Culture is Everywhere: A Call for Intentionally Cultural Evaluation

Juhyun Oh[◇], Inha Cha[†], Michael Saxon[‡], Hyunseung Lim[◇], Shaily Bhatt[*], Alice Oh[◇]

[◇]KAIST [†]Georgia Tech [‡]UC Santa Barbara [*]Carnegie Mellon University

411juhyun@kaist.ac.kr

Abstract

The prevailing "trivia-centered paradigm" for evaluating the cultural alignment of large language models (LLMs) is increasingly inadequate as these models become more advanced and widely deployed. Existing approaches typically reduce culture to static facts or values, testing models via multiple-choice or short-answer questions that treat culture as isolated trivia. Such methods neglect the pluralistic and interactive realities of culture, and overlook how cultural assumptions permeate even ostensibly "neutral" evaluation settings. In this position paper, we argue for **intentionally cultural evaluation**: an approach that systematically examines the cultural assumptions embedded in all aspects of evaluation, not just in explicitly cultural tasks. We systematically characterize the what, how, and circumstances by which culturally contingent considerations arise in evaluation, and emphasize the importance of researcher positionality for fostering inclusive, culturally aligned NLP research. Finally, we discuss implications and future directions for moving beyond current benchmarking practices, discovering important applications that we don't know exist, and involving communities in evaluation design through HCI-inspired participatory methodologies.

1 Introduction

Language model-based applications are growing in adoption across the world. To ensure they are adopted responsibly and effectively, an understanding of their cultural impacts and sensitivities is important. Cultural misalignments in AI can perpetuate stereotypes, marginalize underrepresented voices, and fail to address the needs of diverse user communities (Blodgett et al., 2020). In response, the NLP and ML communities have begun to focus on culturally-aligned NLP, a subfield that aims to develop and evaluate systems capable of understanding and appropriately applying cultural knowledge in context (Adilazuarda et al., 2024; Liu et al., 2024; Zhou et al., 2025). The overarching goal is to create NLP systems (Bhatt and Diaz, 2024) that can effectively respond to and operate within varied cultural settings. In this paper, we concentrate specifically on evaluation, as it increasingly shapes the direction of LLM development and deployment across diverse cultural contexts.

A key challenge, however, is that any decision in the evaluation pipeline—no matter how technical or routine—can carry cultural assumptions or consequences. For example, the tasks selected for evaluation often reflect the developers' cultural context, which may not align with the needs of users from different backgrounds (Hershcovich et al., 2022). Metrics assumed to be universal, such as what counts as "well-structured" writing, can vary significantly across cultures. Even expectations around interaction style and communication can differ (Folk et al., 2025; Ge et al., 2024), affecting how users perceive model outputs.

Despite this, the community often overlooks these *cultural contingencies*, focusing attention only on the most obvious or explicit cultural questions (those labeled as "cultural tasks" or "multilingual settings"). As a result, most current evaluation practices reduce culture to static facts, trivia, or proxies like nationality—primarily testing models through isolated factual questions or their performance on culturally-cued prompts (Zhou et al., 2025; Mukherjee et al., 2024). While knowledge of cultural facts is important, it fails to recognize the cultural contingencies embedded in seemingly "neutral" evaluation choices.

In this position paper, we argue that **every evaluative choice should be examined for culturally contingent considerations**, not just those in explicitly cultural domains. We argue for a shift toward ***intentionally cultural evaluation***: a systematic approach that foregrounds cultural context throughout the evaluation process. By this, we mean making the cultural context of every evaluative decision

explicit and deliberate, rather than leaving cultural influences implicit or accidental.

We therefore systematically distinguish and discuss three key aspects of evaluation: (1) *what* is being evaluated, (2) *how* evaluation is carried out, and (3) *under what circumstances* evaluation decisions are made to illuminate where and how culturally contingent considerations arise in each. In doing so, we surface the cultural factors embedded in evaluation choices. We further highlight the importance of researcher positionality, noting that evaluation is not neutral: it is shaped by who defines the tasks and benchmarks, and by systemic pressures that often privilege English-centric or high-resource perspectives. These dynamics constrain the development of research agendas grounded in diverse local contexts, perpetuating inequities and limiting the inclusiveness of NLP research.

This work is structured as follows: Following this description of **how** any evaluative choice could have a cultural component, we discuss **why** only the obvious choices around "cultural tasks" or "multilingual settings" get considered. We then systematically characterize the ways we can recognize a cultually-contingent decision, including what the evaluative target is (section 2), how evaluative goals are framed (section 3), and in what circumstances (section 4) the desideratum is defined. We then lay out a research practice that realizes these goals with situated researchers (section 5) and the implications of this work (section 6).

Contributions. We find most evaluations reflect a narrow set of cultural assumptions, shaped by those who define the tasks and metrics. The design of "what" gets evaluated is frequently informed by dominant Anglocentric perspectives, reifying specific knowledge types and communicative norms while marginalizing others. We show that standard computational practices, such as static reference examples or aggregate metrics, are poorly equipped to assess culturally grounded variation, and argue for reimagining these methods to support more flexible, context-sensitive judgments of model quality. Crucially, we contend that culture is not just content but interactional: it emerges dynamically through language use, social roles, and situational expectations. As such, evaluating only static outputs misses key aspects of cultural behavior.

Finally, we call for greater reflection on the positionality of those evaluating. Evaluation of cultural competence in NLP is not neutral — it is shaped by the positionality of researchers and by systemic biases embedded in the broader AI/ML ecosystem. Researchers from lower-resource or non-Anglophone contexts often face pressure to conform to English-centric benchmarks to gain visibility, placing additional burdens on their work and constraining the development of research agendas grounded in local cultural contexts. This marginalization limits the diversity of perspectives represented in NLP and reinforces existing inequities.

We propose building blocks for a systematic approach that centers cultural context throughout the evaluation process. Futher, we suggest implications for moving beyond decontexutalized methodologies toward more situated and culturally responsive methods, surfacing "unknown unknowns," and co-constructing evaluation practices with affected communities. We ground our suggestions using findings from HCI studies. In doing so, we support a broader shift in NLP evaluation toward thick evaluation (Qadri et al., 2025)—an approach that prioritizes context-sensitive, community-aligned assessments of AI systems.

2 What to evaluate

To move toward culturally intentional evaluation, we must ask: *What tasks contain important, culturally contingent considerations?* Current evaluations suffer from (a) overly narrow conceptions of 'cultural' tasks and (b) externally imposed definitions of relevance, thus failing to capture true cultural competence in real-world contexts.

2.1 Narrow Definitions of 'Cultural Tasks'

The reductive "culture as trivia" approach neglects ways that complex culturally-contingent interaction patterns or system behavior expectations should be integrated into evaluation design.

Current evaluation practices lack recognition of the fact that **even if a task is not framed as "evaluation of cultural alignment", it may still be culturally contingent and non-universal**. Widely-used benchmarks such as MMLU (Hendrycks et al., 2021) and HELM (Liang et al., 2023), designed to assess foundational LLM performance, are often presented as culturally neutral and universally applicable. However, recent analyses demonstrate that performance on these benchmarks in fact requires considerable culturally contingent knowledge and assumptions. Singh et al. (2025) found that 28% of MMLU dataset requires culturally-sensitive knowl-

edge to answer correctly, demonstrating that accounting for cultural context can change system rankings.

This recognition compels us to expand our understanding of "cultural tasks" to include tasks whose successful execution depends on cultural context, knowledge, norms, and user expectations. **Rather than narrowly defining cultural elements and then gradually designing complex applications, the cultural evaluation community should prioritize capturing real-world user scenarios that embed cultural competence.** This broader lens also helps surface the "unknown unknowns" of culturally aligned model behavior—dimensions of interaction or expectation that remain invisible under narrow definitions, yet become salient when real-world cultural variation is considered.

This shift will incentivize models capable of the "deep" cultural adaptations described by Resnicow et al. (1999)—adaptations requiring understanding of underlying social norms and contexts beyond surface-level knowledge. A seemingly universal task like email writing demands nuanced cultural competence for effective cross-cultural communication. For example, in Korean professional settings, emails to hierarchical superiors typically begin with weather remarks and well-wishes to establish appropriate relational context, illustrating how cultural competence extends beyond isolated knowledge to contextual application.

2.2 Task selection reflects Western priorities

Cultural NLP evaluation often embeds implicit biases in determining which tasks are deemed relevant or valuable. These biases extend beyond model outputs to the earlier and more foundational layer of task selection. While prior work on cultural dominance has largely focused on disparities in model responses across cultures (Wang et al., 2024), we highlight the overlooked bias in deciding what tasks to evaluate in the first place. As Hershcovich et al. (2022) argue through the concept of "Aboutness," cultural context shapes what is considered important. Yet, current benchmarks often treat tasks as culturally neutral, applying them uniformly without regard for differing communicative goals, linguistic norms, or practical needs.

NLP evaluations routinely prioritize tasks rooted in English-speaking, Western contexts—often by adapting existing English benchmarks and framing non-English efforts as merely closing a "performance gap." This bias is reinforced when task selection is based on user interaction data (Bhatt and Diaz, 2024), which overwhelmingly reflects usage patterns in the U.S. and other Western nations (Zhao et al., 2024).

This narrow framing has significant consequences. Tasks meaningful primarily in Western contexts are often overrepresented. Western-prevalent tasks—like sentiment analysis of beer reviews (Ji et al., 2020)—may be irrelevant in contexts where alcohol is prohibited. Even widely-used tasks, such as long-form news summarization, may hold less value in cultures where news is already concise (e.g., Korea). Conversely, and more critically, **tasks crucial in other cultural contexts receive disproportionately less attention**. English text refinement for non-English speakers—a vital need for millions globally—is one such example, often overlooked in mainstream evaluation.

Furthermore, for a specific task, the *topics* on which users engage with LLMs vary significantly across cultural contexts. For example, Tamkin et al. (2024) analyzed multilingual user interactions with Claude.ai and found that topics such as economic and social issues, or culturally specific content like anime, were more prevalent in non-English conversations compared to English ones. Similarly, Kirk et al. (2024) found that identity factors such as race, region, and gender have predictive power on the kinds of topics users choose to discuss with LLMs, even when conversation framing is controlled. This extends to handling sensitive content: alignment for religious issues in a Western context (often focusing on Christianity) differs vastly from needs in India (where Christianity is a minority religion) (Bhatt et al., 2022), and innocuous gestures in one culture can be offensive elsewhere (Yerukola et al., 2025).

To address these biases in task selection, we need evaluation frameworks that actively incorporate diverse cultural perspectives in task selection and design. This requires moving beyond simply adapting Western benchmarks toward building evaluation methodologies that emerge from and reflect the authentic needs and priorities of diverse user communities. Only then can we develop LLMs that truly serve the global population rather than inadvertently reinforcing existing power imbalances in digital communication.

3 How to evaluate

Having established *what* to evaluate, we now address *how* to evaluate these diverse desiderata. Sometimes, *what* (section 2) can be feasibly evaluated, is constrained by limitations in the *how*.

A major challenge in large-scale cultural evaluation is "values pluralism," the existence of diverse, sometimes fundamentally irreconcilable perspectives (Berlin, 1969). As datasets grow to encompass more diverse sub-groups, core differences in perspective can render the aggregation across samples less meaningfully representative of a coherent "culture" as a whole (Diaz and Madaio, 2024). This pluralism creates significant challenges for developing equitable and representative evaluation methods.

3.1 Reference examples are limiting

This challenge of values pluralism manifests even in the simplest domains and evaluation metrics, such as multiple-choice evaluation. For instance, *value alignment* research, which aims to move beyond evaluating culture as mere trivia, often captures *culture as perspective* using demonstrative examples of culturally variable preferences on personality, political, and opinion questions, typically through questionnaires. For example, AlKhamissi et al. (2024) frame cultural alignment in LMs as the distributional similarity of LM answers to national populations on surveys like the World Values Survey (Inglehart et al., 2000).

While such work also seeks to adapt model affinity using interventions like persona-based prompting (Li et al., 2024b), the reliance on multiple-choice opinion outputs is problematic. These outputs from LMs are notoriously noisy; Khan et al. (2025) show how variations of opinions along value scales vary just as much under semantically-irrelevant stylistic modifications of the prompt as they do under cultural conditioning. Further, *even when LMs authentically represent a distinct cultural perspective in their outputs, these questionnaire-based methods may miss them.* This calls into question the fundamental construct validity of questionnaire-based evaluations (O'Leary-Kelly and Vokurka, 1998; Davis, 2023).

Static sets of exemplars can be problematic with more sophisticated metrics, too. Rich, context-dependent trained metrics can vary in unpredictable and task-dependent ways, with system scores that are completely contradictory with the same metric across different tasks. For example, Lum et al. (2024) note how simple "trick tests" of gender bias are not only not predictive of performance within a real-world task—such as generating English learning lessons and writing bedtime stories—but scores on these unrelated real-world tasks couldn't predict each other.

At the same time, reference-based evaluation, the most common approach, has important limitations when applied to cultural assessment. Many culturally problematic behaviors—such as blind spots, stereotypical responses, or the severity of inappropriate outputs—cannot be adequately captured by comparing model outputs to fixed reference answers. Reference-based metrics encode only what is predefined as "correct," making it difficult to detect implicit biases, undesirable defaults, or consistent failures across related inputs (Saxon et al., 2024). For example, Myung et al. (2024) show that models frequently default to a narrow cultural artifact (e.g., repeatedly mentioning "Seblak" for West Java queries), a pattern that standard evaluations tend to overlook.

3.2 Quality notions are culturally contingent

The primary manifestation of values pluralism in evaluation is that what constitutes "good" behavior or desirable performance in an LM is itself culturally contingent and inherently subjective. LM evaluation often seeks to assess "good" outputs, but there is no objective "good" when preferences are diverse and deeply rooted in cultural contexts.

Consider, for example, what patterns in responses to opinion questions make them distinctly American? Johnson et al. (2022) discuss how a propensity of ChatGPT to frame discussions of gun control legislation around individual liberties is a predominantly American position. However, this stance is neither uniquely nor comprehensively American, since many Americans prioritize public safety. Relying on a single viewpoint misses internal diversity; robust evaluation should assess an LM's ability to articulate multiple perspectives *within* and *between* societies.

Defining critical concepts for evaluation is also problematic. Lee et al. (2024) found significant disagreement on what constitutes hate speech even among English-speaking countries. If consensus is elusive even within the English language, universal classifiers or metrics for culturally-embedded tasks become questionable. This suggests that a bespoke metric tuned to the preferences of each culture being tested might be necessary.

Furthermore, the interpretation and use of evaluation scales are culturally variable. Studies show Chinese and Japanese raters prefer midpoint satisfaction scores, unlike Americans who readily provide high scores (Lee et al., 2002). This can be partly explained by underlying cultural values, as individualism, for instance, correlates with less midpoint bias on rating scales, irrespective of the question's topic (Chen et al., 1995). This "extreme response style" (Chun et al., 1974) impacts online ratings across various domains and cultures (Barbro et al., 2020; Alanezi et al., 2022), inevitably influencing LM training and evaluation via human preference feedback.

Culturally variable preferences extend to nuanced desiderata like writing styles. Western readers often prefer concise, linear writing over dialectical styles sometimes favored in East Asia (Kaplan, 1966; Shahid et al., 2024), with variations even within the Anglosphere (Oprea and Magdy, 2020). Complex qualitative desiderata like naturalness, engagingness (Zhong et al., 2022), or likeability (Liu et al., 2023) are culturally variable and hard to transfer across languages, especially from WEIRD contexts. Naive transfer risks unfairly penalizing outputs aligned with local, non-WEIRD norms.

3.3 Standard metrics are improperly situated

The limitations of standard evaluation metrics—such as accuracy, F1, or ROUGE—are increasingly apparent in the context of cultural alignment. These metrics are decontextualized by design and assume a single correct or optimal output, an assumption fundamentally misaligned with culturally situated interactions that require a plurality of valid responses (Qadri et al., 2025). Evaluating against a single reference point not only obscures cultural complexity but also risks penalizing culturally attuned variations that fall outside dominant norms. Even culturally-specific metrics, if not properly situated, tend to capture preferences in isolation while overlooking the broader spectrum of context-dependent user behavior, interaction styles, and emergent practices that constitute genuine culture. Addressing these limitations calls for a fundamental shift in our evaluation paradigms—moving beyond incremental tweaks toward deeper, structural changes, such as the pluralistic frameworks proposed by Sorensen et al. (2024).

Beyond investing in diverse representative *samples*, we need diverse representative *metrics* in order to fully model the diverse needs of culturally diverse users. Metrics are able to express many desiderata in a way that samples alone cannot.

4 In what circumstances to evaluate

NLP evaluation often misses the inherent cultural contingency of human-AI interaction patterns.

4.1 Culture is more than language

Language often serves as a proxy for cultural variation (Wang et al., 2024; Li et al., 2024a). Typical experimental setups involve posing identical culturally-grounded queries or tasks in multiple languages and examining if model performance remains consistent (Myung et al., 2024; Shafayat et al., 2024; Jin et al., 2024). As demonstrated in these studies, language selection plays a significant role in evaluation-tracked performance. However, disparity in performance by language exists even on ostensibly non-cultural tasks such as common concept image generation (Saxon and Wang, 2023). Although verifying consistent performance across languages is important, simple performance matching over aligned translated inputs misses critical cultural information.

Language differences effect subtle aspects of outputs such as information density, factual completeness, and nuance. A recent study (Shafayat et al., 2024) evaluating the factuality of model outputs for culturally-oriented questions found that changing the language altered both the number and informational density of generated facts. This highlights a significant limitation of current "objective" metrics, which often overlook these nuanced linguistic and informational differences. Qualitative and comparative methods are essential to accurately capture these subtle yet important changes.

Second, language usage is fundamentally shaped by social and cultural factors, affecting both linguistic form and communicative style. Evaluations currently overlook how effectively models handle these culturally embedded linguistic norms (Hovy and Yang, 2021; Hershcovich et al., 2022). For example, Korean features a complex honorific system reflecting social hierarchies (Brown, 2015). Evaluating model outputs in Korean contexts thus requires assessing not just informational correctness but also whether responses adhere to culturally appropriate norms of politeness and formality, considerations less prominent in languages like English. Similarly, an LLM responding with the Korean phrase "좋은 질문이야!" (Good question!) might

User reaction to ChatGPT's informal Korean output

"When you speak informally to ChatGPT, it now replies informally too, haha. I used to think of ChatGPT as my assistant, but when it suddenly spoke informally, I felt a bit offended, lol. I guess now I need to start thinking of it as more of a friend." [a]

[a] Originally posted in Korean on a public online forum. Source: https://www.clien.net/service/board/park/18463114

Figure 1: A Korean user reflects on ChatGPT's unexpected use of informal speech, noting a shift in their perceived social relationship with the model. This illustrates the importance of speech-level appropriateness in culturally sensitive language generation.

be grammatically correct, but may feel overly direct or unnatural, mirroring English conversational patterns rather than typical Korean interactional styles. Such mismatches clearly indicate failures of cultural alignment, even if the task's primary goal (e.g., answering a question) is met. Current evaluations typically restrict consideration of linguistic nuances to tasks like translation, neglecting them in instruction-following or question-answering scenarios where task-specific metrics dominate.

Intentionally cultural evaluation can be achieved by moving beyond narrow performance metrics to consider: (1) nuanced linguistic and informational differences that emerge across languages, and (2) the critical influence of social and cultural contexts on language form and use. These concerns extend beyond isolated linguistic choices and individual utterances, pointing toward broader, culturally-shaped interaction patterns. This brings us to the next evaluation circumstance: interaction style.

4.2 Interaction Style should be evaluated.

Since the introduction of LLMs, especially ChatGPT and other web-based agents, conversational interactions have rapidly become the "default" interaction style for human-LLM engagement. This shift towards conversational, general-purpose chatbot models has fundamentally altered the landscape of evaluation, necessitating a more nuanced understanding of how interaction patterns themselves are culturally situated. Therefore, to evaluate LLMs for cultural alignment, we need to consider environmental and cultural differences not only in language but also at the interaction level. However, current cultural NLP research largely overlooks these nuanced interactional dynamics.

Cultural dynamics profoundly shape these human-AI interactions. **Users from different backgrounds vary in their input styles**, such as prompt directness across high and low-context cultures (Haoyue and Cho, 2024). Misinterpreting these culturally-specific instruction cues can cause LLMs to misunderstand intent and reduce conversation quality (Chaves and Gerosa, 2021), creating disadvantages, especially in multi-turn interactions. Concurrently, **users hold culturally grounded expectations for the AI's behavior and role**, including politeness—as seen with Korean users seeking workarounds to ensure models maintain formality Figure 1—and the desired relational nature of the interaction, with some East Asian users seeking more rapport than typically task-focused Western users (Folk et al., 2025; Ge et al., 2024). How LLM manages these interactional styles significantly impacts user satisfaction and perceived quality.

However, **the way these cultural interaction style differences affect model performance is a major gap in current evaluation frameworks.** While many studies report performance variations across languages (Myung et al., 2024; Shafayat et al., 2024; Jin et al., 2024), the specific impact of culturally diverse interaction patterns remains largely unexplored. We lack comprehensive datasets representing diverse human-model interactions across cultures. Despite efforts like LMSYS (Zheng et al., 2023), Chatbot Arena (Chiang et al., 2024), and WildChat (Zhao et al., 2024) collect "in-the-wild" interactions of users, these collections remain dominated by Western perspectives (53.7% of WildChat logs are English queries, with 21.6% of IP addresses from the United States and more than 40% from Western countries).

This research gap is particularly concerning given that models demonstrate high sensitivity to prompt structure and phrasing (Dominguez-Olmedo et al., 2024; Zhu et al., 2023; Pezeshkpour and Hruschka, 2024; Zhuo et al., 2024; Salinas and Morstatter, 2024). Users whose natural communication patterns diverge from those dominant in training data may face consistent disadvantages in model performance and responsiveness, effectively experiencing a "cultural prompt engineering tax" that others do not. Current approaches often place adaptation burdens on users rather than models (e.g., "if the model isn't performing well, you're not prompting it correctly.") This expectation, that users should conform to the model's

preferred communication patterns rather than vice versa, demands critical rethinking.

Such cultural misalignments can have severe impacts, for example user alienation, trust erosion, and system abandonment by users from specific cultural backgrounds (Adilazuarda et al., 2024). This can create a self-reinforcing cycle: models become increasingly optimized for the cultural interaction patterns of those who continue to use them, while simultaneously becoming less accessible to others. Moreover, this dynamic risks what Jones et al. (2025) describe as "hegemonic interactional norms," where models trained predominantly on English-language data from Western contexts implicitly impose particular communication patterns on users from different backgrounds.

Therefore, evaluation frameworks must evolve to account for culturally diverse interaction styles. This means asking not only whether a model performs well overall, but whether it does so equitably across different cultural patterns of engagement. Addressing this requires: (1) collecting data on how users from diverse backgrounds naturally interact with LLMs—including turn-taking, request styles, and conversational repair; (2) analyzing how cultural expectations shape perceptions of response quality; and (3) developing interaction-focused metrics that assess a model's adaptability, identifying and mitigating performance disparities across interaction styles.

5 Situated Researchers

Beyond the technical questions of what and how to evaluate cultural alignment lies a deeper set of socio-political questions concerning who performs this evaluation and within what kind of research ecosystem. The very practice of culturally-aligned evaluation is shaped by the positionality of researchers and the systemic biases embedded within the broader AI/ML community.

Researchers from non-Anglophone cultures face an implicit pressure: to gain visibility and legitimacy, their work must often first engage with English-centric tasks and benchmarks. Even when the researchers have a specific issue that they want to deal with, that they found in their linguistic context, this reality imposes an extra layer of labor to either 91) do parallel research (e.g., building two sets of dataset; one of their own the other English) or (2) first start with English to establish as a legitimate task and then move on to their own languages.

However, language cannot be separated from culture. Just translating the problem at hand to English, or finding a similar problem in English may not be sufficiently useful for the actual problem they first started out. For example, relationship inference based on dialogues between two Koreans may be uniquely difficult due to the linguistic characteristics of Korean, such as frequent omission of the Subject, or Terms of Address that have unique social connotations etc., while it is less of a problem in other languages. This kind of research ecosystem might actually be the real bottleneck of developing culturally aligned LLMs. It potentially hinders the development of research agendas truly grounded in diverse local contexts.

The field's reliance on standardized benchmarks (e.g., GLUE, BigBench, MMLU) to characterize model capability and research value reinforces a subtle form of epistemic injustice. Knowledge systems and problem formulations rooted in non-dominant contexts are often treated as peripheral—framed as "extensions" like "benchmarks for X language"—rather than valued on their own terms. This reflects an implicit belief in the authority of dominant research centers to define legitimate knowledge, pressuring global researchers to conform by translating or adapting to English-centric benchmarks. In doing so, the current system risks marginalizing diverse epistemologies while treating English not merely as a lingua franca, but as the default arbiter of relevance and validity.

A meaningful shift in NLP evaluation thus requires more than new datasets or metrics. Evaluative choices—what to measure, how, and why—are shaped by positionalities, not objective truths. Focusing on simple trivia to characterize culture, while treating all "non-cultural tasks" as universal hides bias behind a false veneer of objectivity. We need to recognize our positionality and seek out culturally-contingent aspects to all evaluation domains, and embrace the inherently cultured nature of LM use. Only through this kind of multi-layered reflection can we hope to build NLP systems that are not only culturally meaningful but also globally inclusive. One approach is to ensure that dataset construction and evaluation criteria are informed by the specific social, linguistic, and cultural contexts in which they are developed. By embedding these contextual considerations into the design of evaluation protocols, we move toward developing NLP systems that are not only more culturally attuned but also more inclusive on a global scale.

6 Implications and Future Directions

Beyond decontexualized measures. While existing benchmarks serve as useful tools for comparing models' general abilities (section 2), they often fall short in evaluating how models perform in real-world, culturally situated contexts. Inspired by behavioral testing approaches like CheckList (Ribeiro et al., 2020), which systematically probe linguistic capabilities through targeted test cases, we propose extending existing "universal" benchmarks with explicit dimensions of cultural capability. By incorporating tests for "cultural alignment failures"—such as how models handle culturally specific communication norms, contextually appropriate responses, or regionally relevant content.

At the same time, as we discuss in subsection 3.3, reference-based evaluation has limitations when applied to cultural assessment because it cannot adequately capture undesired default behaviors or measure the severity of the "inappropriateness," not reflected in the reference. To address these gaps, we need benchmark designs that move beyond static references and instead take a more holistic view of model behavior. This includes explicitly assessing the acceptability and severity of cultural misalignments, and systematically surfacing patterns of bias or insensitivity. Only with such frameworks can we meaningfully assess—and ultimately improve—models' cultural understanding at scale.

Discovering "Unknown Unknowns" As discussed in Section 2, a major challenge in evaluation is surfacing "unknown unknowns"—culturally meaningful tasks, interaction behaviors, and preferences that we currently do not know exist. To uncover these gaps, we need richer data on real user interactions, especially from underrepresented cultures. While resources like WildChat (Zhao et al., 2024), LMSys (Zheng et al., 2023), and Anthropic's Clio project (Tamkin et al., 2024) provide useful insights, current datasets remain limited in both cultural coverage and openness. More open, diverse interaction data is urgently needed.

Addressing unknown unknowns also requires methodological support. Collaborating with HCI researchers can help; for example, interactive systems have been developed to visualize data gaps and guide human-in-the-loop data collection (Yeh et al., 2025). Beyond building technical tools, the field needs empirical studies on why NLP researchers overlook these gaps and what practical interventions can help. Understanding these barriers is key to building more reflective and culturally robust evaluation practices.

Toward Stakeholder-Centered Evaluation Design. To support more culturally responsive evaluation practices, we must begin by identifying and centering those most directly impacted by NLP systems (subsection 2.1). Following Smith et al. (2024), evaluation should involve stakeholders who can define appropriate behavior in context.

HCI research highlights the need for culturally sustaining practices that foreground community voices from the start (Anderson-Coto et al., 2024). Value-sensitive and participatory design approaches further warn against universalist assumptions, emphasizing that values and evaluation standards must be situated in specific cultural contexts (Friedman, 1996; Borning and Muller, 2012).

Recent research at the intersection of NLP and HCI, show that engaging stakeholders in the evaluation process can help surface overlooked dimensions of cultural representation, such as missingness or connotation (Qadri et al., 2025). Building on this, we advocate for frameworks that go beyond simply diversifying annotators: stakeholders should help define tasks, criteria, and standards for evaluation through collaborative processes. This participatory approach better aligns NLP evaluation with the real needs and values of affected communities.

7 Conclusion

We have argued that evaluation in language technology is never culturally neutral, and that every choice—explicit or implicit—carries cultural consequences. Our analysis shows that conventional evaluation practices, from task and metric selection to benchmarking standards, often obscure or marginalize diverse cultural realities. To move beyond these limitations, we advocate for **culturally intentional evaluation**: an approach that makes cultural context visible, explicit, and central at every stage of the evaluation pipeline. By centering positionality, engaging with affected communities, and embracing context-sensitive, "thick" evaluation practices, the NLP community can develop more equitable, representative, and impactful language technologies. We hope this work catalyzes further reflection and action, inviting researchers to critically reexamine and reimagine the cultural assumptions embedded in their evaluation practices, and to co-create more inclusive and responsive models for the world's linguistic diversity.

Limitations

While we advocate for a theory-driven and culturally intentional approach to evaluation in NLP, several limitations should be noted. First, this paper does not aim to be an exhaustive survey of all work in evaluations of cultural alignment or related fields. Readers seeking comprehensive overviews may refer to recent surveys such as Pawar et al. (2024) or Liu et al. (2024). Additionally, our primary focus is on the evaluation of LLMs, which means that broader issues in language technology and culture are not discussed in detail.

As a position paper, our aim is to provoke discussion and outline future research directions, rather than to offer comprehensive solutions or empirical evaluations. We encourage further work that operationalizes these principles in a broader range of cultural, linguistic, and technological settings.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Khaled Alanezi, Nuha Albadi, Omar Hammad, Maram Kurdi, and Shivakant Mishra. 2022. Understanding the impact of culture in assessing helpfulness of online reviews. *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 308–315.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating Cultural Alignment of Large Language Models. *Preprint*, arXiv:2402.13231.

Maria J. Anderson-Coto, Julie Salazar, John Louis-Strakes Lopez, R. Mishael Sedas, Fabio Campos, Andres S. Bustamante, and June Ahn. 2024. Towards culturally sustaining design: Centering community's voices for learning through participatory design. *International Journal of Child-Computer Interaction*, 39:100621.

Patrick A. Barbro, Susan M. Mudambi, and David Schuff and. 2020. Do country and culture influence online reviews? an analysis of a multinational retailer's country-specific sites. *Journal of International Consumer Marketing*, 32(1):1–14.

Isaiah Berlin. 1969. Four essays on liberty.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16055–16074, Miami, Florida, USA. Association for Computational Linguistics.

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.

Alan Borning and Michael Muller. 2012. Next steps for value sensitive design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1125–1134, New York, NY, USA. Association for Computing Machinery.

Lucien Brown. 2015. Honorifics and politeness. *The handbook of Korean linguistics*, pages 303–319.

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758.

Chuansheng Chen, Shin-ying Lee, and Harold W Stevenson. 1995. Response style and cross-cultural comparisons of rating scales among east asian and north american students. *Psychological Science*, 6(3):170–175.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Kl Chun, John B. Campbell, and Jong Hae Yoo. 1974. Extreme response style in cross-cultural research. *Journal of Cross-Cultural Psychology*, 5:465 – 480.

Ernest Davis. 2023. Benchmarks for Automated Commonsense Reasoning: A Survey. *ACM Comput. Surv.*, 56(4):81:1–81:41.

Fernando Diaz and Michael Madaio. 2024. Scaling Laws Do Not Scale. *Preprint*, arXiv:2307.03201.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37:45850–45878.

Dunigan P Folk, Chenxi Wu, and Steven J Heine. 2025. Cultural variation in attitudes toward social chatbots. *Journal of Cross-Cultural Psychology*, 56(3):219–239.

Batya Friedman. 1996. Value-sensitive design. *interactions*, 3(6):16–23.

Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How culture shapes what people want from ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Luna Luan Haoyue and Hichang Cho. 2024. Factors influencing intention to engage in human–chatbot interaction: examining user perceptions and context culture orientation. *Universal Access in the Information Society*, pages 1–14.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, and 1 others. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. 2000. World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*.

Yunjie Ji, Hao Liu, Bolei He, Xinyan Xiao, Hua Wu, and Yanhua Yu. 2020. Diversified multiple instance learning for document-level multi-aspect sentiment classification. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7012–7023.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, pages 2627–2638.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The Ghost in the Machine has an American accent: Value conflict in GPT-3. *arXiv preprint:2203.07785*.

Graham M Jones, Shai Satran, and Arvind Satyanarayan. 2025. Toward cultural interpretability: A linguistic anthropological framework for describing and evaluating large language models. *Big Data & Society*, 12(1):20539517241303118.

R. Kaplan. 1966. Cultural thought patterns in intercultural education. *Language Learning*, 16:1–20.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs. *Preprint*, arXiv:2503.08688.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Preprint*, arXiv:2404.16019.

Jerry W Lee, Patricia S. Jones, Yoshimitsu Mineyama, and Xinwei Esther Zhang. 2002. Cultural differences in responses to a likert scale. *Research in nursing & health*, 25 4:295–306.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. 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 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *ArXiv*, abs/2402.10946.

Victoria R Li, Yida Chen, and Naomi Saphra. 2024b. ChatGPT Doesn't Trust Chargers Fans: Guardrail Sensitivity in Context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6327–6345, Miami, Florida, USA. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.

Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2023. X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects. In *North American Chapter of the Association for Computational Linguistics*.

Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander D'Amour. 2024. Bias in language models: Beyond trick tests and toward ruted evaluation. *arXiv preprint arXiv:2402.12649*.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. *arXiv preprint arXiv:2406.11661*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Scott W O'Leary-Kelly and Robert J Vokurka. 1998. The empirical assessment of construct validity. *Journal of operations management*, 16(4):387–405.

Silviu Vlad Oprea and Walid Magdy. 2020. The effect of sociocultural variables on sarcasm communication online. *Proceedings of the ACM on Human-Computer Interaction*, 4:1 – 22.

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. *Preprint*, arXiv:2411.00860.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025. The case for" thick evaluations" of cultural representation in ai. *arXiv preprint arXiv:2503.19075*.

Ken Resnicow, Tom Baranowski, Jasjit S Ahluwalia, and Ronald L Braithwaite. 1999. Cultural sensitivity in public health: defined and demystified. *Ethnicity & disease*, 9(1):10–21.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *arXiv preprint arXiv:2401.03729*.

Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. 2024. Benchmarks as microscopes: A call for model metrology. *arXiv preprint arXiv:2407.16711*.

Michael Saxon and William Yang Wang. 2023. Multilingual conceptual coverage in text-to-image models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4831–4848.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-fact: Assessing factuality of multilingual llms using factscore. *arXiv preprint arXiv:2402.18045*.

Farhana Shahid, Maximilian Dittgen, Mor Naaman, and Aditya Vashistha. 2024. Examining human-ai collaboration for co-writing constructive comments online. *ArXiv*, abs/2411.03295.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, and 5 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Jessie J Smith, Aishwarya Satwani, Robin Burke, and Casey Fiesler. 2024. Recommend me? designing fairness metrics with providers. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2389–2399.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and 1 others. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, and 1 others. 2024. Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Catherine Yeh, Donghao Ren, Yannick Assogba, Dominik Moritz, and Fred Hohman. 2025. Exploring empty spaces: Human-in-the-loop data augmentation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Akhila Yerukola, Saadia Gabriel, Nanyun Peng, and Maarten Sap. 2025. Mind the gesture: Evaluating ai sensitivity to culturally offensive non-verbal gestures. *Preprint*, arXiv:2502.17710.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, and 1 others. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Peng Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Conference on Empirical Methods in Natural Language Processing*.

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

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, and 1 others. 2023. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, pages 57–68.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A Use of AI Assistant

We used ChatGPT web assistant (ChatGPT Pro) [1] to refine the writing of the manuscript.

[1] https://chatgpt.com/