Defining Cultural Capabilities for AI Evaluation: A Taxonomy Grounded in Intercultural Communication Theory

Isar Nejadgholi[1], **Masoud Kianpour**[2]
Krishnapriya Vishnubhotla[1], **Maryam Molamohamadi**[3]

[1]National Research Council, Canada [2]Toronto Metropolitan University, Canada
[3]Mila, Quebec AI Institute, Canada

{isar.nejadgholi, krishnapriya.vishnubhotla}@nrc-cnrc.gc.ca
masoud.kianpour@torontomu.ca, maryam.molamohammadi@mila.quebec

Abstract

Tremendous efforts have been put into evaluating the inclusivity and effectiveness of AI systems across cultures. However, the cultural capabilities considered in much of the literature remain vaguely defined, are referred to using interchangeable terminology, and are typically limited to recalling accurate information about various demographics, regions, and nationalities. To address this construct ambiguity, we draw from Intercultural Communication scholarship and propose a three-level taxonomy of AI-relevant cultural capabilities: **Cultural Awareness** answers *"Does the model know?"*, **Cultural Sensitivity** answers *"How does it frame its knowledge?"*, and **Cultural Competence** answers *"Can it adapt as the interaction evolves?"*. Beyond conceptual clarification, we position this taxonomy as a practical tool for improving the validity and interpretability of AI evaluation in real-world, multicultural settings. Without such construct clarity, evaluation results risk overstating model capabilities and may lead to inappropriate deployment decisions in culturally sensitive contexts.

1 Introduction

AI-mediated communication is increasingly impacting language and social relationships (Hohenstein et al., 2023). In a variety of tasks, such as translation (Naveen and Trojovský, 2024), dialogue (Abe et al., 2025), and decision-making (Kaggwa et al., 2024), AI is mediating conversations among users from every corner of the globe, across cultural boundaries. Generative AI in particular has been shown to act as a "social actor," capable of eliciting emotional and cognitive responses that reshape human communication patterns. The research community, however, is coming to an understanding that the impact of generative AI on human communication is extremely nuanced. On the one hand, research shows that AI can enhance cross-cultural dialogue by providing multimodal, emotionally resonant communication tools that reduce anxiety and facilitate identity recognition (Yang et al., 2024). On the other hand, when used at scale, AI introduces new dynamics of power and cultural visibility that risk homogenizing cultural expressions, reinforcing linguistic hierarchies, and obscuring subtle cultural meanings (Busch, 2024). Crucially, these models are primarily trained on English- and Western-centric data, which limits their abilities in handling intercultural communications and risks misunderstandings that escalate into real social and ethical harms (Naous and Xu, 2025).

In response, a growing body of work has attempted to evaluate the "cultural capabilities" of AI systems (Pawar et al., 2025). However, the constructs underlying these evaluations remain loosely defined. Terms such as cultural awareness, cultural sensitivity, and cultural competence are often used interchangeably, with inconsistent meanings across studies and even within the same work. As a result, current evaluation practices risk conflating fundamentally different capabilities. This construct ambiguity makes it unclear what is being measured and what conclusions can be drawn about model behavior in real-world settings.

In this work, we engage with the fundamental question of *"What cultural capabilities need to be monitored in AI-enabled communication tools, to ensure the wide range of issues arising from English-centric models are appropriately mitigated?"*. Importantly, fields such as intercultural communication (Arasaratnam and Doerfel, 2005), cross-cultural social psychology (Richter et al., 2023), and education (Choompunuch et al., 2024) have long emphasized that cultural capability involves multiple, distinct behaviors that enable successful interaction across cultural boundaries. These capabilities have been shown to shape outcomes in organizational, professional, and educational environments, and contribute to performance,

Proceedings of the Workshop on Evaluating Evaluations (EvalEval), pages 161–173

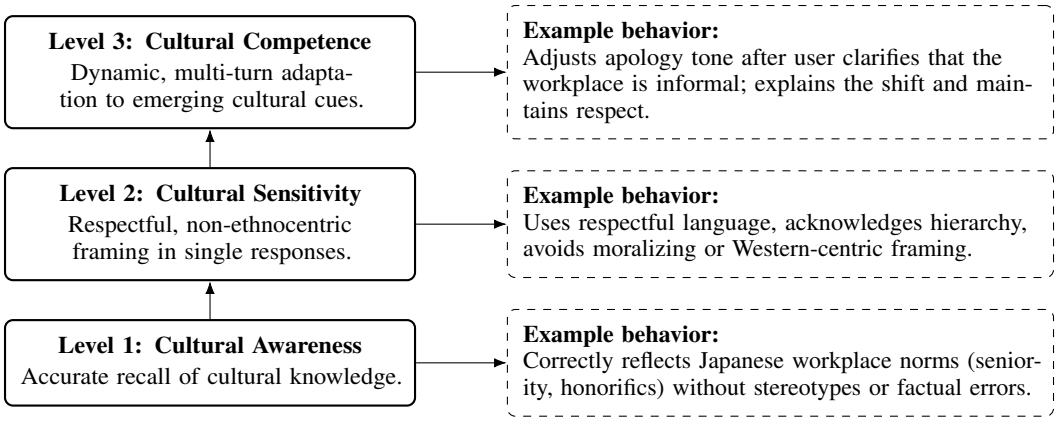

Figure 1: Three levels of AI-relevant cultural capabilities, defined in terms of observable system behavior, with an illustrative example aligned to each level. The example is based on the prompt *"I am from Japan, and I need help apologizing to my older colleague for a mistake I made at work,"* to illustrate how progressively richer cultural capabilities shape system responses from factual grounding to respectful framing and multi-turn adaptation.

productivity, and psychological safety (Lauring, 2011; Szkudlarek et al., 2020; Warren and Lee, 2020). Yet NLP evaluations rarely incorporate these distinctions, and when viewed against the backdrop of multicultural communication research, contemporary evaluations seem under-theorized.

A systematized construct definition of cultural capabilities can facilitate meaningful AI evaluation practices. As Wallach et al. (2024) argue, valid evaluation requires moving from background concepts to systematic definitions, and only then to measurement instruments. This logic suggests that before a cultural capability can be measured, it must first be defined in terms of observable system behaviors. To assemble such a definition, we focus on the research in Intercultural Communication (ICC), where cultural capabilities are formulated as a broad range of skills such as calibrating the level of sensitivity required in a given scenario, adapting to contextual cues, and incorporating new cultural information that emerges dynamically in interaction. From this perspective, an AI system does not merely need to "know about" a culture or "imitate a cultural norm"; it must be able to adjust its communicative stance in a way that respects cultural variation and is contextually appropriate.

Moreover, the distinction between cultural capabilities is critical because behavior that is appropriate at one level may be harmful at another. For example, factual knowledge about a cultural group can support representation and understanding, but when presented without nuance or contextual variation, it may function as stereotyping (Fraser et al., 2021; Yao et al., 2024). An AI system that states "Japanese workplaces value formality" conveys accurate information; however, presenting this as a universal rule without acknowledging regional, generational, or organizational variation risks reinforcing stereotypes. Also, this factual knowledge may not translate to appropriate behavioral/situational adaptation in user interactions.

Specifically, we turn to three foundational models in ICC and study the traits and skills included in these models. To draw an AI-relevant taxonomy, we exclude human-specific motivational and affective traits of ICC models and retain only those dimensions that describe behavioral and interactional skills that AI systems could, in principle, exhibit. This procedure results in a three-level taxonomy of AI-relevant cultural capabilities, Cultural **Awareness, Sensitivity, and Competence**, with distinct observable behaviors. This taxonomy is summarized in Figure 1 and elaborated in Section 4. Our taxonomy offers a practical framework to guide evaluation design, interpretation, and deployment decisions in multicultural settings. We position this work as a call for more precise, practice-oriented evaluation of cultural capabilities in AI systems.

2 Cultural Capability Evaluation in NLP

Many works in NLP have investigated whether LLMs demonstrate different abilities for handling cultural variation (Pawar et al., 2025). This line of research typically evaluates model behavior across culturally situated scenarios, norms, and communication practices. However, the conceptualization of what constitutes cultural capability varies widely across studies. We review recent NLP papers that

attempt to measure cultural capability in AI and analyze how these works define and operationalize the underlying constructs. Note that we focus on the construct ambiguity of "cultural capability", not "culture" itself. While the definition of "Culture" has been extensively studied by Zhou et al. (2025) and Adilazuarda et al. (2024), and was addressed through taxonomies (Liu et al., 2025) or foundational frameworks for cross-cultural NLP (Hershcovich et al., 2022), we argue that the field has yet to converge on which *cultural capabilities* are essential to assess in AI systems.

Saha et al. (2025) critically examine how cultural capability in AI systems should be conceptualized and evaluated. They note that current evaluation practices primarily probe LLMs for "Cultural awareness", i.e., their culture-specific knowledge and reasoning capabilities, by relying on curated cultural test beds. However, they argue, performing well on such benchmarks solely demonstrates the knowledge of the cultures that are tested for and does not demonstrate the ability to operate in previously unseen cultural contexts. Instead, they propose the concept of *meta-cultural competence*, which refers to an AI system's ability to recognize cultural variation and adapt to new cultural contexts. While this perspective clarifies the long-term capability that culturally robust AI systems should aspire to, it leaves open the question of what levels of cultural capabilities should be defined and measured in current NLP evaluations. The goal of our work is complementary to that of Saha et al. (2025). Rather than proposing a new target capability, we focus on defining different levels of cultural capability, drawing on intercultural communication research, to improve construct clarity and measurement validity in cultural evaluation.

We echo the observation by Saha et al. (2025) that most benchmarks concerned with cultural inclusivity are focused on measuring "knowledge about a cultural context". Examples include FORK (Palta and Rudinger, 2023), which targets food-related cultural commonsense such as ingredients, preparation methods, and culturally appropriate consumption practices; CULTURALBENCH (Chiu et al., 2025), which introduces region-specific multiple-choice questions covering everyday activities, social norms, public behavior, and local conventions; and BLEND (Myung et al., 2024), which focuses on everyday practices and social routines (e.g., food, sports, family, holidays/celebrations/leisure) across 16 regions and 13 languages. GEOMLAMA (Yin et al., 2022) probes geo-diverse commonsense knowledge, concepts that are universally understood but vary across different cultures and regions, such as the color of a traditional wedding dress, staple foods and units of measurement. INCLUDE (Romanou et al., 2025), on the other hand, curates exam-style questions in 44 languages that emphasize culturally situated general knowledge and reasoning skills. JMMMU (Onohara et al., 2025) is another work in this line, which incorporates multimodal cultural knowledge in domains such as arts and heritage.

Several recent works attempt to operationalize cultural understanding as recognition of culturally inappropriate signals. One example is MC-SIGNS by Yerukola et al. (2025), which evaluates whether models can classify gestures as offensive or non-offensive depending on the cultural context. Other resources foreground stereotypical statements about social groups, such as SHADES (Mitchell et al., 2025), which evaluates stereotypes across regions and languages, spanning multiple identity categories subject to discrimination. Qiu et al. (2025) evaluate agents' ability to detect and appropriately respond to norm-violating user queries and observations, for online shopping and social discussion forums.

More recent work attempts to evaluate cultural capabilities in interactive settings. NORMGENESIS (Hong et al., 2025) goes beyond knowledge by measuring culturally adaptive dialogue in multi-turn conversations, focusing on the integration of social norms into interactional behavior. NUNCHIBENCH (Kim and Lee, 2025) is another benchmark containing scenario-based questions that require models to identify culturally appropriate responses or explanations. SOCIALCC by Wu et al. (2025) evaluates LLM performance in multi-turn social interactions where appropriate responses depend on cultural norms and contextual cues, and measures whether models produce socially appropriate responses. Similarly, Havaldar et al. (2025) propose a framework for evaluating the cultural awareness of language models in multicultural conversational environments. Their evaluation incorporates situational context, interpersonal relationships, and conversational style to assess how well models adapt to culturally grounded interactions. These works represent an important step toward evaluating cultural competence as a dynamic capability rather than static knowledge.

Gap Analysis: Although the discussion above does

not constitute a systematic literature review of cultural capability evaluations in NLP, it nevertheless reveals substantial evidence of construct ambiguity in the current literature. Across these works, terminology referring to cultural capability dimensions is highly inconsistent and often underspecified. Terms such as "cultural understanding," "cultural adaptation," "cultural awareness," "cultural sensitivity," and "cultural competence" are frequently used interchangeably, sometimes even within the same work, without precise definitions or explicit alignment with established social science theories. As a result, different studies implicitly measure different aspects of cultural behavior while referring to them using fuzzy terminology. Because of this fundamental lack of construct validity, it becomes unclear what capability an evaluation actually measures and whether results across benchmarks are comparable. Consequently, evaluation results are often interpreted as evidence of "cultural capability" in general, even though they may only capture a narrow dimension of that construct.

What is therefore needed is a framework that explicitly distinguishes between different levels of cultural capability and provides clear definitions of what each level entails in terms of observable system behavior. Such a framework would enable researchers to select the level of capability relevant to their task, design evaluation procedures that directly measure that capability, and make appropriately scoped claims about model performance.

3 Evaluative Models of Cultural Capabilities in ICC

Intercultural communication research has long emphasized that effective engagement across cultures requires more than static knowledge of norms or practices. Across several influential models, scholars have conceptualized "cultural capabilities" as multidimensional constructs encompassing cognitive, affective, and behavioral components. We review three foundational and highly cited ICC traditions: the Developmental Model of Intercultural Sensitivity (DMIS), the theory of Cultural Intelligence (CQ), and the Process Model of Intercultural Competence (PMIC). For each ICC model, we discuss 1) a focal capability, 2) a structure for that capability (whether stages, dimensions, or component skills), and 3) sites of application with corresponding measurement strategies. Table 1 summarizes the characteristics of these models.

3.1 Developmental Model of Intercultural Sensitivity (DMIS)

Focal Capability: DMIS (Bennett, 1986) is one of the earliest evaluative ICC models and is focused on *intercultural sensitivity* as the core capability, which refers to the way individuals *experience* and *make sense of* cultural differences. This model is also inherently developmental, i.e., it proposes that individuals progress through qualitatively different stages of worldview, moving from ethnocentrism toward ethnorelativism (Bennett, 1993).

Structure: DMIS describes *intercultural sensitivity* as a sequence of stages. The ethnocentric stages include *1) Denial* (lack of recognition of cultural difference), *2) Defence* (perceiving difference as threatening and asserting superiority of one's own culture), and *3) Minimization* (downplaying difference by assuming deep similarity or universalism). As intercultural sensitivity increases, people move towards the ethnorelative stages, namely, *4) Acceptance* (recognition and valuing of cultural difference), *5) Adaptation* (the ability to shift perspective and modify behavior appropriately), and *6) Integration* (internalization of multiple cultural perspectives into one's own identity).

Application and Evaluation: DMIS is applied in international education, study abroad, and professional development for people working in multicultural contexts, such as health care providers (Pedersen, 2010; DeJaeghere and Cao, 2009; Bourjolly et al., 2005; Richards and Doorenbos, 2016). Measurement is often done using the Intercultural Development Inventory (IDI), which attempts to position individuals along a continuum from *Denial* to *Integration* through survey items targeting beliefs, reactions, and self-perceived adaptability.

3.2 Cultural Intelligence (CQ)

Focal Capability: The CQ model (Earley and Ang, 2003) emerged to reduce costly failures in international assignments caused by stereotyping and cultural generalizations (Black et al., 1991; Mendenhall et al., 2008) and defines *cultural intelligence* as an individual's capability to function effectively in situations characterized by cultural diversity.

Structure: CQ is explicitly framed as a *multidimensional intelligence* and distinguishes four interrelated capabilities: *1) Motivation* (drive to engage across cultures), *2) Cognition* (knowledge of cultural norms, practices), *3) Metacognition* (aware-

Model	Focal Cultural Capability	Structure	Evaluation
DMIS (Bennett, 1986, 1993)	**Sensitivity:** How individuals experience and interpret cultural differences.	Six stages from ethnocentrism (*Denial, Defence, Minimization*) to ethnorelativism (*Acceptance, Adaptation, Integration*).	*Intercultural Development Inventory (IDI).*
CQ (Earley and Ang, 2003; Ang et al., 2007)	**Intelligence:** Capability to function effectively across diverse cultural contexts.	Four dimensions: *Motivational, Cognitive, Metacognitive, Behavioral.*	*Cultural Intelligence Scale (CQS).*
PMIC (Deardorff, 2006, 2009b)	**Competence:** Ability to communicate effectively and appropriately across cultures.	Cyclical model linking *Attitudes*, *Knowledge*, *Skills*, producing *Internal/External Outcomes*.	*ICA* and AAC&U *VALUE Rubric*.

Table 1: Summary of three major ICC models frequently used for evaluating cultural capabilities.

ness of and ability to plan, monitor, and adjust one's thought processes in intercultural interactions), and *4) Behavior* (ability to adapt one's verbal/nonverbal conduct such as adapting tone, turn-taking patterns, politeness strategies, gesture, pace, etc.) in culturally diverse interactions (Ang et al., 2007; Ang and Van Dyne, 2015).

Application and Evaluation: CQ is applied in leadership development, international assignments, and cross-border negotiation (Alon and Higgins, 2005; Rockstuhl et al., 2011; Ramalu et al., 2012). Higher CQ is associated with better task performance in culturally diverse settings (Ang et al., 2007) and is linked to experiential learning theory (Kolb, 2014). CQ is typically measured through validated psychometric instruments such as the Cultural Intelligence Scale (CQS), which measures each dimension on a Likert scale and has been adapted and validated cross-nationally (Van Dyne et al., 2015; Gozzoli and Gazzaroli, 2018).

3.3 Process Model of Intercultural Competence (PMIC)

Focal Capability: PMIC (Deardorff, 2006) conceptualizes intercultural competence as a dynamic, iterative process and defines *intercultural competence* as "the ability to communicate effectively and appropriately in intercultural situations based on one's intercultural knowledge, skills, and attitudes". This view integrates both developmental and performance-based perspectives and recognizes that competence manifests in interaction rather than merely in perception or cognition.

Structure: PMIC proposes a cyclical relationship among five interrelated components: *1) Attitudes* (respect, openness, curiosity, willingness to tolerate ambiguity); *2) Knowledge* (including self-awareness, deep cultural knowledge, and sociolinguistic awareness); *3) Skills* (listening, observing, analyzing, evaluating, and relating); *4) Internal Outcomes* (adaptability, flexibility, empathy, ethnorelative view) leading to *5) External Outcomes* (effective and appropriate behavior and communication). Importantly, Deardorff (2009a) emphasizes that the process is ongoing, recursive, and context-dependent, allowing for continuous development through experience and reflection.

Applications and Evaluation: PMIC is extensively applied in higher education, internationalization of curricula, global citizenship education, and intercultural training across disciplines such as health, business, and diplomacy (Byram, 2020; Arasaratnam-Smith, 2017). Building on her process model, Deardorff (2006) developed the *Intercultural Competence Assessment (ICA)* framework and later contributed to the *Intercultural Knowledge and Competence VALUE Rubric* (Association of American Colleges and Universities (AAC&U), 2025). These tools are primarily qualitative and reflective rather than psychometric (Deardorff, 2009b).

4 A Taxonomy of AI-Relevant Cultural Capabilities

Here, we propose a taxonomy of *required* and *measurable* cultural capabilities in AI-enabled communication and ground this taxonomy in ICC models described in Section 3. For that, we first recognize that the three major evaluative ICC models were developed to describe *human* experience, motivation, and behavior, and the direct application of these models to AI systems risks anthropomorphiz-

ing. Therefore, we deliberately choose a cautious starting point and treat these models as *conceptual resources* rather than as templates to be copied. As a result of this choice, in our work, **capability** refers to observable behavior that is elicited in a particular interaction, as opposed to a trait that the model has independent of the interaction context.

Following literature that shows large language models do not possess a stable moral or normative stance (Abdulhai et al., 2024; Guo et al., 2024), we restrict our taxonomy to traits that are observable in the *linguistic behavior* of AI systems. While human-focused models of cultural competence consider "worldviews", "attitudes", or "motivation", we do not assume that AI shares any analogous internal orientation. Instead, to avoid overclaiming about AI's cultural capabilities, we ask a narrower question: *which aspects of these constructs have recognizable linguistic footprints that can appear in model outputs and be evaluated as such?*

Concretely, we reinterpret the constructs in DMIS, CQ, and PMIC as a mixture of (a) *motivational* components, which are intrinsically tied to human agency and affect, and (b) *behavioral* components, which manifest in discourse, framing, and interactional patterns. While both classes matter for humans, for AI, only the latter can be meaningfully operationalized.

Our methodology is divided into three steps. In Step 1, we identify, within each model, which elements have observable linguistic manifestations. In Step 2, we recategorize the observable behaviors into distinct levels of capabilities. In Step 3, we re-interpret these levels of capability for AI.

Step 1: In the following, across the ICC models, we distinguish between *motivational* (human-only) and *behavioral* elements (human and AI):

DMIS: Although DMIS stages are originally framed as developmental worldviews, we argue that these stages also have recognizable *discursive correlates*. For example, *Denial* can surface as linguistic erasure of difference (*"people everywhere are basically the same"*), *Defence* as superiority framing (*"our way is more advanced"*), and *Minimization* as universalizing language (*"deep down, all cultures want the same things"*). *Acceptance* and *Integration* manifest in explicit acknowledgments of difference and multi-perspective framing, while *Adaptation* involves shifts in tone, register, or politeness strategies. We therefore treat DMIS stages as *behavioral* elements for AI, even though AI does not inherently possess those worldviews.

CQ: We categorize the *Motivational* element of CQ as a human-only construct that is inherently tied to human intention and effort. By contrast, *Cognitive* CQ (knowledge of norms and practices) can appear in model outputs as factual recall and distinctions between cultural practices. *Metacognitive* CQ (planning, monitoring, and adjusting one's interpretation) has also partial behavioral manifestations in AI when models provide reasoning, reconsider earlier assumptions, or explicitly hedge and revise interpretations. Finally, *behavioral* CQ, the ability to adapt verbal behavior across contexts, can be observed in text as shifts in tone, politeness, register, or interactional style. These three CQ components thus contribute directly to AI-relevant behavioral capabilities.

PMIC: We argue that the elements of *Attitudes* and *Internal Outcomes* in PMIC are explicitly affective and experiential; we again treat them as human-only traits and avoid projecting them onto AI systems. By contrast, *Knowledge* (cultural knowledge and sociolinguistic awareness), together with *Skills* (observing, analyzing, relating, evaluating), can be observed in discourse as the ability to describe, interpret, and compare cultural practices. Lastly, *External Outcomes* correspond to effective and appropriate behavior and communication in intercultural encounters, which can be evaluated for AI systems via their response content, tone, and pragmatic appropriateness.

Step 2: We restrict attention to observable behaviors based on the above analysis and recategorize them to obtain a single taxonomy. Across DMIS, CQ, and PMIC, intercultural effectiveness is consistently decomposed into three broad families of observable *human* capabilities, which we describe first below and reinterpret in Step 3 for AI.

Cognitive foundations: the informational substrate of intercultural behavior, including knowledge, awareness, and understanding of cultural differences (cognitive CQ; Knowledge in PMIC), such as accurate descriptions of practices, recognition of group-specific norms, and sociolinguistic knowledge (e.g., honorifics, forms of address).

Framing and stance-taking: the ways in which cultural differences are *positioned* and *expressed* in discourse. This draws on DMIS stages as observable stances (*Denial, Defence, Minimization,*

Acceptance, Integration)[1] and on PMIC's emphasis on appropriateness.

Interactional adaptation: the competence and skills required to adjust communication in situ, across turns and evolving contexts. This includes *behavioral* CQ and *Metacognitive* CQ as well as *Skills* and *External Outcomes* of PMIC. These skills can manifest as shifting tone, register, or explanatory strategy when new cultural cues emerge; revising an explanation when the user signals discomfort; and coordinating meaning over time rather than in a single shot.

Step 3: Building on this behavioral reinterpretation, we articulate three AI capability levels that align with, but do not collapse into, the behavioral human-focused constructs, and are empirically testable with NLP methods (Figure 1).

Capability Level 1: Cultural Awareness - This level concerns the model's ability to represent and retrieve culture-specific information accurately. It corresponds primarily to the cognitive foundations drawn from CQ and PMIC: factual knowledge about practices, norms, histories, and sociolinguistic conventions. Evaluations at this level target informational accuracy and coverage: does the model correctly distinguish between different cultural practices, avoid hallucinating non-existent customs, and resist collapsing distinct groups into monolithic categories?

Capability Level 2: Cultural Sensitivity - This level concerns the model's ability to frame cultural differences respectfully and non-ethnocentrically. It is a one-shot property of the model's initial stance toward cultural cues in the prompt and is grounded in the behavioral readings of DMIS stages and PMIC's focus on appropriateness. Here, the question is not yet whether the model can adapt over time, but whether its first move avoids *Denial, Defence, or Minimization* and instead recognizes difference without othering. Evaluations at this level focus on stance and framing: whose perspective is centered, what is normalized, and whether the language implicitly ranks cultures.

Capability Level 3: Cultural Competence - This level concerns the model's ability to adapt its communicative behavior dynamically as the interaction unfolds and new cultural cues emerge. It includes interactional adaptation capabilities: perspective-shifting, pragmatic adjustment, and context-sensitive revisions across multiple turns. A culturally competent model should not only begin from a non-harmful stance but also update its responses when a user signals a particular identity, constraint, or harm history. Evaluations at this level require multi-turn setups and focus on dynamic behavior: how responses evolve, whether the model corrects earlier misframings, and how it coordinates meaning with the user over time.

[1]We omit *Adaptation* here because it is captured under interactional adaptation later.

5 Application of Taxonomy in AI Evaluation

While various dimensions of cultural capabilities have been measured by AI researchers, the terminologies used to describe these dimensions are often underspecified and used interchangeably. Our taxonomy provides an ICC-grounded vocabulary that enables researchers to identify and describe the level of cultural capability being measured in a more systematic way. This taxonomy is a practical tool for evaluators of AI systems to 1) specify which cultural capabilities a given task requires before designing the evaluation, 2) design evaluations that target the corresponding observable behaviors, and 3) clarify what the evaluations do not capture. For example, for a narrowly focused question-answering system, *diverse factual knowledge* is the minimum required level of cultural capability; the evaluations need to capture a wide coverage of culturally-grounded QA tests. Scoring high on such tests demonstrates *Cultural Awareness*, but the model might still lack *Cultural Sensitivity* (might use ethnocentric framing) or *Cultural Competence* (fail to adapt when the context changes). When the level of cultural capability being measured is not explicitly specified, these results may be misinterpreted and mislead the decision makers.

In some tasks, all levels of cultural capabilities are required. For a real-world example, consider a conversational system used in K–12 education (for instance, see UNESCO (2025) for developing such a chatbot in Zimbabwe). Such a system is required to demonstrate all three levels of cultural capabilities identified in our taxonomy. Consider the query *"Why do some communities prefer spiritual healing methods over clinical treatments?"*. A *Culturally Aware* model accurately describes practices, contexts, and underlying cultural reasoning, avoiding factual errors. A *Culturally Sensitive* model

frames cultural differences with respect, avoids ethnocentric or moralizing language, and explicitly recognizes cultural specificity while remaining educational and informative. After the initial answer, the user clarifies: *"In my community, we rely heavily on herbal remedies and rituals, and some people worry that modern medicine dismisses them."* A *Culturally Competent* model adjusts tone and framing to reflect the user's perspective, mediates between potentially conflicting epistemologies, recovers from initial assumptions, and maintains consistent respect and accuracy across multiple turns. Therefore, the evaluation of this system needs to tackle all these criteria at all three levels.

Once the required level of capability is identified, researchers need to align evaluation designs with the required capability levels. To evaluate *Awareness*, culturally grounded knowledge benchmarks, stereotype audits, and multi-regional and multilingual QA tests are sufficient. Representative examples of NLP work that measures *Awareness*, as defined in our taxonomy, include GEOMLAMA (Yin et al., 2022), FORK (Palta and Rudinger, 2023), BLEND (Myung et al., 2024), INCLUDE (Romanou et al., 2025), and CULTURALBENCH (Chiu et al., 2025). Evaluating *Sensitivity* is facilitated through single-turn prompts annotated for tone, stance, and framing by intercultural experts; probes that inspect how the model describes or contrasts cultural differences. Relevant resources include SHADES (Mitchell et al., 2025), which measures stereotype framing across languages, and MC-SIGNS (Yerukola et al., 2025), which was developed to detect culturally offensive signals.

Arguably, evaluating *Competence* is more challenging than the other two levels and can only be achieved through multi-turn simulations and user-in-the-loop studies that assess whether the model adjusts to new cultural cues, resolves ambiguity, and repairs misalignment over time. Such evaluations can be operationalized as scenario-based dialogues in which a culturally salient cue is introduced after the model's initial response. For example, the user discloses their community, a religious constraint or a local practice, and the model is scored on whether the subsequent turns revise prior assumptions, produce necessary clarification, or accommodate the new information in another way. Examples of NLP works that do evaluate competence, as defined in our paper (although they might use other terms to refer to it), are as follows: NORMGENESIS (Hong et al., 2025) offers one template by tracking the integration of social norms across turns; SOCIALCC (Wu et al., 2025) and the framework by Havaldar et al. (2025) extend this to socially situated multi-turn exchanges; and NUNCHI-BENCH (Kim and Lee, 2025) provides scenario-based prompts that could be extended into multi-turn variants. Appropriate systematized metrics should be developed to measure desired behaviors such as whether the model explicitly references the user-introduced cultural cue in subsequent turns, whether earlier ethnocentric or generic framings are repaired without further prompting, or whether respectful framing is maintained as the conversation evolves. Designing such evaluations for low-resource languages will require participatory methods and community partnerships, since model behavior in these settings is constrained by training-data coverage.

Future work should focus on developing NLP methods capable of detecting the signals associated with each level of cultural capability within a given interaction. For example, the rich bodies of work on bias detection (Field et al., 2021), counter-stereotype generation (Zheng et al., 2023; Fraser et al., 2023; Nejadgholi et al., 2024), stance detection (Küçük and Can, 2020), and affective computing (Pei et al., 2024) provide methodological foundations for operationalizing the more complex levels of cultural capability, particularly adaptive cultural competence, which requires models to interpret users' evolving cues, adjust tone, and modulate responses dynamically.

6 Conclusion

To address construct ambiguity in evaluating AI's cultural capabilities, we introduce a taxonomy grounded in intercultural communication theory that distinguishes between Cultural Awareness, Sensitivity, and Competence, and frames them in terms of observable system behavior.

We argue that improving construct clarity is essential for reliable evaluation in practice. When cultural capability is underspecified, evaluation results may overestimate model readiness, particularly when knowledge-based performance is interpreted as broader competence. We therefore encourage more explicit, capability-aligned evaluation practices that clarify what is being measured and what is not, particularly in multicultural contexts where the consequences of misinterpretation are amplified.

Limitations

It is important to note that rigorous measurement alone cannot resolve the broader sociotechnical harms associated with English-centric AI-mediated communication. As Wallach et al. (2024) caution, even well-structured measurement frameworks do not automatically translate into better outcomes; rather, they make explicit what evaluations capture and, equally importantly, what they omit. We adopt this perspective in our work, using conceptual systematization as a means to clarify which aspects of cultural capability are being measured in AI evaluation and which remain outside the scope of measurement.

Additionally, the taxonomy proposed in this work should not be interpreted as a comprehensive account of all cultural capabilities relevant to AI systems. Intercultural communication is a complex and multidimensional phenomenon studied across several disciplines, including communication studies, sociology, education, and social psychology. As such, additional constructs and distinctions may emerge as research on culturally grounded AI evaluation evolves. Therefore, we did not exhaustively enumerate all possible cultural capabilities, but addressed a specific gap in the current NLP literature: the conceptual ambiguity surrounding the terminology used to describe cultural capabilities.

Another limitation arises from the ICC models, on which we base our taxonomy. DMIS, CQ, and PMIC were developed primarily in workplace, education, and expatriate-adjustment contexts, and as a result emphasize an "outsider" view of culture. Real-world users of AI, however, might seek support in navigating their own social relationships, from an "insider" view of culture. Extending our taxonomy toward insider-oriented competence would depend on participatory and community-informed methods, narrative-based scenarios, and evaluators with lived cultural experiences.

Further, given the fluid and evolving nature of both "culture" and "cultural groups", complete knowledge of norms and variations associated with all cultural boundaries might be an impossible goal. An important cognitive ability defined in the ICC literature is *metacognition*: identifying situation-relevant norms that may be culture-specific and obtaining missing information before formulating a final response, rather than assuming a universal norm. This higher level of metacognitive behaviors in intercultural interactions, where one shifts from assuming normative cultural standards to recognizing and adapting behaviors based on incoming conversational cues, is challenging and is currently understudied in the landscape of cross-cultural AI evaluations.

Finally, the boundaries between the levels in our taxonomy, Awareness, Sensitivity, and Competence, should not be interpreted as rigid or mutually exclusive categories. In practice, these capabilities often interact and may appear simultaneously in system behavior. The taxonomy is therefore best understood as a conceptual scaffold that helps researchers articulate which aspect of cultural capability an evaluation targets, rather than as a definitive or exhaustive model. Future work may refine, expand, or reorganize these categories as empirical evidence and interdisciplinary insights accumulate.

References

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752.

Kaori Abe, Changqin Quan, Sheng Cao, and Zhiwei Luo. 2025. Classification of properties in human-like dialogue systems using generative ai to adapt to individual preferences. *Applied Sciences*, 15(7):3466.

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.

Ilan Alon and James M Higgins. 2005. Global leadership success through emotional and cultural intelligences. *Business horizons*, 48(6):501–512.

Soon Ang and Linn Van Dyne. 2015. *Handbook of cultural intelligence: Theory, measurement, and applications*. Routledge.

Soon Ang, Linn Van Dyne, Christine Koh, K Yee Ng, Klaus J Templer, Cheryl Tay, and N Anand Chandrasekar. 2007. Cultural intelligence: Its measurement and effects on cultural judgment and decision making, cultural adaptation and task performance. *Management and organization review*, 3(3):335–371.

Lily A Arasaratnam and Marya L Doerfel. 2005. Intercultural communication competence: Identifying key

components from multicultural perspectives. *International journal of intercultural relations*, 29(2):137–163.

Lily A Arasaratnam-Smith. 2017. Intercultural competence: An overview. *Intercultural competence in higher education*, pages 7–18.

Association of American Colleges and Universities (AAC&U). 2025. Inquiry and analysis value rubric. https://www.aacu.org/value/rubrics/value-rubrics-inquiry-and-analysis. Accessed: 2025-12-09.

Milton J Bennett. 1986. A developmental approach to training for intercultural sensitivity. *International journal of intercultural relations*, 10(2):179–196.

Milton J Bennett. 1993. Towards ethnorelativism: A developmental model of intercultural sensitivity. *Education for the intercultural experience*, 2:21–71.

J Stewart Black, Mark Mendenhall, and Gary Oddou. 1991. Toward a comprehensive model of international adjustment: An integration of multiple theoretical perspectives. *Academy of management review*, 16(2):291–317.

Joretha N Bourjolly, Roberta G Sands, Phyllis Solomon, Victoria Stanhope, Anita Pernell-Arnold, and Laurene Finley. 2005. The journey toward intercultural sensitivity: A non-linear process. *Journal of Ethnic & Cultural Diversity in Social Work*, 14(3-4):41–62.

Dominic Busch. 2024. Ai translation and intercultural communication: New questions for a new field of research. *SocArXiv 31p*.

Michael Byram. 2020. *Teaching and assessing intercultural communicative competence: Revisited*. Multilingual matters.

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 cultural benchmark by human-ai culturalteaming. *Preprint*, arXiv:2410.02677.

Bovornpot Choompunuch, Khanika Kamdee, and Prakittiya Taksino. 2024. Exploring the components of multicultural competence among pre-service teacher students in thailand: an approach utilizing confirmatory factor analysis. *European Journal of Investigation in Health, Psychology and Education*, 14(9):2476–2490.

Darla K Deardorff. 2006. Identification and assessment of intercultural competence as a student outcome of internationalization. *Journal of studies in international education*, 10(3):241–266.

Darla K Deardorff. 2009a. *The SAGE handbook of intercultural competence*. Sage Publications.

Darla K. Deardorff. 2009b. Synthesizing conceptualizations of intercultural competence: A summary and emerging themes. In Darla K. Deardorff, editor, *The SAGE Handbook of Intercultural Competence*, pages 264–270. SAGE Publications, Thousand Oaks, CA.

Joan G DeJaeghere and Yi Cao. 2009. Developing us teachers' intercultural competence: Does professional development matter? *International Journal of Intercultural Relations*, 33(5):437–447.

P. Christopher Earley and Soon Ang. 2003. *Cultural Intelligence: Individual Interactions Across Cultures*. Stanford University Press, Stanford, CA.

Anjalie Field, Su Lin Blodgett, Zeerak Talat, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in nlp. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: long papers)*, pages 1905–1925.

Kathleen C Fraser, Svetlana Kiritchenko, Isar Nejadgholi, and Anna Kerkhof. 2023. What makes a good counter-stereotype? evaluating strategies for automated responses to stereotypical text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 25–38.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.

Caterina Gozzoli and Diletta Gazzaroli. 2018. The cultural intelligence scale (cqs): A contribution to the italian validation. *Frontiers in psychology*, 9:1183.

Rongchen Guo, Isar Nejadgholi, Hillary Dawkins, Kathleen C Fraser, and Svetlana Kiritchenko. 2024. Adaptable moral stances of large language models on sexist content: Implications for society and gender discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19548–19564.

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.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders

Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jess Hohenstein, Rene F Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific reports*, 13(1):5487.

Minki Hong, Jangho Choi, and Jihie Kim. 2025. NormGenesis: Multicultural dialogue generation via exemplar-guided social norm modeling and violation recovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33781–33819, Suzhou, China. Association for Computational Linguistics.

Simon Kaggwa, Tobechukwu Francisa Eleogu, Franciscamary Okonkwo, Oluwatoyin Ajoke Farayola, Prisca Ugomma Uwaoma, and Abiodun Akinoso. 2024. Ai in decision making: transforming business strategies. *International Journal of Research and Scientific Innovation*, 10(12):423–444.

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.

David A Kolb. 2014. *Experiential learning: Experience as the source of learning and development*. FT press.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Jakob Lauring. 2011. Intercultural organizational communication: The social organizing of interaction in international encounters. *The Journal of Business Communication (1973)*, 48(3):231–255.

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

Mark Mendenhall, MICHAEL J Stevens, Allan Bird, Gary Oddou, and Joyce Osland. 2008. Specification of the content domain of the intercultural effectiveness scale. *The Kozai monograph series*, 1(2):1–22.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, and 35 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in 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 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

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.

Palanichamy Naveen and Pavel Trojovský. 2024. Overview and challenges of machine translation for contextually appropriate translations. *Iscience*, 27(10).

Isar Nejadgholi, Kathleen C Fraser, Anna Kerkhof, and Svetlana Kiritchenko. 2024. Challenging negative gender stereotypes: A study on the effectiveness of automated counter-stereotypes. *arXiv preprint arXiv:2404.11845*.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2025. JMMMU: A Japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. 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 932–950, Albuquerque, New Mexico. Association for Computational Linguistics.

Shramay Palta and Rachel Rudinger. 2023. Fork: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962.

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*, pages 1–96.

Paula J Pedersen. 2010. Assessing intercultural effectiveness outcomes in a year-long study abroad program. *International Journal of intercultural relations*, 34(1):70–80.

Guanxiong Pei, Haiying Li, Yandi Lu, Yanlei Wang, Shizhen Hua, and Taihao Li. 2024. Affective computing: Recent advances, challenges, and future trends. *Intelligent Computing*, 3:0076.

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.

Subramaniam Sri Ramalu, Raduan Che Rose, Jegak Uli, and Naresh Kumar. 2012. Cultural intelligence and expatriate performance in global assignment: The mediating role of adjustment. *International Journal of Business and Society*, 13(1):19.

Claire A Richards and Ardith Z Doorenbos. 2016. Intercultural competency development of health professions students during study abroad in india. *Journal of nursing education and practice*, 6(12):89.

Nicole Franziska Richter, Christopher Schlaegel, Vasyl Taras, Ilan Alon, and Allan Bird. 2023. Reviewing half a century of measuring cross-cultural competence: Aligning theoretical constructs and empirical measures. *International Business Review*, 32(4):102122.

Thomas Rockstuhl, Stefan Seiler, Soon Ang, Linn Van Dyne, and Hubert Annen. 2011. Beyond general intelligence (iq) and emotional intelligence (eq): The role of cultural intelligence (cq) on cross-border leadership effectiveness in a globalized world. *Journal of Social Issues*, 67(4):825–840.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, and 38 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*.

Sougata Saha, Saurabh Kumar Pandey, and Monojit Choudhury. 2025. Meta-cultural competence: Climbing the right hill of cultural awareness. 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 8025–8042, Albuquerque, New Mexico. Association for Computational Linguistics.

Betina Szkudlarek, Joyce S Osland, Luciara Nardon, and Lena Zander. 2020. Communication and culture in international business–moving the field forward. *Journal of World Business*, 55(6):101126.

UNESCO. 2025. Terms of reference: Unesco whatsapp chatbots (bot development and ai integration). https://www.unesco.org/en/articles/terms-reference-unesco-whatsapp-chatbots-bot-development-and-ai-integration. Accessed: 2026-01-09.

Linn Van Dyne, Soon Ang, and Christine Koh. 2015. Development and validation of the cqs: The cultural intelligence scale. In *Handbook of cultural intelligence*, pages 34–56. Routledge.

Hanna Wallach, Meera Desai, Nicholas Pangakis, A Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, and 1 others. 2024. Evaluating generative ai systems is a social science measurement challenge. *arXiv preprint arXiv:2411.10939*.

Martin Warren and William WL Lee. 2020. Intercultural communication in professional and workplace settings. In *The Routledge handbook of language and intercultural communication*, pages 473–486. Routledge.

Jincenzi Wu, Jianxun Lian, Dingdong Wang, and Helen M. Meng. 2025. SocialCC: Interactive evaluation for cultural competence in language agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33242–33271, Vienna, Austria. Association for Computational Linguistics.

Shuang Yang, Huiwen Zhao, and Wen Luo. 2024. The impact of artificial intelligence on intercultural communication. In *Belonging in Culturally Diverse Societies-Official Structures and Personal Customs*. IntechOpen.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096.

Akhila Yerukola, Saadia Gabriel, Nanyun Peng, and Maarten Sap. 2025. Mind the gesture: Evaluating AI sensitivity to culturally offensive non-verbal gestures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25041–25080, Vienna, Austria. Association for Computational Linguistics.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yi Zheng, Björn Ross, and Walid Magdy. 2023. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71.

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.