Toward Culturally Grounded Natural Language Processing

Sina Bagheri Nezhad
Independent Researcher
Seattle, WA, USA
sina.bagherinezhad@gmail.com

Abstract

Multilingual NLP is often treated as a route to global inclusion, but linguistic coverage and cultural competence frequently diverge. This paper synthesizes over 50 papers spanning multilingual performance inequality, cross-lingual transfer, culture-aware evaluation, cultural alignment, multimodal benchmarks, benchmark-design critique, and community-grounded data practices. Across this literature, training data coverage remains important, but outcomes are also shaped by tokenization, prompt language, translated benchmark design, culturally grounded supervision, modality, and who authors or validates evaluation data. We argue that culturally grounded NLP should move beyond treating languages as isolated rows in benchmark tables and instead model communicative ecologies: the institutions, scripts, domains, modalities, and communities through which language is used. We propose a layered evaluation and reporting agenda centered on representation audits, mixed elicitation, ecological validity, community validation, adaptation provenance, within-language variation, and maintenance of living cultural resources.

1 Introduction

Multilingual NLP is increasingly presented as a route to global inclusion, yet a growing body of work shows that linguistic coverage and cultural competence are not the same thing. Recent surveys note that culture in NLP is multi-scalar, dynamic, and typically operationalized only through partial proxies, while sociocultural critiques argue that many current evaluations treat culture as a static property rather than a situated social process (Adilazuarda et al., 2024; Liu et al., 2025a; Pawar et al., 2025; Zhou et al., 2025). In practice, a model can be fluent in many languages and still mis-handle local norms, misread culturally specific entities, or privilege a dominant worldview in text, multimodal, and interactional settings (Rystrøm et al., 2025; Naous et al., 2024; Naous and Xu, 2025; Belay et al., 2025; Nayak et al., 2024; Havaldar et al., 2025).

This concern matters because multilingual NLP has advanced rapidly. Work on global coverage and performance inequality shows that most languages remain weakly represented in resources, benchmarks, and deployed systems (Joshi et al., 2020; Blasi et al., 2022). Even when evaluation expands to 200+ languages, large disparities persist (Adelani et al., 2024; Team et al., 2022). A complementary line of research studies the drivers of cross-lingual transfer, highlighting pretraining distribution, lexical overlap, script, tokenization, and typology as important predictors (Philippy et al., 2023; Limisiewicz et al., 2023; Hämmerl et al., 2025; Bagheri Nezhad and Agrawal, 2024; Bagheri Nezhad et al., 2025). But these predictors alone do not explain whether models are aligned with the values, interactional norms, or situated knowledge of the communities they are meant to serve.

Recent culture-focused benchmarks and adaptation methods make that gap visible. On text evaluation, translated benchmarks can preserve source-culture assumptions, cultural ranking changes appear when evaluation is restricted to culturally sensitive subsets, and survey-style or value-oriented probes reveal substantial cultural dominance in default model behavior (Singh et al., 2025; Wang et al., 2024; Zhao et al., 2024; AlKhamissi et al., 2024; Masoud et al., 2025; Xu et al., 2025). In multimodal and interactional settings, models struggle with culturally grounded VQA, emotion understanding, metaphor interpretation, conversational adaptation, and region-specific question answering (Nayak et al., 2024; Schneider et al., 2025; Winata et al., 2025; Maji et al., 2025; Yang et al., 2025; Havaldar et al., 2025). At the same time, adaptation methods based on native preference data, culture-aware prompting, synthetic critique data,

Proceedings of the 4th Workshop on Cross-Cultural Considerations in NLP (C3NLP 2026), pages 119–131

and cultural learning show that these failures are not immutable, but they also reveal how dependent progress is on locally grounded supervision (Guo et al., 2025; Liu et al., 2025a; Feng et al., 2025).

We use *communicative ecologies* to refer to the social and technical conditions through which language is produced, interpreted, translated, annotated, evaluated, and deployed, including institutions, scripts, domains, modalities, and communities of practice.

This synthesis integrates evidence from over 50 papers spanning multilingual performance inequality, cross-lingual transfer, culture-aware evaluation, cultural alignment, multimodal benchmarking, benchmark-design critique, and community-grounded data practices. We prioritize recent ACL Anthology, TACL/CL, and C3NLP-adjacent work that either evaluates cultural behavior, explains multilingual performance variation, proposes culture-aware adaptation methods, or documents community-grounded data practices. We make three contributions. First, we connect the multilingual transfer literature to recent culture-oriented evaluation work, showing where the two have often been discussed separately. Second, we identify recurring empirical patterns across the literature: scale matters, but benchmark design, tokenizer behavior, prompt language, local supervision, and modality all materially affect cultural performance. Third, we propose a research agenda for culturally grounded NLP centered on communicative ecologies rather than isolated language labels. Figure 1 summarizes this shift.

Because the cited literature is broad and recent, our goal is not a formal meta-analysis or an exhaustive survey. Instead, we synthesize representative findings that jointly explain why multilinguality, cultural evaluation, and participatory localization should be studied together.

2 From Multilinguality to Cultural Competence

2.1 Coverage, transfer, and the limits of language-only evaluation

The multilingual literature has established two robust facts. First, language technology is unequally distributed. Resource concentration, benchmark availability, and deployment all remain heavily skewed toward a small subset of languages (Joshi et al., 2020; Blasi et al., 2022). Second, widening language coverage does not remove disparity by itself. When evaluation is extended to hundreds of languages and dialects, performance gaps remain large and often correlate with resource concentration or infrastructure advantages (Adelani et al., 2024; Team et al., 2022; Aji and Cohn, 2025).

Research on cross-lingual transfer helps explain part of this pattern. Reviews and empirical analyses point to pretraining data, lexical sharing, script, tokenization, and other similarity signals as important determinants of multilingual transfer (Philippy et al., 2023; Limisiewicz et al., 2023; Hämmerl et al., 2025; Bagheri Nezhad and Agrawal, 2024; Bagheri Nezhad et al., 2025). These findings are valuable, but they still mostly answer a language-centric question: *when does one language help another?* They say much less about whether the model understands local norms, values, or region-specific knowledge once it has transferred technically.

That distinction matters because token or script overlap can support transfer while still leaving culturally grounded behavior under-specified. A model may reuse familiar subwords and achieve respectable benchmark accuracy, yet still fail when asked to interpret a culturally loaded metaphor, a local emotional expression, a region-specific entity, or a social norm that is not well represented in pretraining data (Naous et al., 2024; Naous and Xu, 2025; Belay et al., 2025; Yang et al., 2025; Nyandwi et al., 2025). Multilinguality, then, is necessary but not sufficient for cultural competence.

2.2 Culture is usually operationalized through proxies

A second lesson from recent surveys is conceptual. "Culture" is rarely modeled directly; instead, it is approximated through countries, languages, prompt languages, values surveys, food, rituals, emotions, conversational style, or locally salient entities (Adilazuarda et al., 2024; Liu et al., 2025a; Pawar et al., 2025). This is unavoidable to some extent, but it also creates risk. A benchmark may appear "cross-cultural" while actually testing only one thin slice of culture, such as national-value questionnaires or culture-specific trivia.

Recent conceptual work argues that this is not just a measurement inconvenience but a theoretical problem. Zhou et al. (2025) show that many cultural NLP setups rely on static and often nationalized proxies, while Liu et al. (2025a) emphasize that culture spans within-human, between-human, and extra-human dimensions. In related work,

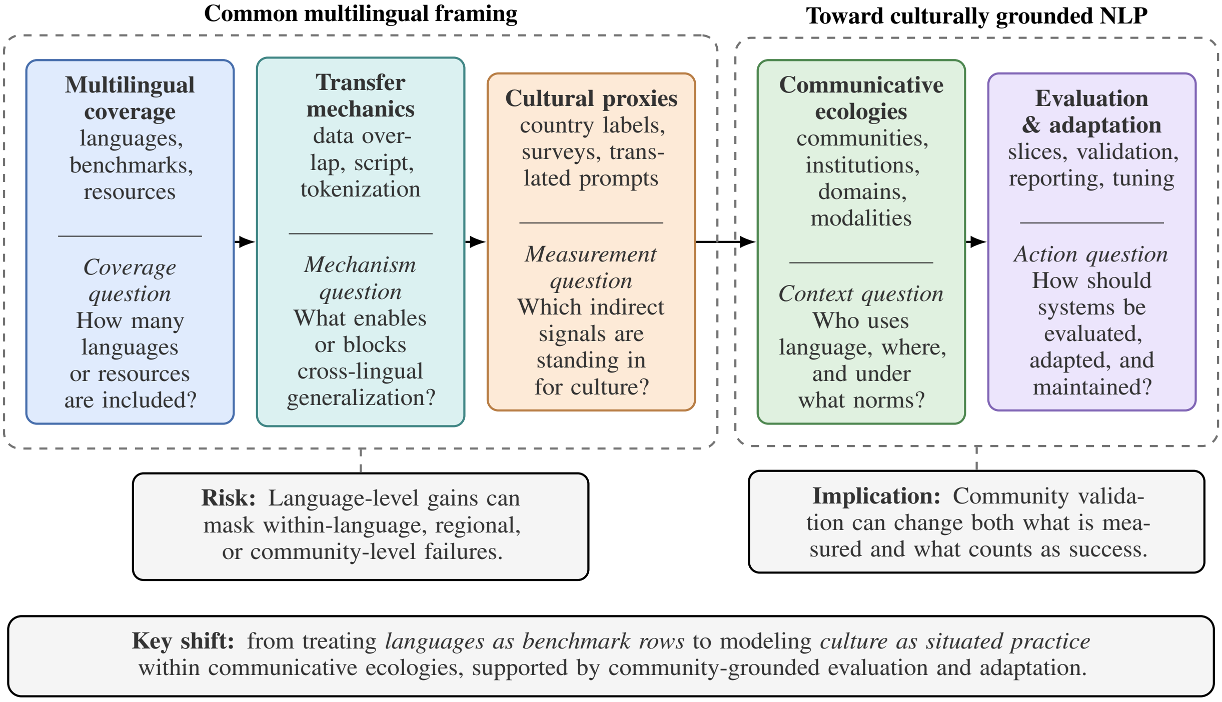

Figure 1: From multilingual coverage to culturally grounded NLP. The first three stages reflect common multilingual framing: increasing language coverage, explaining cross-lingual transfer, and approximating culture through indirect proxies. The paper argues for a further shift toward communicative ecologies, where language use is situated in communities, institutions, domains, and modalities, and where evaluation and adaptation are grounded in slice-based analysis, community validation, and explicit reporting.

alignment studies operationalize culture through prompt language, value surveys, or persona framing and demonstrate that model behavior can shift under these interventions (AlKhamissi et al., 2024; Masoud et al., 2025; Xu et al., 2025). The implication is not that such proxies are useless, but that they should be read as partial windows into broader communicative ecologies rather than as full representations of culture.

3 What Recent Benchmarks and Adaptation Work Show

3.1 Text and value-oriented evaluation

Recent text-centric evaluation work makes the multilingual-versus-multicultural gap especially clear. Global-MMLU shows that translated benchmarks can preserve linguistic and cultural assumptions from English source questions, and that model rankings may shift when evaluation is restricted to culturally sensitive subsets (Singh et al., 2025). Related benchmarks such as CDEval and WorldValuesBench probe cultural dimensions or value distributions more directly, revealing that models can produce systematically different alignments across countries and social domains (Wang et al., 2024; Zhao et al., 2024). CulturalBench and NunchiBench further show that locally grounded, human-written questions expose failures that are hard to detect with standard multilingual QA or classification tasks (Chiu et al., 2025; Kim and Lee, 2025).

This gap is also visible in work on cultural bias. In Arabic, Naous et al. (2024) show that multilingual and monolingual models can prefer Western-associated entities and generate culturally inappropriate associations. Naous and Xu (2025) trace part of this bias to the interaction between pretraining distributions, linguistic phenomena, and frequency-based tokenization, illustrating that technical design choices can have culturally asymmetric effects. More broadly, Rystrøm et al. (2025) find that multilingual ability does not reliably predict cultural alignment, reinforcing the argument that language coverage and cultural competence should be evaluated separately.

Text evaluation also reveals that alignment can move, but unevenly. Prompt language and persona prompting affect cultural outputs in both value-oriented and survey-based setups (AlKhamissi et al., 2024; Masoud et al., 2025). Self-pluralising or culture-joint tuning can improve alignment to

Theme	Representative literature	Main implication
Coverage and disparity	(Joshi et al., 2020; Blasi et al., 2022; Adelani et al., 2024; Team et al., 2022)	Aggregate multilingual progress still leaves many languages and communities poorly served.
Transfer factors	(Philippy et al., 2023; Limisiewicz et al., 2023; Hämmerl et al., 2025; Bagheri Nezhad and Agrawal, 2024; Bagheri Nezhad et al., 2025)	Pretraining, script, and tokenizer behavior shape transfer, but they do not by themselves measure cultural fitness.
Culture definitions and surveys	(Adilazuarda et al., 2024; Liu et al., 2025a; Pawar et al., 2025; Zhou et al., 2025)	"Culture" is multi-faceted and usually operationalized through incomplete proxies.
Text and value-oriented evaluation	(Wang et al., 2024; Zhao et al., 2024; Singh et al., 2025; Kim and Lee, 2025; Chiu et al., 2025; Rystrøm et al., 2025)	Model rankings and failure modes change on culturally sensitive, locally authored, or culturally situated subsets.
Alignment and adaptation	(AlKhamissi et al., 2024; Masoud et al., 2025; Xu et al., 2025; Guo et al., 2025; Liu et al., 2025a; Feng et al., 2025)	Cultural behavior can shift with prompts or targeted supervision, but gains depend on grounded data and task framing.
Multimodal, interactional, and local tasks	(Nayak et al., 2024; Schneider et al., 2025; Yang et al., 2025; Belay et al., 2025; Maji et al., 2025; Winata et al., 2025; Villa-Cueva et al., 2025; Nyandwi et al., 2025; Havaldar et al., 2025; Ochieng et al., 2025; Aji and Cohn, 2025; Ferawati et al., 2024; Pranida et al., 2025; Naous et al., 2024; Naous and Xu, 2025)	Local knowledge, interactional norms, region-specific entities, and within-language variation remain difficult across settings and modalities.

Table 1: Representative lines of work that jointly motivate a culture-grounded view of multilingual NLP.

multiple cultures without fully collapsing general performance, but the underlying problem remains open because cultures are internally diverse and task demands differ (Xu et al., 2025). The shared lesson is that cultural performance is not a fixed model property; it is partly a function of how the task is framed, authored, and prompted.

3.2 Multimodal, interactional, and local evaluation

Culture is not expressed only in text, and recent multimodal work makes this plain. CulturalVQA, GIMMICK, WorldCuisines, DRISHTIKON, and CaMMT all show that models struggle with geographically diverse food, dress, artifacts, rituals, and region-specific visual cues (Nayak et al., 2024; Schneider et al., 2025; Winata et al., 2025; Maji et al., 2025; Villa-Cueva et al., 2025). Work on multimodal metaphors and culturally grounded MLLMs reaches a similar conclusion: local entities and cultural references remain brittle, especially when they are rare in dominant training sources (Yang et al., 2025; Nyandwi et al., 2025).

Interactional and low-resource settings reveal additional failure modes. CULEMO shows that cross-cultural emotion understanding varies substantially across languages and resource conditions (Belay et al., 2025). Culturally-Aware Conversations argues that many existing benchmarks are misaligned with the actual conversational situations in which cultural adaptation matters, and demonstrates that even strong models struggle with stylistic sensitivity and subjective correctness (Havaldar et al., 2025). In low-resource real-world scenarios, qualitative assessment can surface weaknesses that automatic metrics miss, especially in code-mixed or socially contextualized tasks (Ochieng et al., 2025). Regionally focused benchmarks such as LORAXBENCH and culturally nuanced story-generation tasks for Javanese and Sundanese also show that cultural reasoning, formality, and local narrative competence are uneven even within one geopolitical region (Aji and Cohn, 2025; Pranida et al., 2025).

These findings also connect back to dataset construction. Synchronizing annotation guidelines across cultures is itself a methodological challenge, as shown by work on multilingual annotation design (Ferawati et al., 2024). In other words, cultural competence is not only something to test at the end of the pipeline; it is also shaped by how annotation, adjudication, and benchmark authorship are organized from the start.

3.3 Adaptation and alignment methods

Recent alignment work suggests that targeted intervention can help. Early cultural-alignment studies showed that model responses shift when prompted with a dominant local language or culture-specific persona information (AlKhamissi et al., 2024). More recent methods go further by introducing culture-aware training signals. CARE uses multilingual human preference data from native speakers to improve cultural awareness (Guo et al., 2025); CLCA uses simulated social interactions to adapt models toward target cultural values (Liu et al., 2025b); CulFiT synthesizes multilingual critiques to support fine-grained cultural training (Feng et al., 2025); and CultureSPA demonstrates that pluralistic culture alignment can be improved through self-generated culture-related supervision (Xu et al., 2025).

These methods are promising, but their success underscores a broader point: cultural competence does not arise automatically from multilingual scale. It often requires native raters, culture-specific task framing, or carefully designed supervision (Guo et al., 2025; Havaldar et al., 2025). The data and design burden is therefore part of the science, not just an implementation detail.

3.4 Evaluation paradigms, representation, and ecological validity

Recent work increasingly argues that the main bottleneck is not only model capability but also what current evaluations choose to treat as culture. Oh et al. (2025) describe a "trivia-centered" paradigm in which culture is reduced to static facts, survey answers, or decontextualized commonsense questions. Kabir et al. (2025) show that this matters empirically: open-ended assessment can yield different conclusions from closed multiple-choice evaluation, because it lets models express partial knowledge, uncertainty, or culturally plausible alternatives that a fixed answer set suppresses. Meanwhile, Wu et al. (2025) survey 105 cultural-alignment benchmarks and find that representation is heavily concentrated in majority populations, dominant languages, and national-level categories, with much less attention to subcultural variation or minority perspectives. Alkhamissi et al. (2026) extend the critique by arguing that many benchmarks rely on impoverished concepts of culture and would benefit from stronger anthropological grounding.

These measurement questions are tightly linked to task realism. Cultural failures often appear most clearly when a model must act in a situated environment rather than answer a standalone prompt. Qiu et al. (2025) show that LLM web agents have markedly weaker cultural and social awareness in live browsing settings than in simpler, non-agentic setups. Likewise, recent multimodal video benchmarks indicate that cultural competence depends on joint interpretation of scene context, gesture, speech, objects, and event structure, not just isolated text or static images (Shafique et al., 2025). Ecological validity is therefore not a downstream deployment issue; it is part of the evaluation construct itself.

Native-speaker and community-validated datasets sharpen the point further. DaKultur shows that automatically translated data are inadequate for evaluating Danish cultural awareness and that native-speaker data materially improve both human acceptance and the usefulness of automatic evaluation (Müller-Eberstein et al., 2025). Community-engaged resources such as HESEIA and SAFARI similarly demonstrate that co-design changes which stereotypes, harms, and everyday contexts are even represented in the benchmark (Ivetta et al., 2025; Verma et al., 2026). Representation, in other words, is not only a matter of *how many* languages or regions appear, but of *whose categories, harms, and communicative practices* define the task.

Finally, recent efforts to scale culture-related resources reveal a tension between breadth and grounding. CultureInstruct and newer work on scaling cultural resources suggest that larger multicultural instruction collections can improve performance on cultural benchmarks, yet they also raise familiar questions about provenance, normalization, and loss of local detail (Pham et al., 2025; Stepanyan et al., 2026). The research challenge is therefore not just to collect more cultural data, but to preserve enough social context that scaling does not erase the very phenomena the field claims to model.

4 Synthesis: Recurring Findings Across the Literature

4.1 Data quantity remains necessary but insufficient

Across the multilingual literature, training data coverage remains one of the strongest predictors of performance (Joshi et al., 2020; Blasi et al., 2022; Adelani et al., 2024; Bagheri Nezhad and Agrawal, 2024; Bagheri Nezhad et al., 2025). However, recent work shows that the path from coverage to performance is mediated by tokenizer allocation, literal and non-literal token alignability, prompt language, and the interaction between local linguistic phenomena and pretraining composition (Limisiewicz et al., 2023; Hämmerl et al., 2025; Naous and Xu, 2025). More data is therefore necessary, but it does not guarantee that a model will reason appropriately about local culture once deployed.

4.2 Benchmark design often imports external cultural assumptions

A second recurring pattern is that benchmark construction itself can flatten culture. Translation from English source tasks can import assumptions about curricula, answer spaces, or what counts as common knowledge (Singh et al., 2025). Survey-style probes can capture useful signals, but they also narrow culture to the specific dimensions encoded in the survey instrument (Wang et al., 2024; Zhao et al., 2024; Masoud et al., 2025). Even locally grounded benchmarks face design choices around question writing, annotation, and disagreement handling (Chiu et al., 2025; Kim and Lee, 2025; Havaldar et al., 2025; Ferawati et al., 2024). The implication is that culture-sensitive evaluation must report not only scores, but also the assumptions embedded in benchmark construction.

4.3 Local supervision and participation matter

A third pattern is methodological. When native speakers, culturally diverse raters, or locally grounded supervision are included, performance and diagnostic value tend to improve (Guo et al., 2025; Liu et al., 2025a; Feng et al., 2025; Havaldar et al., 2025). Conversely, low-resource studies repeatedly show that generic metrics or imported annotation schemes can obscure the very failures that matter in practice (Ochieng et al., 2025; Aji and Cohn, 2025). Participation is therefore not only an ethical add-on; it changes what the field can validly claim to have measured.

4.4 Culture is distributed across modalities and within-language variation

The broadest lesson from recent work is that culture is distributed. It appears in images, food, artifacts, emotional repertoires, social interaction, metaphors, politeness registers, and local entities, not only in standardized written text (Nayak et al., 2024; Schneider et al., 2025; Yang et al., 2025; Belay et al., 2025; Maji et al., 2025; Winata et al., 2025; Villa-Cueva et al., 2025; Nyandwi et al., 2025). It also appears within languages, through dialect, register, code-mixing, and local narrative conventions (Aji and Cohn, 2025; Pranida et al., 2025; Ochieng et al., 2025). This means that culture-sensitive NLP cannot be reduced to a one-language-one-culture mapping.

4.5 From language lists to communicative ecologies

Taken together, these findings suggest a shift in framing. Languages should not be treated as isolated rows in a benchmark spreadsheet, but as elements of communicative ecologies: institutions, scripts, education systems, media networks, annotation pipelines, and communities of practice. Some multilingual transfer studies already hint at this by showing that contextual covariates beyond raw data amount can help explain performance (Bagheri Nezhad et al., 2025). Culture-oriented work extends the point by showing that what matters is not only *which language* is present, but *how that language is socially situated* in training, evaluation, and deployment (Liu et al., 2025a; Zhou et al., 2025; Rystrøm et al., 2025).

5 Toward Better Evaluation Protocols

The literature now supports a practical conclusion: evaluating cultural competence should be treated as a layered protocol rather than a single benchmark score. Table 2 summarizes concrete shortcuts and stronger alternatives. The key idea is to separate questions that are often collapsed in current practice: who is represented, how the task elicits cultural knowledge, how realistic the evaluation setting is, who validates outputs, and what kinds of data changed the model.

Representation audits should accompany every benchmark. At minimum, papers should report whether items are translated or native-authored, which language varieties are included, whether prompts target dominant or minoritized groups, and how local validation was handled. This matters because country and language labels can hide major within-language gaps, as shown by DIALECTBENCH and low-resource narrative tasks (Faisal et al., 2024; Aji and Cohn, 2025; Ochieng et al., 2025). Benchmark audits further suggest that majority-focused scope is still the norm, so missing-group analysis should be treated as core evaluation rather than appendix material (Wu et al., 2025; Alkhamissi et al., 2026).

Elicitation should also be methodologically diverse. Closed-form surveys and multiple-choice QA remain useful for comparability, but they should be complemented with open generation,

Protocol layer	Common shortcut	Stronger practice	Representative literature
Representation audit	Country/language labels only	Report authorship, language variety, translation pipeline, and whether majority/minority or subcultural groups are represented	(Wu et al., 2025; Alkhamissi et al., 2026; Faisal et al., 2024)
Elicitation diversity	Multiple-choice or Likert only	Combine closed items with open generation, pairwise judgments, and qualitative error analysis	(Oh et al., 2025; Kabir et al., 2025; Havaldar et al., 2025)
Ecological validity	Static text-only QA	Add conversation, web-agent, image, video, and region-specific task slices	(Qiu et al., 2025; Shafique et al., 2025; Nayak et al., 2024)
Community validation	Expert-only or automatic scoring	Include native-speaker review, disagreement analysis, and participatory co-design	(Müller-Eberstein et al., 2025; Ivetta et al., 2025; Verma et al., 2026)
Adaptation reporting	"Culture-tuned" as a black box	Publish supervision provenance, target population, and trade-offs across groups and tasks	(Guo et al., 2025; Pham et al., 2025; Stepanyan et al., 2026)

Table 2: A practical protocol for evaluating cultural competence beyond language-only leaderboards.

pairwise preference judgments, disagreement analysis, and qualitative error coding. Recent work demonstrates that different elicitation styles surface different cultural behaviors and that static trivia-style formats are too narrow for many deployment settings (Oh et al., 2025; Kabir et al., 2025; Havaldar et al., 2025). Mixed protocols can better distinguish absence of knowledge, culturally plausible variation, and direct norm violation.

A third requirement is to evaluate situated use. This includes multimodal tasks, dialect and code-mixed inputs, region-specific entities, conversation, and agentic interaction. The point is not that every paper must evaluate every modality, but that claims about cultural competence should be limited to the settings actually tested. Results from CulturalVQA, WorldCuisines, video benchmarks, and web agents show that performance can degrade sharply when models must integrate local cues in more realistic contexts (Nayak et al., 2024; Winata et al., 2025; Shafique et al., 2025; Qiu et al., 2025).

Finally, adaptation papers should document the social provenance of their supervision. When models improve through native preference data, co-designed datasets, or large-scale culture-specific instructions, those data are part of the causal explanation and should be reported as such. Participatory work in low-resource MT and design-inspired NLP already offers a vocabulary for this, while newer datasets show how community involvement changes both task design and measured harms (Nekoto et al., 2020; Caselli et al., 2021; Guo et al., 2025; Ivetta et al., 2025; Verma et al., 2026). A culture-aware benchmark without provenance or stakeholder documentation may still be useful, but its claims should be correspondingly narrower.

A final implication is temporal maintenance. Cultural resources should be versioned, refreshed, and revalidated as social norms, salient entities, and public events change. Static benchmark release is especially brittle for culturally grounded tasks because local salience can shift faster than generic linguistic competence, and because new communities or subcultures may become visible only after deployment. Recent benchmark-audit and resource-scaling work suggests that ongoing collection and community review are necessary if culture-aware evaluation is to remain current rather than freeze one historical snapshot (Wu et al., 2025; Stepanyan et al., 2026; Verma et al., 2026; Ivetta et al., 2025).

6 A Research Agenda for Culturally Grounded NLP

Distinguish multilingual coverage from cultural competence. Reporting aggregate multilingual accuracy is not enough. Evaluations should explicitly separate language coverage from culture-sensitive performance, for example by contrasting translated versus native-authored items, culturally sensitive versus culturally agnostic subsets, and monolingual versus code-mixed or interactional settings (Singh et al., 2025; Ryström et al., 2025; Ochieng et al., 2025).

Publish richer contextual metadata and benchmark slices. Benchmarks should document script practices, domain concentration, register, source authorship, translation pipeline, represented communities, and community validation. The goal is not to create one scalar "culture score," but to make the assumptions behind evaluation auditable (Liu et al., 2025a; Zhou et al., 2025; Wu et al., 2025; Alkhamissi et al., 2026).

Use mixed elicitation protocols. Closed-form questionnaires are useful for comparability, but culture-sensitive evaluation should combine closed and open-ended elicitation, pairwise judgments, and qualitative analysis, while distinguishing genuine norm violation from reasonable local variation (Oh et al., 2025; Kabir et al., 2025; Qiu et al., 2025).

Prefer native-authored and locally validated data where possible. Recent work shows that native preference data, culturally diverse raters, locally authored prompts, and community-grounded datasets surface different errors and can improve adaptation quality (Guo et al., 2025; Chiu et al., 2025; Kim and Lee, 2025; Müller-Eberstein et al., 2025; Ivetta et al., 2025). This suggests that participation should be treated as core research infrastructure rather than optional review.

Evaluate multimodality and within-language variation. Future benchmarks should move beyond text-only settings and beyond standardized language varieties. Multimodal artifacts, local entities, politeness registers, dialects, and narrative traditions are central to real deployment contexts (Nayak et al., 2024; Winata et al., 2025; Maji et al., 2025; Aji and Cohn, 2025; Pranida et al., 2025; Faisal et al., 2024; Shafique et al., 2025).

Treat alignment as ongoing localization rather than one-time correction. Cultural adaptation is likely to remain task-specific, community-specific, and dynamic. Promising methods already exist, but they should be understood as part of a continuous localization process that must accommodate pluralism and change over time (AlKhamissi et al., 2024; Liu et al., 2025a; Feng et al., 2025; Xu et al., 2025; Pham et al., 2025; Stepanyan et al., 2026).

Make participation and governance part of evaluation. Community-engaged data creation and participatory design should be treated as part of the experimental setup, not just project framing. Earlier work in low-resource MT and positive-impact NLP already provides methodological guidance, and newer culturally grounded datasets show why this matters for which harms and categories are surfaced (Nekoto et al., 2020; Caselli et al., 2021; Ivetta et al., 2025; Verma et al., 2026).

Maintain cultural resources as living infrastructure. Cultural knowledge, norms, and harms are dynamic. Benchmarks should therefore support refresh cycles, re-annotation, and versioned slices rather than one-off releases that implicitly fix culture at collection time. Audit and scaling studies suggest that this ongoing maintenance is necessary both for coverage and for preventing majority defaults from becoming entrenched as de facto standards (Wu et al., 2025; Alkhamissi et al., 2026; Stepanyan et al., 2026; Verma et al., 2026).

7 Conclusion

The central lesson of recent work is straightforward: multilingual capability does not automatically yield multicultural competence. The field now has substantial evidence, across text, values, emotion, conversation, multimodality, and low-resource evaluation, that models can transfer technically while still failing culturally (Rystrøm et al., 2025; Singh et al., 2025; Belay et al., 2025; Schneider et al., 2025). Building more culturally grounded NLP systems will require more than scaling data. It will require better benchmark design, richer contextual metadata, community-grounded supervision, and evaluation protocols that recognize languages as part of larger communicative ecologies.

For the ACL/C3NLP community, the immediate implication is methodological rather than purely rhetorical. Culture-sensitive NLP should become a standard reporting axis alongside multilingual accuracy, robustness, and safety. Papers that claim broad generalization should specify whether they tested translated or native-authored items, standard or non-standard varieties, static or interactive settings, and which communities validated the data and labels. Without that shift, the field risks continuing to reward models that are globally legible yet locally brittle. With it, multilingual NLP can move closer to systems that are not merely widespread, but socially usable across the communities they reach.

8 Limitations

This is a synthesis paper rather than a new experimental study, so its claims depend on the coverage and quality of the underlying literature. Although the evidence base spans more than 50 papers, it is still uneven across regions, modalities, and tasks. Many benchmarks are recent, and some target a small number of countries or languages even when their conceptual claims are broad.

A second limitation is that the literature itself often operationalizes culture through partial proxies such as countries, prompt languages, survey instruments, or region-specific artifacts. We follow the literature in discussing these measures, but we do not claim that any one of them captures culture exhaustively. Our use of the term *communicative ecologies* is intended precisely to resist over-identifying culture with any single variable.

A third limitation is temporal. Culture-aware evaluation and alignment are moving quickly, particularly in workshop, multimodal, and low-resource settings. The research agenda proposed here should therefore be read as a living program rather than a closed taxonomy.

9 Ethical Considerations

The main ethical risk in operationalizing culture for NLP is reification. Country, language, script, or survey averages can be analytically useful while still flattening internal diversity, especially for Indigenous, diasporic, minoritized, or non-standard language communities (Zhou et al., 2025; Liu et al., 2025a). We therefore recommend treating such variables as partial descriptors rather than as direct stand-ins for culture.

A second risk concerns optimization targets. If model development focuses only on predictors that maximize average performance, it may reinforce existing inequalities by favoring languages and communities that already benefit from strong representation, stable orthographies, or abundant culturally relevant supervision (Joshi et al., 2020; Blasi et al., 2022). Culture-sensitive evaluation should thus prioritize disparity reporting, documentation, and community relevance in addition to mean scores.

Finally, participatory alignment is not a magic solution. Communities are internally diverse, disagreement is often substantive, and cultural norms change over time. Native preference learning and local annotation should therefore be paired with transparency about whose judgments are represented and what disagreements remain unresolved (Guo et al., 2025; Havaldar et al., 2025; Ferawati et al., 2024).

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Alham Fikri Aji and Trevor Cohn. 2025. LORAXBENCH: A multitask, multilingual benchmark suite for 20 Indonesian languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17421–17446, Suzhou, China. Association for Computational Linguistics.

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

Mai Alkhamissi, Yunze Xiao, Badr AlKhamissi, and Mona T. Diab. 2026. Hire your anthropologist! rethinking culture benchmarks through an anthropological lens. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1218–1235, Rabat, Morocco. Association for Computational Linguistics.

Sina Bagheri Nezhad and Ameeta Agrawal. 2024. What drives performance in multilingual language models? In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 16–27, Mexico City, Mexico. Association for Computational Linguistics.

Sina Bagheri Nezhad, Ameeta Agrawal, and Rhitabrat Pokharel. 2025. Beyond data quantity: Key factors driving performance in multilingual language models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 225–239, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025. CULEMO: Cultural lenses on emotion - benchmarking LLMs for cross-cultural emotion understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 18894–18909, Vienna, Austria. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. CulFiT: A fine-grained cultural-aware LLM training paradigm via multilingual critique data synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22413–22430, Vienna, Austria. Association for Computational Linguistics.

Kiki Ferawati, Wan Jou She, Shoko Wakamiya, and Eiji Aramaki. 2024. Synchronizing approach in designing annotation guidelines for multilingual datasets: A COVID-19 case study using English and Japanese tweets. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 32–41, Bangkok, Thailand. Association for Computational Linguistics.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32866–32895, Suzhou, China. Association for Computational Linguistics.

Katharina Hämmerl, Tomasz Limisiewicz, Jindřich Libovický, and Alexander Fraser. 2025. Beyond literal token overlap: Token alignability for multilinguality. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 756–767, Albuquerque, New Mexico. Association for Computational Linguistics.

Shreya Havaldar, Young Min Cho, Sunny Rai, and Lyle Ungar. 2025. Culturally-aware conversations: A framework & benchmark for LLMs. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 220–229, Suzhou, China. Association for Computational Linguistics.

Guido Ivetta, Marcos J Gomez, Sofía Martinelli, Pietro Palombini, M Emilia Echeveste, Nair Carolina Mazzeo, Beatriz Busaniche, and Luciana Benotti. 2025. HESEIA: A community-based dataset for evaluating social biases in large language models, co-designed in real school settings in Latin America. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25095–25117, Suzhou, China. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the checkbox: Challenging closed-style evaluations of cultural alignment in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24–51, Suzhou, China. Association for Computational Linguistics.

Kyuhee Kim and Sangah Lee. 2025. Nunchi-bench: Benchmarking language models on cultural reasoning with a focus on Korean superstition. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15328–15342, Vienna, Austria. Association for Computational Linguistics.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025a. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025b. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134, Vienna, Austria. Association for Computational Linguistics.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, and Sriparna Saha. 2025. DRISHTIKON: A multimodal multilingual benchmark for testing language models' understanding on Indian culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1289–1313, Suzhou, China. Association for Computational Linguistics.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Max Müller-Eberstein, Mike Zhang, Elisa Bassignana, Peter Brunsgaard Trolle, and Rob Van Der Goot. 2025. DaKultur: Evaluating the cultural awareness of language models for Danish with native speakers. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 50–58, Albuquerque, New Mexico. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6423–6443, Albuquerque, New Mexico. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, and 28 others. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Jean De Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual multimodal LLMs with cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24187–24231, Suzhou, China. Association for Computational Linguistics.

Millicent Ochieng, Varun Gumma, Sunayana Sitaram, Jindong Wang, Vishrav Chaudhary, Keshet Ronen, Kalika Bali, and Jacki O'Neill. 2025. Beyond metrics: Evaluating LLMs effectiveness in culturally nuanced, low-resource real-world scenarios. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 230–247, Vienna, Austria. Association for Computational Linguistics.

Juhyun Oh, Inha Cha, Michael Saxon, Hyunseung Lim, Shaily Bhatt, and Alice Oh. 2025. Culture is everywhere: A call for intentionally cultural evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19156–19168, Suzhou, China. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004.

Viet Thanh Pham, Zhuang Li, Lizhen Qu, and Gholamreza Haffari. 2025. CultureInstruct: Curating multi-cultural instructions at scale. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9207–9228, Albuquerque, New Mexico. Association for Computational Linguistics.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Salsabila Zahirah Pranida, Rifo Ahmad Genadi, and Fajri Koto. 2025. Culturally-nuanced story generation for reasoning in low-resource languages: The case of Javanese and Sundanese. In *Proceedings of the 5th*

Workshop on Multilingual Representation Learning (MRL 2025), pages 369–384, Suzhuo, China. Association for Computational Linguistics.

Haoyi Qiu, Alexander Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of LLM web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3978–4005, Albuquerque, New Mexico. Association for Computational Linguistics.

Jonathan Hvithamar Rystrøm, Hannah Rose Kirk, and Scott Hale. 2025. Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 74–85, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025. GIMMICK: Globally inclusive multimodal multitask cultural knowledge benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9605–9668, Vienna, Austria. Association for Computational Linguistics.

Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, Mohammed Irfan Kurpath, Yahya Hmaiti, Go Inoue, Jean Lahoud, Md. Safirur Rashid, Shadid Intisar Quasem, Maheen Fatima, Franco Vidal, Mykola Maslych, Ketan Pravin More, Sanoojan Baliah, Hasindri Watawana, Yuhao Li, Fabian Farestam, and 10 others. 2025. A culturally-diverse multilingual multimodal video benchmark & model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19998–20022, Suzhou, China. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Hayk Stepanyan, Aishwarya Verma, Andrew Zaldivar, Rutledge Chin Feman, Erin MacMurray van Liemt, Charu Kalia, Vinodkumar Prabhakaran, and Sunipa Dev. 2026. Scaling cultural resources for improving generative models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 6695–6709, Rabat, Morocco. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Aishwarya Verma, Laud Ammah, Olivia Nercy Ndlovu Lucas, Andrew Zaldivar, Vinodkumar Prabhakaran, and Sunipa Dev. 2026. SAFARI: A community-engaged approach and dataset of stereotype resources in the sub-Saharan African context. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 359–370, Rabat, Morocco. Association for Computational Linguistics.

Emilio Villa-Cueva, Sholpan Bolatzhanova, Diana Turmakhan, Kareem Elzeky, Henok Biadglign Ademtew, Alham Fikri Aji, Vladimir Araujo, Israel Abebe Azime, Jinheon Baek, Frederico Belcavello, Fermin Cristobal, Jan Christian Blaise Cruz, Mary Dabre, Raj Dabre, Toqeer Ehsan, Naome A Etori, Fauzan Farooqui, Jiahui Geng, Guido Ivetta, and 16 others. 2025. CaMMT: Benchmarking culturally aware multimodal machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22423–22441, Suzhou, China. Association for Computational Linguistics.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.

Meng-Chen Wu, Si-Chi Chin, Tess Wood, Ayush Goyal, and Narayanan Sadagopan. 2025. Incorporating diverse perspectives in cultural alignment: Survey of evaluation benchmarks through a three-dimensional framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17026–17061, Suzhou, China. Association for Computational Linguistics.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. Self-pluralising culture alignment for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6859–6877, Albuquerque, New Mexico. Association for Computational Linguistics.

Senqi Yang, Dongyu Zhang, Jing Ren, Ziqi Xu, Xiuzhen Zhang, Yiliao Song, Hongfei Lin, and Feng Xia. 2025. Cultural bias matters: A cross-cultural benchmark dataset and sentiment-enriched model for understanding multimodal metaphors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26301–26317, Vienna, Austria. Association for Computational Linguistics.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.