Culture is Everywhere: A Call for Intentionally Cultural Evaluation

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

[◇]KAIST [†]Georgia Institute of Technology
[‡]University of Washington [*]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 yet 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 that can effectively respond to and operate within varied cultural settings (Bhatt and Diaz, 2024). 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 (Zhou et al., 2025) or their performance on culturally-cued prompts (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

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19156–19168
November 4-9, 2025

the cultural context of every evaluative decision explicit and deliberate, rather than leaving cultural influences implicit or accidental.

To challenge the current focus on only the most obvious choices like explicitly cultural tasks or multilingual settings, we systematically distinguish and discuss three key aspects of evaluation: (1) **what** is evaluated (section 2), (2) **how** it is evaluated (section 3), and (3) **in what circumstances** (section 4) the desideratum is defined. We also examine the critical role of **researcher positionality** in shaping these evaluative choices (section 5). Finally, we outline the broader implications of our proposed approach for NLP research and practice (section 6).

Contributions. Our work offers several key contributions. We characterize the cultural contingencies in evaluation and propose building blocks for intentionally cultural evaluation. 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 argue that culture in evaluation is not merely static content to be measured but is fundamentally tied to the circumstances of evaluation. We show how culture is both embedded in the very language of evaluation and enacted through culturally-contingent interactional patterns. As such, evaluating only static outputs misses key aspects of cultural alignment.

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.

Further, we suggest implications for moving beyond decontextualized 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"

Current evaluation practices suffer from a flawed dichotomy. On one hand, explicitly "cultural" tasks are often reduced to testing factual knowledge, a "culture as trivia" approach (Zhou et al., 2025) that neglects the complex interaction patterns and behavioral expectations core to cultural competence. On the other hand, 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. However, recent analyses reveal demonstrate that performance on these benchmarks in fact requires considerable culturally contingent knowledge and assumptions. Singh et al. (2025) found that 28% of MMLU requires culturally-sensitive knowledge to answer correctly, demonstrating that accounting for cultural context can change system rankings.

We argue that **cultural tasks should be expanded to include any task whose successful execution depends on cultural context, knowledge, norms, and user expectations.** In domains traditionally treated as "value"-oriented, such as social bias or moral reasoning, culture-adapted benchmarks (*e.g.*, Jeong et al. (2022), Jin et al. (2024a)) have long embedded linguistic and socio-cultural conventions specific to their cultural contexts, reflecting the well-recognized influence of culture on these tasks. In contrast, tasks that are typically categorized as general capabilities, such as email

writing or instruction-following, are often assumed to be culturally neutral and are evaluated without regard to contextual norms. Yet these tasks can be highly culturally contingent. For example, in Korean professional communication, emails to hierarchical superiors often begin with seasonal greetings or weather remarks, a practice rarely reflected in English-centric benchmarks. Performance on such subtle forms of localization remains largely unevaluated. While existing adaptations in value-oriented domains represent important progress, they overlook the broader set of tasks that also require cultural awareness to be executed appropriately.

2.2 Task selection reflects Western priorities

Cultural evaluation also embeds implicit biases in determining which tasks are considered relevant or valuable. 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. In practice, 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. First, tasks meaningful primarily in Western contexts are often overrepresented. For example, sentiment analysis of beer reviews is irrelevant where alcohol is prohibited (Ji et al., 2020), and long-form news summarization may hold less value in cultures where news is already concise. Second, and more critically, tasks crucial in other cultural contexts are underrepresented. English text refinement for non-native speakers—a vital need for millions globally—is one such example, often overlooked in mainstream evaluation.

Even the topics and categories underlying evaluation reflect Western assumptions. For example, ostensibly universal notions like fairness and harm have largely been operationalized through Western categories, such as skin tone or race, leaving harms rooted in non-Western cultural contexts effectively unmeasured (Qadri et al., 2025; Dammu et al., 2024).[1] Yet, research shows that in practice, users from different cultural contexts engage with LLMs around very different concerns. Tamkin et al. (2024) demonstrate that non-English conversations more often center on issues like economics, social concerns, or culturally specific content (*e.g.,*anime). Similarly, Kirk et al. (2024) find that identity factors such as race and region have predictive power on the kinds of topics users discuss with LLMs, even when conversation framing is controlled.

To address these biases in task selection, we must move beyond simply adapting Western benchmarks and instead **build evaluation methodologies that emerge from and reflect the authentic needs and priorities of diverse user communities.** A practical next step is to co-design evaluation tasks with these communities, ensuring they reflect real-world priorities and cultural norms. More broadly, identifying what we might call the "unknown unknowns"—culturally significant capabilities, interaction patterns, and potential concerns that remain invisible to outside researchers—is crucial to developing LLMs that serve the global population without reinforcing existing power imbalances.

3 How to evaluate

Having established *what* to evaluate, we now address *how* to evaluate these diverse desiderata. Sometimes, *what* 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 shapes how we can define and measure cultural alignment.

3.1 Definitions of "good" are culturally contingent

The primary manifestation of values pluralism in evaluation is that **what constitutes "good" behavior or desirable performance in a language model is itself culturally contingent and inherently subjective**. LM evaluation often seeks to as-

[1]For example, the GPT-4 Technical Report's system card frames safety challenges primarily around Western-centric categories (*e.g.*, race and gender).

sess "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. There are considerable populations of Americans who prioritize public safety over individual liberties in this debate and vice-versa. This illustrates a fundamental limitation: relying on a single viewpoint to model cultural representativeness will inevitably exclude significant internal diversity. A more robust, albeit expensive, approach would involve demonstrating a language model's ability to understand and articulate the range of diverse perspectives that exist *within* and *between* societies.

This challenge extends to defining concepts critical for evaluation. Lee et al. (2024) show that annotators across English-speaking countries (US, UK, Australia, Singapore, South Africa) disagree significantly on what constitutes hate speech. Given that even related cultural contexts cannot agree on such a critical concept, this raises fundamental questions about developing universal classifiers or metrics for hate speech to evaluate language models in culturally-embedded tasks. 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 themselves are subject to cultural variation, directly impacting how "goodness" is expressed and measured. (Lee et al., 2002) find that Chinese and Japanese raters prefer midpoint satisfaction scores, as opposed to Americans readily providing high scores. Individual endorsement of individualism also leads to less midpoint bias on questions that are otherwise unrelated to culture (Chen et al., 1995), suggesting that these culturally contingent values directly impact the meaning of scales for "non-cultural" tasks.

This phenomenon of "extreme response style" (Chun et al., 1974) between different cultures impacts a variety of domains, including online product and helpfulness ratings between Europe and Asia (Barbro et al., 2020) and differences in hotel and restaurant reviews in the Middle East and Anglosphere (Alanezi et al., 2022). Such culturally contingent values inevitably impact model performance as diverse human preference feedback, reflecting these varied response styles, is collected and used for training or evaluation.

Beyond explicit opinions and scale use, culturally variable preferences exist for more nuanced desiderata like writing styles. Western readers, for instance, often have a stronger preference for concise and linear writing over more dialectical writing styles sometimes favored in some East Asian countries (Kaplan, 1966; Shahid et al., 2024). Even within the Anglosphere, variations in national cultures drive differences in online communication styles (Oprea and Magdy, 2020). Unlike simpler lexical or semantic similarity metrics (Zhang et al., 2019), more complex, qualitative desiderata such as naturalness, engagingness, and understandability (Zhong et al., 2022), or likeability and interestingness (Liu et al., 2023), are extremely culturally variable and difficult to transfer across languages. Transferring these complex desiderata across languages is particularly challenging, as researchers cannot readily build on work developed predominantly in English and Western, Educated, Industrialized, Rich, and Democratic (WEIRD) contexts. A naive transfer risks unfairly penalizing outputs that align well with expected local cultural norms but deviate from WEIRD ones.

Acknowledging this cultural contingency is not to suggest an uncritical acceptance of all cultural norms. Rather, it is a necessary step to move past the current state of cultural ignorance and avoid the "perspectival homogenization" (Fazelpour and Fleisher, 2025) of models to a single dominant viewpoint.

3.2 Reference examples alone cannot express culture as practice

Given that definitions of "good" are culturally contingent and subject to pluralistic interpretations, evaluation paradigms that heavily rely on static reference examples or aggregated demonstrative samples inevitably struggle to capture this complexity. Such methods often implicitly assume a singular or dominant notion of correctness or preference, which, as discussed above, is an oversimplification.

This challenge manifests itself even in the simplest domains and evaluation metrics, such as in 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 language models as the distributional similarity of models' 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., 2024), the reliance on multiple-choice opinion outputs is problematic. These multiple-choice opinion outputs from language models 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 models 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 cannot predict each other. These limitations point to the need for alternative evaluation designs that foreground cultural variation directly, rather than relying solely on static exemplars or task-specific metrics.

3.3 Standard metrics are improperly situated

Beyond diversifying representative *samples*, we also need diverse representative *metrics*. Metrics can encode many desiderata in ways that samples alone cannot. However, conventional measures like accuracy or F1 assume a single correct answer and thus penalize culturally valid variation. Comparing model outputs to fixed "correct" references can miss problematic defaults, blind spots, or subtle stereotypes. For example, Myung et al. (2024) highlight models repeatedly defaulting to narrow cultural artifacts (*e.g.*, "Seblak" in West Java queries), a phenomenon invisible to standard qualitative metrics. This calls for a shift from singular metrics to multi-dimensional ones. For example, Qadri et al. (2025) show that evaluating cultural representation requires moving beyond factual accuracy to assess richer categories like the *missingness* of iconic elements or the *coherence* of cultural symbols. Developing metrics that capture such fine-grained, socially-grounded dimensions is essential for moving beyond a simple pass/fail judgment of cultural alignment. Addressing this gap requires moving beyond incremental tweaks toward pluralistic and structural alternatives (Sorensen et al., 2024).

4 In what circumstances to evaluate

The circumstances of an evaluation are not culturally neutral. Yet current practices often fail to account for the deep cultural contingency embedded in two fundamental dimensions: (a) the language in which evaluation is conducted and (b) the interactional context it assumes.

4.1 Language use is culturally situated

Language is not a neutral vehicle for universal meanings; it embeds and enacts culture. The same concept can be realized through different linguistic forms depending on context, shaped by social and cultural norms that affect both form and style. Yet current evaluations often overlook this (Hovy and Yang, 2021; Hershcovich et al., 2022), treating language as a simple variable rather than a cultural site. This oversight manifests in two common evaluation paradigms. First, in explicit "cultural tasks," language often serves as a flat proxy for a culture, with evaluations focusing on task-specific performance parity (*e.g.*, scoring knowledge about that culture) (Myung et al., 2024; Shafayat et al., 2024; Jin et al., 2024b). Second, in seemingly universal tasks evaluated in a multilingual setting, language is treated merely as a constraint. In both cases, the methodology is confined to measuring whether a model's task performance remains consistent across different languages, obscuring the rich cultural information encoded within linguistic choices themselves.

A direct consequence of this methodological oversight is the failure to evaluate whether models respect the social and cultural norms embedded in language. For example, Korean has a complex honorific system reflecting social hierarchies (Brown, 2015). Evaluation in such contexts must assess not only informational correctness but also whether responses adhere to culturally appropriate politeness

and formality—considerations less prominent in languages like English. A model response like "좋은 질문이야!" (*Good question!*) may be grammatically correct yet pragmatically awkward, reflecting English conversational norms rather than Korean interactional expectations. 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.

Furthermore, using language as a proxy for culture is problematic because the mapping is not one-to-one (Pawar et al., 2024; Lee et al., 2023); a single language can be used across many cultures, and a single culture can encompass multiple languages. This ambiguity challenges us to cautiously interpret performance gaps, recognizing that they can stem from a model's lack of cultural competence, the linguistic properties of the language itself, or an inseparable combination of both. For instance, Saxon and Wang (2023) demonstrate that performance disparities by language exist even on ostensibly non-cultural tasks such as common concept image generation. This shows that language itself is a powerful variable, making it difficult to isolate "cultural knowledge" as the sole factor behind performance differences in multilingual evaluations.

This critique does not dismiss the importance of performance parity across languages, which remains a crucial goal for multicultural equity. Rather, we argue that our approach to achieving it must be fundamentally expanded. A more robust evaluation framework would therefore address two distinct but related goals. First, it must move beyond informational correctness to assess pragmatic and cultural appropriateness, judging whether an utterance respects the social norms and communicative styles embedded in a language. Second, it must enrich the concept of "parity" itself, moving beyond task-specific metrics to include qualitative consistency. This involves using meta-metrics to track whether the granularity, amount, and quality of information remain stable across different linguistic contexts (Shafayat et al., 2024). Together, these two advancements would shift evaluation from merely verifying if a model works in a language to assessing how well it communicates within that language's cultural context.

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.

4.2 Interaction patterns 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 isolated, decontextualized statements but also **in the dynamics of interaction.** 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 im-

pacts 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., 2024b), 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). 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. This tax manifests at a fundamental level, with models often failing to reply consistently in the user's chosen language (Marchisio et al., 2024), forcing users to bear the extra cost of explicitly prompting "Reply in Language X". Moreover, models consistently show degraded comprehension on code-switched text, a natural communication pattern in many multilingual or non-English speaking communities (Mohamed et al., 2025). 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.

The field's reliance on standardized benchmarks (*e.g.*, GLUE (Wang et al., 2018), BigBench (Srivastava et al., 2023), MMLU (Hendrycks et al., 2021)) 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.

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. The pres-

sure arises because work on English is routinely treated as a universal baseline rather than as research on one specific language—a tendency that the *Bender Rule* directly critiques (Bender, 2011). From an evaluation-research standpoint, this reality imposes an extra layer of labor: scholars must either (1) conduct 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 superficially analogous English task often fails to address the phenomenon that originally motivated the research. For example, inferring social relationships from Korean dialogue is uniquely difficult due to the linguistic characteristics of Korean, such as frequent omission of the sentence subject, or Terms of Address that have unique social connotations, while it is less of a problem in other languages. In this sense, the global research ecosystem itself might actually be the primary bottleneck to developing genuinely *culturally-aligned* language models. It potentially hinders the development of research agendas truly grounded in diverse local contexts.

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 must (i) acknowledge our positionalities, (ii) seek out culturally-contingent aspects across *all* evaluation domains, and (iii) embed local social, linguistic, and cultural expertise into dataset construction and protocol design. 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.

6 Implications and Future Directions

Beyond decontextualized 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 section 3, reference-based evaluation has fundamental limitations for cultural assessment: it cannot capture undesired default behaviors or accommodate the culturally contingent definitions of "good" we identified in subsection 3.1. We need evaluation frameworks that accommodate multiple valid perspectives simultaneously rather than forcing consensus on a single metric. This requires benchmark designs that move beyond static references to holistically assess the acceptability and severity of cultural misalignments while systematically surfacing patterns of bias or insensitivity.

While employing LLM-as-a-judge systems is one promising method to evaluate such nuances without a single correct answer, as explored for cultural QA (Arora et al., 2025), we caution that this is not a silver bullet. Even when trained evaluators or LLM-as-a-judge systems are used, the cultural frame embedded in the judge itself risks re-inscribing a dominant perspective, creating epistemic circularity.

Discovering "Unknown Unknowns" Human-centric evaluation of LLMs often masks LLMs' unique, non-human errors; likewise, researchers evaluating cultures as outsiders risk overlooking problems they do not know exist. As discussed in section 2, surfacing these "unknown unknowns"—culturally meaningful tasks, interaction behaviors, and preferences currently invisible to us—is a major evaluation challenge.This includes both task-level gaps and interaction-level misalignments where cultural communication patterns affect model performance across user communities (section 4).

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 cultural coverage and openness.

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). The field also needs empirical studies on why researchers overlook these gaps and what interventions can help build more culturally robust evaluation practices section 5.

Toward Stakeholder-Centered Evaluation Design. To support more culturally responsive evaluation practices, we must identify and center those most directly impacted by LLMs (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 evaluation standards must be situated in specific cultural contexts (Friedman, 1996; Borning and Muller, 2012).

Recent research at the intersection of NLP and HCI shows that engaging stakeholders can surface overlooked dimensions of cultural representation (Qadri et al., 2025). We advocate for frameworks where stakeholders help define tasks, criteria, and evaluation standards through collaborative processes, moving beyond simply diversifying annotators. 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.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00509258 and No. RS-2024-00469482, Global AI Frontier Lab)

Contribution

Juhyun Oh initiated and led the project, structured and wrote the manuscript, and contributed to the position and all sections.
Inha Cha initiated the project, contributed to the overall conceptualization, integrated STS and HCI perspectives into the implications, drafted the initial introduction, and refined all sections of the manuscript.
Michael Saxon contributed to the position, wrote initial drafts for Section 3, and contributed to revisions in other sections.
Hyunseung Lim contributed to the initial drafts for Section 4, and provided feedback on other sections.
Shaily Bhatt contributed to the overall conceptualization, wrote first drafts of section 2, and provided feedback on other sections.
Alice Oh advised the project and helped revise the manuscript.

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.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11772–11817, Vienna, Austria. Association for Computational Linguistics.

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.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Isaiah Berlin. 1969. Four essays on liberty.

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.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. 2024. "they are uncultured": Unveiling covert harms and social threats in LLM generated conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20339–20369, Miami, Florida, USA. Association for Computational Linguistics.

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. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 341–357.

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.

Sina Fazelpour and Will Fleisher. 2025. The value of disagreement in ai design, evaluation, and alignment. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2138–2150.

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*.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024a. Kobbq: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024b. 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.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, pages 35–46.

Victoria R Li, Yida Chen, and Naomi Saphra. 2024. 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*.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

Amr Mohamed, Yang Zhang, Michalis Vazirgiannis, and Guokan Shang. 2025. Lost in the mix: Evaluating llm understanding of code-switched text. *arXiv preprint arXiv:2506.14012*.

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*.

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.

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*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

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*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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.

A Use of AI Assistant

We used ChatGPT web assistant (ChatGPT Pro) [2] to refine the writing of the manuscript.

[2] https://chatgpt.com/