader calls within anthropology and cultural studies to move beyond the nation state as the default unit of cultural analysis and instead to address the multiplicity and fluidity of cultural formations (Gupta and Ferguson, 1992; Appadurai, 1996).

BLEND demonstrates progress by combining intra- and cross-national perspectives, and Jiraibench (Xiao et al., 2025a) investigates how meanings change between discourse types, platforms, and diasporic communities. Jiraibench is an

| Benchmark | Platform Bias | Nation-State | Annotation | Moral Simplification | Consensus | Lack of Context |
|---|---|---|---|---|---|---|
| BBQ | ✓ | – | ✓ | – | – | – |
| BLEND | ✓ | ✓ | ✓ | – | – | – |
| BOLD | ✓ | – | ✓ | – | – | – |
| CASA | ✓ | – | ✓ | ✓ | ✓ | – |
| CDEval | ✓ | ✓ | ✓ | ✓ | ✓ | – |
| Civics | ✓ | – | ✓ | ✓ | – | ✓ |
| CoCo-LM | ✓ | – | – | – | – | – |
| Crows-Pairs | ✓ | – | ✓ | – | – | – |
| CuLEmo | – | – | ✓ | – | – | ✓ |
| CulturalAtlas | ✓ | ✓ | – | – | – | ✓ |
| CulturalBench | ✓ | – | ✓ | – | – | – |
| LLM-GLOBE | ✓ | ✓ | – | – | – | – |
| M-Bias | ✓ | – | ✓ | – | – | – |
| M5 | ✓ | – | – | – | – | – |
| Moral Bench | ✓ | – | ✓ | ✓ | ✓ | ✓ |
| Moral Stories | ✓ | – | ✓ | ✓ | ✓ | ✓ |
| NormAd | ✓ | – | ✓ | ✓ | ✓ | – |
| SeaExam | ✓ | ✓ | ✓ | – | – | – |
| Social Bias Frames | ✓ | – | ✓ | – | – | – |
| Social Chemistry 101 | ✓ | – | ✓ | ✓ | ✓ | – |
| X-FACTR | ✓ | ✓ | – | – | – | – |

Table 1: **Presence of Methodological Limitations Across 20 Cultural Benchmarks.** Each column indicates the presence (✓) or absence (–) of a limitation defined in Section 5: **Platform Bias** (§5.1), **Nation-State** (§5.2), **Annotation** (§5.3), **Moral Simplification** (§5.4), **Consensus** (§5.5), **Lack of Context** (§5.6).

exploratory first step; sustained research is needed to develop richer, more situated evaluations that reflect contested, plural, and evolving cultural life.

## 5.3 Representation Fallacy Through Annotation

Many benchmarks assume that a few annotators can represent the cultural norms of a group. This affects BBQ, BLEND, CASA, CDEval, SEAExam, LLM-GLOBE and others. SEAExam collapses Singapore, Malaysia, and Thailand into one label, and LLM-GLOBE rarely reports annotator demographics. Treating such labels as comprehensive removes the variation by class, gender, generation, and religion. Using LLMs as "cultural judges" without external validation amplifies the problem.

As Abu-Lughod (1991) and Mahmood (2005) note, representation is not only who speaks, but also how voices are situated within power relations. Little is understood of how annotation choices, sampling practices, or positionality of annotators shape the contours of what counts as 'culture' in these data sets. This methodological opacity risks reinforcing normative assumptions, perpetuating a narrow view of cultural diversity that can marginalize less visible voices and mask tensions within communities. Without reflexive attention to which experiences are foregrounded and which are silenced, benchmarks may inadvertently produce a shallow or exclusionary portrait of cultural life.

## 5.4 Moral Simplification through Survey Format Overload

Many "cultural reasoning" benchmarks reduce Culture-as-Preference and Culture-as-Dynamics to survey responses, collapsing nuanced judgments into Likert scales or binaries. CASA, CDEval, Civics, Moral Bench, Moral Stories, NormAd, and Social Chemistry 101 rely on simplified prompts that flatten context. Moral Bench, grounded in Moral Foundations Theory built from WEIRD samples, treats ethics as consensus scores and projects a narrow frame as universal. Anthropological work shows why this fails: moral life is context-dependent, narratively situated, and historically contingent, not a set of fixed rules (Zigon, 2007; Laidlaw, 2013).

## 5.5 Assumption of Cultural Consensus

Many benchmarks treat aggregated judgments as correct answers, assuming cultural agreement on norms, which eliminates the contested nature essential to Culture-as-Dynamics. CASA, CDEval, Moral Bench, Moral Stories, NormAd, and Social Chemistry 101 use majority votes as ground truth. Social Chemistry 101 collects explanations showing diverse reasoning, but still trains models on the most common answers. This treats disagreement as noise, rather than recognizing culture as a site of ongoing negotiation. As Gal (2015) and Tsing (2005) emphasize, disagreement is a fundamental feature of cultural life, not a flaw.

A more promising directions treat culture as situated and temporal, emphasizing multiplicity over

universality. BLEND (Myung et al., 2025) advances this by evaluating culturally situated knowledge in 16 regions and 13 languages using contemporary community-informed data, and SEACrowd (Lovenia et al., 2024) shows what large-scale community collaboration can achieve with more than 400 Southeast Asian contributors.

## 5.6 Decontextualized Prompts and Abstracted Norms

Cultural benchmarks often test models on questions disconnected from historical and political context, which strips away the situational grounding necessary for Culture-as-Dynamics. For example, BBQ presents stereotype questions without a grounding in social histories, CASA assumes universal norms, and CulturalAtlas reduces culture to decontextualized trivia (Zhou, 2024). By isolating questions from their environments, these benchmarks treat cultural reasoning as context-free, overlooking how norms are formed and contested within specific conditions.

Some benchmarks move toward scenario-based tasks that embed challenges in locally meaningful contexts (Bravansky et al., 2025). However, these often rely on generic scenarios that lack microcontextual cues such as tone, history, or power dynamics (Blodgett et al., 2020). The most promising direction involves observing model behavior rather than just output. Following Li et al. (2025), who found that models with minimal stated bias still exhibit *significant sociodemographic disparities* in simulations, the evaluation must recognize the gaps between the stated results and contextual decisions.

## 6 Recommendations

These six design choices arise from the drive for scalability and quantifiability of NLP, which conflicts with the anthropological notions of situated practice. Here is our list of recommendations for ways forward.

### 6.1 Social Science Collaboration

Our central recommendation is to establish ongoing, methodical collaborations that position social scientists as integral partners throughout every stage of AI and NLP research. Drawing on blueprints developed in interdisciplinary initiatives and early prototypes from our own work, we propose a pipeline that spans problem definition, data collection, benchmark design, annotation, and evaluation, with social scientists actively co-leading each phase. Our ongoing project seeks to formalize such practices, providing concrete steps and governance models for co-designing research agendas, annotation guidelines, and evaluation protocols to ensure that social scientific expertise is embedded and reflected in every methodological decision.

### 6.2 Decoupling Culture from Nation

Decoupling culture from the nation-state within AI systems requires moving beyond the assumption that national borders neatly align with cultural boundaries and instead recognizing the internal diversity, hybridity, and mobility present within and across nations. Although the category of the nation can remain analytically useful, benchmarks and models should treat it as one layer among many, situating national identity alongside transnational, diasporic, and local cultural affiliations. This can be achieved by designing evaluations that explicitly compare how core concepts (such as freedom, kinship, or authority) are deployed across legal, religious, activist, and everyday discourses, and by collaborating with local and transnational communities to identify relevant categories, axes of variation, and migration histories. Through such pluralistic methods, AI systems can avoid essentializing culture as nation-bound while leveraging the nation category for context-specific inferences, regulatory needs, or critical comparisons where appropriate.

### 6.3 Critical Annotation

A critical annotation approach must move beyond simply measuring annotator disagreement, recognizing that annotation itself is a social and interpretive act deeply shaped by the power relations and positionality within annotation work. While existing tools for calculating inter-annotator agreement help quantify variation, this is only a partial solution: a single annotator's choices reflect not the entirety of a culture or religion, but their own relationship with and perspective on those traditions— a performance influenced by social, economic, and institutional contexts. Drawing from anthropology's tradition of 'writing culture' (Clifford and Marcus, 1986), critical annotation recommends protocols that allow ambiguity and narrative, incorporating continuous scales, narrative storytelling, and space for dissent and self-reflection rather than forcing binary consensus. Methods should encour-

age iterative, community-driven processes where annotation guidelines and ethical dilemmas are refined collaboratively, capturing multiple, contested meanings and the complexity of lived experience.

### 6.4 Participatory and Emic Evaluation

Our recommendations align with HCI and ICTD traditions that treat evaluation as a socio-technical process shaped by power, participation, and local accountability (Suchman, 1987; Costanza-Chock, 2020; Harrison et al., 2011).. Participatory and emic approaches in these fields emphasize that categories and success criteria should be co-defined with affected communities, and that validation can occur through lightweight consultation, iterative co-design, and situated feedback rather than one-shot extraction. Bringing these practices into NLP benchmarking suggests concrete design shifts: consult community members during scenario construction to identify salient roles, harms, and contested norms; validate rubrics against local interpretations; and document how evaluation categories were negotiated (Dell and Kumar, 2016; Anokwa et al., 2009; Wyche, 2020; Harrington et al., 2019).. This reframes "cultural competence" as an accountable design and validation problem rather than a purely technical scaling problem.

Beyond participatory construction, benchmarks must also evaluate models in dynamic, interactive settings that resemble ethnographic observation—BLEND, for instance, progresses through community-informed scenarios, while action-based approaches assess model behavior in simulated environments. The field should deploy ethnographic testing that captures adaptation strategies and norm negotiation, embeds micro-contextual cues, and co-designs scenarios with cultural insiders, tracing how models navigate competing norms and make implicit judgments in open-ended interactions rather than static Q&A.

### 6.5 Mapping Controversy

Tools should be developed to systematically map zones of controversy in datasets, institutionalizing plural 'ground truths' where contestation and disagreement are recognized as valuable data signals. Rather than erasing dissent, benchmarks should record, analyze, and represent sites of active dispute as key sources of interpretive richness. Empowering communities to identify which disagreements matter and leveraging critical annotation practices shifts evaluation away from artificial consensus and allows AI systems to better reflect the multiplicity and contested nature of culture.

To improve on these methods, future benchmarks can systematize the documentation of annotation rationales, support longitudinal annotation so that evolving viewpoints are tracked, and provide transparent metadata about power structures and institutional roles in the annotation process. Involving community members directly in the identification and contextualization of disputes will further root ground truths in lived experience.

### 6.6 Using the Full Taxonomy

Our taxonomy bridges gaps between how NLP benchmarks and anthropology conceptualize culture by distinguishing culture-as-knowledge, culture-as-preference, culture-as-dynamics, and culture-as-bias. We recommend using all four dimensions together, and placing them in dialogue with complementary frameworks, including demographic and semantic proxies Adilazuarda et al., ideational, linguistic, and social dimensions Liu et al., and multimodal and interactional approaches Pawar et al.. Taken together, these perspectives can improve both conceptual clarity and benchmark design, enabling more dynamic, intersectional, and context-aware evaluations of how NLP systems represent and negotiate culture.

## 7 Conclusion

Current NLP benchmarks often reduce culture to trivia, and common fixes like diverse annotators, localized fine-tuning, or disagreement-aware metrics still capture only fragments. Drawing on anthropology, we propose a flexible taxonomy and an interdisciplinary pipeline in which social scientists collaborate with NLP researchers across design, data, modeling, and evaluation. Social theory can guide scenario-based, context-rich prompting and treat annotator disagreement as a signal to interpret rather than noise to suppress, while NLP methods help scale these insights. Sustained collaboration replaces ad hoc interventions with iterative co-design and joint interpretation, yielding benchmarks that reflect fluid identities and models that engage culture as lived, negotiated, and plural practice, improving both bias mitigation and interpretability while offering a practical roadmap for more robust, inclusive, and dynamic cultural evaluation co-produced with the communities they are meant to serve.

## Limitations

Our study provides a comprehensive qualitative critique of existing cultural benchmarks and offers concrete desiderata for building more realistically grounded cultural benchmarks based on deep anthropological tenets. However, we acknowledge some critical limitations. A significant constraint lies in the scope of the benchmarks analyzed. Although we reviewed 20 cultural benchmarks, our selection was necessarily limited to publicly available datasets and widely recognized frameworks. As a result, other relevant benchmarks, particularly those that are emerging or proprietary, may provide additional insights that were not covered in this work. In particular, we did not review all benchmarks that work to address these issues. Several key datasets, such as the PRISM alignment dataset (Kirk et al., 2025), Value Kaleidoscope (Sorensen et al., 2024), and Social Norms in Cinema (Rai et al., 2024), have made significant strides toward overcoming anglocentric biases and embracing pluralistic cultural perspectives. However, we excluded them from our analyses to maintain a bird's eye view focused on categorizing broader types of critique in benchmarks, rather than conducting an exhaustive dataset inventory.

Our study focuses primarily on benchmarking methodologies rather than assessing the real-world impact of cultural benchmarks on LLM applications. Although improving benchmarks is a crucial step towards cultural sensitivity in AI, future research should investigate how these advancements translate to tangible improvements in AI-human interactions across various domains, including education, governance, healthcare and cross-cultural communication. Without real-world deployment analysis, the long-term effectiveness of improved cultural benchmarks remains an open question.

In addition, our position paper focuses on text-based evaluations, omitting vision-oriented culture benchmarks that interrogate the understanding of models of imagery, symbolism, and geospecific visual practices. Multimodal datasets such as CulturalVQA (visual question answering on food, clothing and rituals drawn from 11 countries) (Nayak et al., 2024), CVQA (Romero et al., 2024) (a culturally diverse and multilingual VQA set spanning 28 nations and 26 languages), and World-Cuisine (Winata et al., 2025) (a dataset of world cuisines) demand analytical framework (e.g. visual-semantics grounding, region-specific iconography, and image–text alignment biases) that differ fundamentally from the linguistic constructs underlying our framework. Incorporating them would therefore require distinct taxonomies and fairness diagnostics, which we flag as an important avenue for future multimodal culture work.

We also recognize that we are not the first to introduce a taxonomy of cultural understanding in NLP (Liu et al., 2025c). However, in contrast, our four-part taxonomy distinguishes between culture-as-knowledge (factual/symbolic information), culture-as-preference (shared values and moral attitudes), culture-as-dynamics (performative, enacted, context-dependent practices), and culture-as-bias (patterns of stereotyping and harm reduction). Our framework is not only descriptive, mapping how culture is conceptualized, but also prescriptive, challenging single-axis approaches, and recommending that future benchmarks intentionally combine these dimensions to better capture the complexity and contested nature of cultural life.

We recognize the value of concurrent work toward shared definitions, but claim that our four-part taxonomy and critical evaluation analysis provide a foundation for both understanding and transforming cultural benchmarking, moving from descriptive surveys to actionable recommendations that place contestation and plural ground truths at the heart of NLP evaluation.

Despite these limitations, we believe that this work serves as an important step toward rethinking cultural benchmarking in AI.

## Ethical Statement

This position paper advocates for a rethinking of cultural benchmarks in Large Language Models (LLMs) through an anthropological lens, a perspective that introduces several ethical considerations regarding the framing, methodology, and potential impact of our analysis.

The issue of intellectual ownership and participatory ethics is relevant to this discussion. By advocating for the incorporation of ethnographic data and lived experiences into AI benchmarks, this paper indirectly raises concerns about how cultural knowledge is sourced and used. Ethical AI research must avoid extractive approaches that commodify cultural knowledge without consent or compensation. The recommendations presented in our paper should be interpreted as a call for ethically co-designed benchmarks rather than an endorse-

ment of appropriating cultural insights without the involvement of the communities they pertain to.

Furthermore, this paper discusses bias and misrepresentation in AI, but operates within the constraints of currently available benchmarks. The selection of 20 cultural benchmarks, while intended to be comprehensive, is still shaped by the availability of research, the predominance of sources in English, and our own methodological choices. Any analysis of AI benchmarks carries the risk of reinforcing certain academic and institutional perspectives while overlooking others. Future interdisciplinary research should seek to include voices from non-Western epistemologies, indigenous knowledge systems, and community-driven AI governance models.

In general, this paper seeks to approach cultural benchmarking in AI with ethical responsibility, advocating methodologies that are rigorous and respectful of the diverse human experiences they seek to represent. However, the very act of critiquing, selecting, and proposing cultural benchmarks must itself remain subject to ethical scrutiny, ensuring that the pursuit of cultural sensitivity in AI does not inadvertently replicate the biases and epistemic imbalances it aims to address.

# References

Lila Abu-Lughod. 1991. Writing against culture. In Richard G. Fox, editor, *Recapturing Anthropology: Working in the Present*, pages 137–162. School of American Research Press, Santa Fe, NM.

Lila Abu-Lughod. 1991. Writing against culture. In Richard G. Fox, editor, *Recapturing Anthropology: Working in the Present*, pages 137–162. School of American Research Press, Santa Fe, NM.

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.

H. Samy Alim. 2009. Translocal style communities: Hip hop youth as cultural theorists of style, language, and globalization. In H. Samy Alim, Awad Ibrahim, and Alastair Pennycook, editors, *Global Linguistic Flows: Hip Hop Cultures, Youth Identities, and the Politics of Language*, pages 3–26. Routledge, New York.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Yaw Anokwa, Thomas N. Smyth, Divya Ramachandran, Jahanzeb Sherwani, Yael Schwartzman, Rowena Luk, Melissa Ho, Neema Moraveji, and Brian DeRenzi. 2009. Stories from the field: Reflections on hci4d experiences. *Information Technologies & International Development*, 5(4):101–116.

Arjun Appadurai. 1996. *Modernity at Large: Cultural Dimensions of Globalization*. Public Worlds. University of Minnesota Press, Minneapolis, MN.

Richard Bauman and Charles L. Briggs. 1990. Poetics and performance as critical perspectives on language and social life. *Annual Review of Anthropology*, 19:59–88.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Culemo: Cultural lenses on emotion – benchmarking llms for cross-cultural emotion understanding. *Preprint*, arXiv:2503.10688.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. 2024. Human-centered evaluation of language technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 39–43, Miami, Florida, USA. Association for Computational Linguistics.

Franz Boas. 1911. *The Mind of Primitive Man*. Macmillan, New York.

Michal Bravansky, Filip Trhlik, and Fazl Barez. 2025. Rethinking ai cultural alignment. *Preprint*, arXiv:2501.07751.

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. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *Preprint*, arXiv:2410.02677.

James Clifford and George E. Marcus, editors. 1986. *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley.

Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press.

Kimberlé Crenshaw. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299.

Nicola Dell and Neha Kumar. 2016. The ins and outs of hci for development. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, New York, NY, USA. ACM.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *Preprint*, arXiv:2012.15738.

Shutong Feng. 2024. Towards emotion-aware task-oriented dialogue systems in the era of large language models. In *Proceedings of the 20th Workshop of Young Researchers' Roundtable on Spoken Dialogue Systems*, pages 81–83, Kyoto, Japan. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *Preprint*, arXiv:2402.09369.

Susan Gal. 2015. Politics of translation. *Annual Review of Anthropology*, 44:225–240.

Clifford Geertz. 1973. *The Interpretation of Cultures*. Basic Books, New York.

Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Doubleday, Garden City, NY.

Akhil Gupta and James Ferguson. 1992. Beyond "culture": Space, identity, and the politics of difference. *Cultural Anthropology*, 7(1):6–23.

Christina N. Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

Steve Harrison, Phoebe Sengers, and Deborah Tatar. 2011. Making epistemological trouble: Third-paradigm hci as successor science. *Interacting with Computers*, 23(5):385–392.

Shreya Havaldar, Sunny Rai, Young-Min Cho, and Lyle Ungar. 2025. Culturally-aware conversations: A framework & benchmark for llms. *arXiv preprint arXiv:2510.11563*.

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.

Geert Hofstede. 1980. *Culture's Consequences: International Differences in Work-Related Values*. Sage Publications, Beverly Hills, CA.

Heather A. Horst and Daniel Miller, editors. 2012. *Digital Anthropology*. Routledge, London and New York.

Dell Hymes. 1974. *Foundations in Sociolinguistics: An Ethnographic Approach*. University of Pennsylvania Press, Philadelphia, PA.

Ronald Inglehart and Christian Welzel. 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge University Press, New York.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2024. Moralbench: Moral evaluation of llms. *Preprint*, arXiv:2406.04428.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *Preprint*, arXiv:2411.06032.

Hannah Rose Kirk, Alice Whitefield, Paul Rottger, Alison M. Bean, Katerina Margatina, Ricardo Mosquera-Gomez, and Scott Hale. 2025. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 105236–105344.

James Laidlaw. 2013. *The Subject of Virtue: An Anthropology of Ethics and Freedom.* Cambridge University Press, Cambridge.

Yuxuan Li, Hirokazu Shirado, and Sauvik Das. 2025. Actions speak louder than words: Agent decisions reveal implicit biases in language models. *Preprint*, arXiv:2501.17420.

Q. Vera Liao, Mihaela Vorvoreanu, Hari Subramonyam, and Lauren Wilcox. 2024. Ux matters: The critical role of ux in responsible ai. *Interactions*, 31(4):22–29.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025a. Seaexam and seabench: Benchmarking llms with local multilingual questions in southeast asia. *Preprint*, arXiv:2502.06298.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025b. SeaExam and SeaBench: Benchmarking LLMs with local multilingual questions in Southeast Asia. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6119–6136, Albuquerque, New Mexico. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025c. 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.

Jiangming Liu. 2024a. Model-agnostic cross-lingual training for discourse representation structure parsing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11486–11497, Torino, Italia. ELRA and ICCL.

Xiao Liu, Yansong Feng, and Kai-Wei Chang. 2024. CASA: Causality-driven argument sufficiency assessment. 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 5282–5302, Mexico City, Mexico. Association for Computational Linguistics.

Zhaoming Liu. 2024b. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James Validad 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 Hadiwijaya, 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.

Saba Mahmood. 2005. *Politics of Piety: The Islamic Revival and the Feminist Subject.* Princeton University Press, Princeton, NJ.

Yu Meng, Chenyan Xiong, Payal Baja, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: correcting and contrasting text sequences for language model pretraining. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Daniel Miller. 2018. Digital anthropology. In Felix Stein, editor, *The Open Encyclopedia of Anthropology*. University of Cambridge. Updated 2023.

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.

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)*, pages 1953–1967, Online. 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.

Sherry B. Ortner. 2006. *Anthropology and Social Theory: Culture, Power, and the Acting Subject.* Duke University Press, Durham, NC.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. Civics: Building a dataset for examining culturally-informed values in large language models. *Preprint*, arXiv:2405.13974.

Haoyi Qiu, Alexander R. Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of llm web agents. *Preprint*, arXiv:2410.23252.

Saurabh Rai, Kunal J. Zaveri, Shruti Havaldar, Samarth Nema, Lyle Ungar, and Sharath Chandra Guntuku. 2024. Social norms in cinema: A cross-cultural analysis of shame, pride and prejudice. *arXiv preprint arXiv:2402.11333*.

Abhinav Sukumar 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.

Shaina Raza, Ananya Raval, and Veronica Chatrath. 2024. MBIAS: Mitigating bias in large language models while retaining context. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 97–111, Bangkok, Thailand. Association for Computational Linguistics.

Nuria Rodríguez-Barroso, Eugenio Martínez Cámara, Jose Camacho Collados, M. Victoria Luzón, and Francisco Herrera. 2024. Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, 12:630–648.

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.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Florian Schneider and Sunayana Sitaram. 2024a. M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. *Preprint*, arXiv:2407.03791.

Florian Schneider and Sunayana Sitaram. 2024b. 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.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. 2024. Pedagogical alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13641–13650, Miami, Florida, USA. Association for Computational Linguistics.

Tyler Sorensen, Lisa Jiang, Jena D. Hwang, Sergey Levine, Valentina Pyatkin, Peter West, and Yejin Choi. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947.

Lucy A. Suchman. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, Cambridge, UK.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Anna Lowenhaupt Tsing. 2005. *Friction: An Ethnography of Global Connection*. Princeton University Press, Princeton, NJ.

Supriti Vijay and Daniel Hershcovich. 2024. Can Abstract Meaning Representation facilitate fair legal judgement predictions? In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 101–109, Mexico City, Mexico. Association for Computational Linguistics.

Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. 2024a. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *Preprint*, arXiv:2406.14194.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024b. 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.

David Gray Widder and Tamara Kneese. 2025. Salvage anthropology and low-resource nlp: What computer science should learn from the social sciences. *Interactions*, 32(2):46–53.

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.

Susan Wyche. 2020. Using cultural probes in hci4d/ictd: A design case study from bungoma, kenya. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23.

Yunze Xiao, Tingyu He, Lionel Z. Wang, Yiming Ma, Xingyu Song, Xiaohang Xu, Irene Li, and Ka Chung Ng. 2025a. Jiraibench: A bilingual benchmark for evaluating large language models' detection of human self-destructive behavior content in jirai community. *Preprint*, arXiv:2503.21679.

Yunze Xiao, Lynnette Hui Xian Ng, Jiarui Liu, and Mona T. Diab. 2025b. Humanizing machines: Rethinking llm anthropomorphism through a multi-level framework of design. *Preprint*, arXiv:2508.17573.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. *Preprint*, arXiv:2502.12057.

Xulin Zhou. 2024. Towards a co-creation dialogue system. In *Proceedings of the 20th Workshop of Young Researchers' Roundtable on Spoken Dialogue Systems*, pages 48–49, Kyoto, Japan. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023. Large language models are built-in autoregressive search engines. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2666–2678, Toronto, Canada. Association for Computational Linguistics.

Jarrett Zigon. 2007. Moral breakdown and the ethical demand: A theoretical framework for an anthropology of moralities. *Anthropological Theory*, 7(2):131–150.

# Appendix

## A  Anthropological Accounts of Culture: A Short Primer

Anthropologists have historically studied culture through immersive qualitative methods such as ethnography, long-term fieldwork, and participant observation. Rather than reducing culture to a set of traits or norms, anthropologists emphasize its fluidity, contestation, and entanglement with everyday practices. Thus, the concept of culture has evolved significantly in the social sciences, shifting from fixed categories to fluid and dynamic interpretations.

### A.1  Historical and contextual accounts

Early anthropological work, notably Boas' cultural relativism (Boas, 1911), rejected universal hierarchies and emphasized understanding each culture on its own terms. Boas argued against simplistic "trait-list" models, proposing instead that culture should be understood historically and contextually.

### A.2  Interpretive accounts

Subsequent interpretive approaches, epitomized by Geertz (1973), reframed culture as "webs of significance" woven by human beings. Geertz emphasized that cultural analysis requires "thick description," the detailed interpretation of meanings embedded in everyday actions, rituals, and symbols.

### A.3  Reflexive accounts

The interpretive approach gave way to the "writing culture" turn, where culture is not an object to be discovered "out there," but is constructed through the process of writing about culture (Clifford and Marcus, 1986). This perspective foregrounds that cultural accounts are authored and positioned.

### A.4  Interactional and performative accounts

Building on these foundations, Goffman (1959) introduced the notion of culture as performative and interactional, suggesting that cultural meanings are continuously enacted and negotiated through everyday interactions and social performances.

### A.5  Connection to our taxonomy

In this paper, we operationalize these ideas through four evaluation lenses that align with common NLP task designs: what is known (knowledge), what is valued or preferred (preference), how meaning is enacted in context (dynamics), and where identity-linked harms emerge (bias).

## B  Case Study: CASA vs BLEND

This section contrasts the BLEND and CASA benchmarks, highlighting their distinct strengths in evaluating cultural understanding in large language models (LLMs).

Both benchmarks operationalize culture in distinct ways that reflect different theoretical commitments. BLEND treats culture as practices, objects, and habits that form part of daily life. It evaluates models' ability to recall and articulate such details through factual question–answer pairs covering domains like food, family rituals, holidays, education, and sports. For example, BLEND might ask what dishes are typical at a child's birthday in Mexico or what musical instruments schoolchildren commonly learn in Nigeria. These prompts test whether a model associates the correct cultural referents. This includes example such as tamales for celebratory meals or talking drums in West African settings. This demonstrates how well it reproduces localized knowledge across 16 regions and 13 languages.

CASA, in contrast, operationalizes culture through social norms and moral values, assessing how LLMs behave as socially aware agents rather than repositories of factual knowledge. It situates models in simulated web tasks, such as online shopping or social discussion forums, to examine sensitivity to norm violations and culturally appropriate conduct. For instance, an agent might need to recommend gifts while respecting gender or age conventions in different regions, or navigate a discussion about politeness norms in public spaces by avoiding culturally insensitive replies. Thus, CASA quantifies 'cultural awareness' in terms of how often models recognize and appropriately manage variations in social expectations, such as different comfort with assertive speech or deference to elders, rather than factual precision. Although BLEND challenges the breadth of LLM by introducing multiple details to recall for a culture, CASA may oversimplify by overlooking contestations or differences between cultures.

BLEND excels in assessing **culture-as-knowledge**. It immerses models in realistic conversational scenarios that demand the contextual recall and appropriate application of culturally specific facts. For example, prompts might involve

characters discussing preparations for Lunar New Year or Diwali, testing the model's ability to refer to customary foods, greetings, or rituals in natural dialogue. This tests the integration of factual knowledge, such as understanding that joss paper is burned in Chinese ancestral rituals or that modaks are made during Ganesh Chaturthi.

In contrast, CASA is particularly strong in capturing **culture-as-preference**. It presents models with survey-style statements reflecting social or moral attitudes and assesses whether the models align their responses with the presumed dominant values of particular cultural groups. For example, CASA tasks might ask, "In your culture, is it acceptable to speak loudly in public spaces?" or assess whether a model expresses greater agreement with statements endorsing filial piety in East Asian contexts versus valuing individual autonomy in Western contexts.

## C   Description of Reviewed Benchmarks

Table 2 shows the benchmarks reviewed in this work classified using the taxonomy introduced in §C.

| Benchmark | Description | Citation |
|---|---|---|
| BBQ | Tests bias robustness in QA across nine U.S. demographic dimensions with ambiguous/disambiguated contexts. | Parrish et al. (2022) |
| BLEND | 52.6K culturally grounded Q&A pairs spanning 16 regions and 13 languages for evaluating everyday knowledge. | Myung et al. (2025) |
| BOLD | Benchmarks social biases in open-ended generation across five domains with 23,67=9 prompts. | Dhamala et al. (2021) |
| CASA | Corpus of culturally aware situational actions testing context-appropriate behavioral selection across societies. | Liu et al. (2024) |
| CDEval | Cross-domain suite benchmarking cultural adaptability across news, social media, and literature with expert-verified items. | Wang et al. (2024b) |
| Civics | Corpus of civic-knowledge questions (history, government, social studies) probing democratic processes and civics literacy. | Pistilli et al. (2024) |
| CoCo-LM | Pretraining objective that corrects/contrasts corrupted sequences, improving efficiency on GLUE, SQuAD, and culturally sensitive tasks. | Meng et al. (2021) |
| Crows-Pairs | Minimal-pair corpus measuring stereotype direction and strength across social categories. | Nangia et al. (2020) |
| CuLEmo | Cross-cultural emotion-understanding benchmark across six languages requiring nuanced cultural reasoning. | Belay et al. (2025) |
| CultureAtlas | A fine-grained multicultural dataset (127K assertions, 10K+ cities, 2.5K+ ethnolinguistic groups) that enables cultural-knowledge reasoning and norm-violation detection. | Feng (2024) |
| CulturalBench | 1,227 human-verified questions across 45 regions and 17 topics for evaluating cultural knowledge and adaptability. | Chiu et al. (2024) |
| LLM-GLOBE | Assesses cultural values in LLM outputs via the GLOBE framework, comparing alignment with societal value systems. | Karinshak et al. (2024) |
| M-Bias | Multilingual corpus probing gender, profession, and cultural stereotypes in 50+ languages. | Raza et al. (2024) |
| M5 | Multimodal vision–language benchmark spanning 41 languages to highlight performance disparities. | Schneider and Sitaram (2024b) |
| Moral Bench | Evaluates LLMs' moral reasoning via Moral Foundations Theory, measuring ethical alignment. | Ji et al. (2024) |
| Moral Stories | Branching-narrative dataset for grounded social reasoning over norms, intents, and consequences. | Emelin et al. (2020) |
| NormAd | Measures model adaptability to varying cultural norms through social-acceptability judgments. | Rao et al. (2025) |
| SeaExam | A benchmark derived from Southeast Asian educational exams for assessing LLM proficiency in regional languages and cultural contexts. | Liu et al. (2025b) |
| Social Bias Frames | 150K annotations capturing pragmatic bias implications in language for nuanced social-media detection. | Sap et al. (2020) |
| Social Chemistry 101 | 292K commonsense social norms across 12 dimensions for everyday moral-norm reasoning. | Forbes et al. (2020) |
| X-FACTR | Multilingual fact-checking benchmark with claims in 25 languages to test factual verification and cross-lingual generalization. | Jiang et al. (2020) |

7e34

Table 2: Descriptions and citations for 20 cultural NLP benchmarks, ordered alphabetically by benchmark name.