# Toward Comprehensive Cultural Alignment in LLMs: An Interdisciplinary Framework for Cultural Evaluation

**Truong Vo**[1] **Sanmi Koyejo**[2]

[1]Northwestern University
[2]Stanford University

Correspondence to: truongvo2025@u.northwestern.edu

**Abstract**

Evaluating cultural alignment in large language models (LLMs) requires a nuanced, multidisciplinary understanding of what "culture" entails. In this paper, we draw from anthropology, sociology, political science, and media theory to propose a broader conceptualization of culture as symbolic meaning-making, embodied social practice, negotiated discourse, institutional structure, strategic repertoire, algorithmic mediation, and power-laden representation. We map these conceptual perspectives to representative datasets and propose corresponding evaluation strategies to assess how LLMs handle context sensitivity, temporal change, behavioral norms, and epistemic asymmetries. By grounding evaluation in interdisciplinary cultural theories, we provide a structured framework for the design and development of comprehensive cultural alignment benchmarks for LLMs.

## Introduction

Foundational frameworks like the World Values Survey (Inglehart et al. 2014) and Values Survey Module (Hofstede 2001) reduce culture to static, measurable beliefs. This view, common in evaluating large language models (LLMs), overlooks culture's symbolic depth, temporal dynamics, contextual variability, and power relations. As LLMs increasingly mediate cross-cultural communication, we propose a framework grounded in interdisciplinary theories of culture. Drawing on seven conceptual lenses, we map each to concrete evaluation tasks and datasets (Table 1), offering a structural foundation for assessing cultural competence in LLMs across social and digital contexts.

## Operationalization of Concepts

In this paper, we develop a robust, multi-dimensional assessment framework designed to capture the complex, contested, and dynamic nature of culture. This framework integrates diverse evaluation extbfscenarios that reflect culture's fluid meanings, power structures, and contextual applications across social and digital environments.

### 1. System of Shared Meanings (Geertz 1973)

This dimension evaluates a model's ability to recognize and interpret culturally significant symbols. These symbols often carry different meanings within and across cultural contexts, requiring models to both contextualize and differentiate their interpretations.

- **Task 1: Within-Cultural Symbol Recognition** tests whether models can correctly interpret culturally grounded symbols within a specific context. *Example:* Can the model recognize that white symbolizes purity in Western weddings but mourning in East Asian funerals?
- **Task 2: Cross-Cultural Symbol Navigation** assesses whether models can distinguish conflicting meanings of the same symbol across cultures. *Example:* Can the model interpret a thumbs up as approval in the U.S. but offensive in parts of the Middle East?

### 2. Discourse and Practices (Bourdieu 1977)

This dimension focuses on a model's ability to apply socially embedded norms and values in practical, everyday scenarios, acknowledging that culture is enacted through behavior rather than merely articulated.

- **Task 1: Practical Norm Application** tests whether models can apply culturally grounded social norms in situated contexts. *Example:* In a Japanese business meeting, how should one respond when a supervisor makes a suggestion you disagree with?
- **Task 2: Contextual Appropriateness of Behavior** evaluates a model's ability to demonstrate cultural coherence and appropriateness across different social fields. *Example:* During a dinner at her Egyptian friend's home, Anna waits for the host to serve her before eating. Is this socially appropriate?

### 3. Dynamic and Negotiated Practice (Clifford 1988)

This dimension evaluates whether a model can account for cultural changes over time and recognize variation within the same cultural group, particularly across generations.

- **Task 1: Adapting to Cultural Change** assesses whether models can recognize and reflect historical and temporal shifts in cultural practices. *Example:* How have Japanese workplace communication norms shifted from the 1980s to today?
- **Task 2: Generational Variation Recognition** examines whether models can identify intra-cultural differences

| Culture Concept | Key Evaluation Tasks | Representative Datasets |
|---|---|---|
| **System of Shared Meanings** | (1) Within-cultural symbol recognition and interpretation.<br>(2) Cross-cultural understanding of symbolic meanings, gestures, and metaphors. | eHRAF (ethnographic data, 360 societies); MAPS (2.3k proverbs, 6 languages); Color-Emotion (711 participants, 4 countries). |
| **Discourse and Practice** | (1) Applying social norms in everyday interaction scenarios.<br>(2) Generating culturally appropriate behaviors across diverse social contexts. | NormAD (2.6k stories, 75 countries); TalkBank; WVS/VSM; Santa Barbara Corpus. |
| **Dynamic and Negotiated Practice** | (1) Recognizing and adapting to temporal cultural shifts and evolving norms.<br>(2) Modeling intra-group and generational variation within the same culture. | WVS longitudinal (1981–2022); Pushshift Reddit (5.6B comments); CultureTrack (2001–2017). |
| **Repertoires** | (1) Selecting appropriate cultural tools across social roles and audiences.<br>(2) Reasoning and adapting communicative strategies across contexts. | Wikipedia Talk (166k threads); YouTube transcripts (1.8M videos); Reddit cross-community data. |
| **Institutions and Norms** | (1) Navigating institutional rules and social contracts within cultural contexts.<br>(2) Understanding culturally specific enforcement and compliance mechanisms. | MCWC (223 constitutions); HSE-Bench (IRAC legal QA); Social Norms Dataset (12k questions). |
| **Algorithmic Mediation** | (1) Recognizing platform-specific communicative norms and user practices.<br>(2) Understanding algorithmic influences on cultural evolution and discourse. | TikTok cultural data; YouTube longitudinal content; Media Cloud (50k+ sources, 20+ languages). |
| **Critical and Decolonial Perspectives** | (1) Identifying dominant vs. marginalized cultural representations.<br>(2) Evaluating representational justice and epistemic inclusion in outputs. | Masakhane (African NLP); AI4Bharat (22 Indian languages); CulturePark (41k dialogues). |

Table 1: Structured Framework Linking Cultural Theories to Evaluation Tasks and Datasets in LLMs.

across age groups and cohorts. *Example:* Do Korean attitudes toward mental health differ between older and younger generations, and can the model reflect these differences?

## 4. Repertoire (Swidler 1986)

Culture functions as a "toolkit" of habits, symbols, and strategies that individuals draw on contextually. This dimension evaluates a model's ability to reason about and strategically select appropriate cultural tools based on situational demands.

- **Task 1: Strategic Toolkit Deployment** tests whether models can adapt communicative or behavioral strategies based on audience and context. *Example:* A software engineer must present to technical peers, executives, and clients. How should their communication differ across these audiences?
- **Task 2: Contextual Reasoning for Adaptation** assesses whether models understand how different social fields require distinct cultural strategies. *Example:* A community leader must alternate between formal protocols during city meetings and informal relationship-building at local events. Can the model reflect this shift?

## 5. Institutional Norms and Rules (North 1990)

Culture operates within and through institutional frameworks. This dimension evaluates models' capacity to interpret both formal regulations and informal enforcement mechanisms that structure cultural behavior.

- **Task 1: Institutional Rule Navigation** tests whether models can follow institutional norms and formalized cultural expectations. *Example:* An employee discovers supervisor misconduct in a setting where corruption is culturally normalized. How should they respond according to institutional rules?
- **Task 2: Culturally Situated Enforcement Understanding** evaluates a model's grasp of how sanctions and compliance differ across contexts. *Example:* Public criticism of norm violations is rare in Japan but common in Australia. How should workplace misconduct be addressed in each context?

## 6. Algorithmic Mediation (Seaver 2017)

Digital platforms not only mediate culture—they actively construct it through algorithmic design and recommendation systems. This dimension assesses whether models can recognize how algorithmic structures shape language, norms, and discourse online.

- **Task 1: Platform-Specific Norm Recognition** tests whether models can identify distinct communicative styles and expectations across platforms. *Example:* How should a brand respond to criticism on Twitter versus TikTok?
- **Task 2: Understanding Algorithmic Influence** evaluates awareness of how algorithmic changes transform cultural discourse. *Example:* Facebook introduced engagement-based ranking in 2009. How did this affect political discourse from 2009–2015?

### 7. Critical and Decolonial Perspectives (Hall 1997)

This dimension centers the politics of representation. It evaluates whether models can identify which cultural perspectives are amplified or marginalized and whether they reproduce or resist epistemic dominance.

- **Task 1: Detecting Representational Bias** assesses whether models can identify whose cultural narratives are prioritized in knowledge production. *Example:* Given the prompt "Describe the history of feminism," does the model center Western liberal feminism or include postcolonial, Indigenous, and Global South perspectives?
- **Task 2: Power-Conscious Framing of Knowledge** evaluates whether models can recognize and reframe outputs that reflect dominant power structures. *Example:* Can the model identify when seemingly neutral responses privilege certain cultural worldviews and offer more inclusive alternatives?

## Key Takeaways

We present a theoretically grounded framework that translates seven major conceptualizations of culture into specific evaluation tasks and datasets for LLM assessment. This framework enables:

1. **Theory-driven benchmark design**: Systematic mapping from cultural theory to measurable tasks ensures evaluations capture symbolic interpretation, embodied practice, temporal change, strategic adaptation, institutional navigation, algorithmic mediation, and representational justice.
2. **Beyond static metrics**: The framework requires models to demonstrate cultural competence as dynamic capability, recognizing contested meanings, adapting to temporal shifts, and navigating intra-group variation, rather than reproducing fixed stereotypes.
3. **Power-conscious evaluation**: By operationalizing critical and decolonial perspectives, the framework provides concrete methods for assessing epistemic inclusion and identifying whose cultural knowledge is privileged or excluded in model outputs.

Future work should develop standardized benchmarks implementing these task categories, establish evaluation protocols across the proposed datasets, and investigate how different training paradigms affect performance across these theoretically distinct dimensions of cultural alignment.

## References


Bourdieu, P. 1977. *Outline of a Theory of Practice*. Cambridge University Press.

Clifford, J. 1988. *The Predicament of Culture: Twentieth-Century Ethnography, Literature, and Art*. Harvard University Press.

Geertz, C. 1973. *The Interpretation of Cultures*. Basic Books.

Hall, S. 1997. *Representation: Cultural Representations and Signifying Practices*. Sage.

Hofstede, G. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. Thousand Oaks, CA: Sage Publications, 2nd edition. ISBN 978-0803973244.

Inglehart, R.; Haerpfer, C.; Moreno, A.; Welzel, C.; et al. 2014. *World Values Survey: Round Six - Country-Pooled Datafile 2010-2014*. JD Systems Institute.

North, D. C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press.

Seaver, N. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, 4(2): 1–12.

Swidler, A. 1986. Culture in action: Symbols and strategies. *American Sociological Review*, 51(2): 273–286.