Incorporating Diverse Perspectives in Cultural Alignment: Survey of Evaluation Benchmarks Through A Three-Dimensional Framework

Meng-Chen Wu, Si-Chi Chin, Tess Wood, Ayush Goyal, Narayanan Sadagopan
Amazon.com
{mengchw, sichi, tesswoo, ayushg, sdgpn}@amazon.com

Abstract

Large Language Models (LLMs) increasingly serve diverse global audiences, making it critical for responsible AI deployment across cultures. While recent works have proposed various approaches to enhance cultural alignment in LLMs, a systematic analysis of their evaluation benchmarks remains needed. We propose a novel framework that conceptualizes alignment along three dimensions: Cultural Group (who to align with), Cultural Elements (what to align), and Awareness Scope (how to align: majority-focused vs. diversity-aware). Through this framework, we analyze 105 cultural alignment evaluation benchmarks, revealing significant imbalances: Region (37.9%) and Language (28.9%) dominate Cultural Group representation; Social and Political Relations (25.1%) and Speech and Language (20.9%) concentrate Cultural Elements coverage; and an overwhelming majority (97.1%) of datasets adopt majority-focused Awareness Scope approaches. In a case study examining AI safety evaluation across nine Asian countries (Section 5), we demonstrate how our framework reveals critical gaps between existing benchmarks and real-world cultural biases identified in the study, providing actionable guidance for developing more comprehensive evaluation resources tailored to specific deployment contexts.

1 Introduction

LLMs have become increasingly prevalent in our daily lives, powering applications ranging from virtual assistants and content generation to educational tools and business solutions. Their global deployment has created unprecedented opportunities for cross-cultural interaction and knowledge exchange. However, this widespread adoption also brings significant challenges regarding cultural sensitivity and appropriateness of AI-generated content (Hershcovich et al., 2022). As these models serve diverse users across different cultural contexts, their ability to understand and respect cultural nuances becomes critical for responsible AI deployment.

Aligning AI systems with diverse cultural contexts is critical in their development and deployment. Recent research and real-world incidents highlight the potential consequences of culturally misaligned AI outputs. For instance, LLMs have been criticized for exhibiting Western-centric biases when responding to diverse cultural contexts (Cao et al., 2023; Durmus et al., 2023; Naous et al., 2024), while machine translation systems have faced challenges with culturally-specific terms and expressions (Yao et al., 2024a; Cao et al., 2024c; Naveen and Trojovský, 2024). In business contexts, culturally inappropriate AI-generated marketing content has led to public relations issues and damaged brand reputation (Coscarelli, 2022). These cases underscore how cultural misalignment can lead to misunderstandings, offense, or even harm to specific cultural groups.

Despite these growing concerns, detecting and measuring cultural misalignment presents significant challenges, particularly in capturing the full spectrum of cultural perspectives. While large-scale cultural surveys like the PEW Global Attitudes Survey (Pew Research Center, 2014) and World Values Survey (WVS) (Haerpfer et al., 2022) capture rich distributions of responses across demographic groups, revealing significant intra-cultural variations in values, beliefs, and practices, current cultural alignment research typically reduces these nuanced distributions to single majority responses when constructing evaluation datasets (AlKhamissi et al., 2024; Cao et al., 2023). For instance, studies using WVS data often consider only the most frequent response as ground truth, discarding valuable information about minority viewpoints (Tao et al., 2024; Masoud et al., 2023). This simplification, while computationally convenient, fails to capture the complexity of cultural nuances and may perpetuate the marginalization of minority perspectives. The nascent nature of cultural alignment evaluation, combined with the rapid development of LLMs, highlights the urgent need for more systematic and comprehensive evaluation strategies that can account for both majority and diverse perspectives.

This paper presents, to our knowledge, the first comprehensive analysis of cultural alignment evaluation benchmarks, examining 105 datasets through our proposed framework. This analysis not only reveals the current state of cultural alignment evaluation but also helps understand the coverage and limitations of existing benchmarks. Our key contributions are:

- We propose a novel cultural alignment framework that conceptualizes alignment along three dimen-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 17038–17073
November 4-9, 2025 ©2025 Association for Computational Linguistics

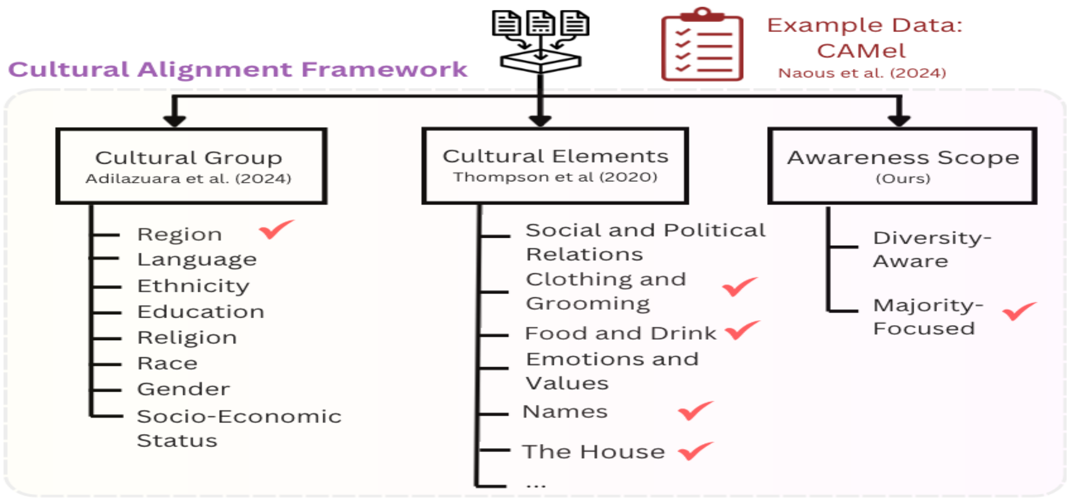

Figure 1: Overview of our cultural alignment framework. The framework categorizes alignment along three dimensions: Cultural Group (who to align with, e.g., Language in CAMel Naous et al. (2024)), Cultural Elements (what to align, e.g., Clothing, Food, Names in CAMel), and Awareness Scope (how to align, e.g., majority-focused in CAMel). This categorization enables systematic analysis of how datasets address cultural alignment across different dimensions.

sions: Cultural Group (who to align with), Cultural Elements (what to align), and Awareness Scope (how to align) (section 2.2).

- We identify significant imbalances in current evaluation practices: region and language dominate Cultural Group representation (37.9% and 28.9%), Social and Political Relations (25.1%) and Speech and Language (20.9%) dominate Cultural Elements coverage, and for Awareness Scope, an overwhelming majority (97.1%) of datasets adopt majority-focused approaches (sections 4.1, 4.2, 4.3). These imbalances risk perpetuating existing biases and failing to ensure comprehensive and fair evaluation across different cultural contexts.

- Through systematic analysis of current evaluation benchmarks, we identify critical directions for advancing cultural alignment: (1) developing application-specific benchmarks guided by our framework when existing datasets provide insufficient coverage; (2) expanding dataset coverage beyond majority-focused approaches and incorporating temporal dynamics; (3) implementing diversity-aware alignment techniques with specialized loss functions and reward models; (4) exploring synthetic data generation strategies (section 6.2).

2 Cultural Alignment Framework

2.1 Definition of Culture

Culture is commonly understood as shared norms, values, and behaviors within a particular group, yet this understanding is an oversimplification. Anthropological perspectives have evolved significantly, from Tylor (1871)'s definition of culture as "that complex whole" learned through social interaction, to more recent views by Anderson-Levitt (2012) emphasizing culture as a process of meaning-making. These perspectives highlight three crucial aspects: culture is social and constantly evolving, emerges from interactions between individuals and groups, and is intertwined with power dynamics. However, as Adilazuarda et al. (2024) note, most NLP research examining culture often leaves it undefined beyond the high level, with cultural features being implicitly specified through dataset choices rather than explicit definition.

These varying definitions of culture have significant implications for cultural alignment in LLMs. A simplified definition that treats culture as a fixed set of values and behaviors risks promoting monolithic views and reinforcing stereotypes. Understanding culture as dynamic, interactive, and power-laden instead suggests that cultural alignment efforts should account for three key factors: variations within cultural groups, influences between different cultures, and the role of power group in shaping cultural practices. This nuanced understanding cautions against uncritically aligning with majority cultural practices and emphasizes the importance of recognizing cultural complexity in alignment efforts.

2.2 Framework Overview

Cultural alignment for LLMs presents a complex challenge that requires careful consideration of multiple dimensions. As shown in Figure 1, we propose a framework that conceptualizes cultural alignment along three primary dimensions: Cultural Group, Cultural Elements, and Awareness Scope. Building upon the cultural proxies taxonomy established by Adilazuarda et al. (2024), our framework adopts their classification of demographic and semantic proxies for the first two di-

Paper	# Papers Surveyed	Framework Dimensions	Cultural Group #	Cultural Elements #	Awareness Scope
Liu et al. (2025)	127	1	-	11*	-
Adilazuarda et al. (2024)	90	2	7	13 / 22§	-
Ours	105	3	12	16 / 25§	✓

Table 1: Comparison of our three-dimensional framework with existing cultural alignment surveys. *Not based on Thompson et al. (2020)'s semantic domains. § Elements found in datasets / Total elements defined.

mensions, while introducing Awareness Scope as a new dimension. Each dimension addresses a critical aspect of cultural alignment, enabling researchers to precisely define alignment targets and develop more culturally aware AI systems.

1. **Cultural Group** defines the target community for alignment. This dimension operates through demographic indicators[1] including Region, Language, Ethnicity, Religion, Gender, Socioeconomic Status, and Online community, either individually or through their intersections (e.g., urban middle-class women). These indicators provide concrete parameters for developing culturally aligned systems in specific contexts, such as aligning an LLM with Southeast Asian communities or Vietnamese-speaking populations.

2. **Cultural Elements** represent the specific aspects of culture to be aligned. These encompass both tangible manifestations (e.g., food, clothing, and arts) and intangible constructs (e.g., values, social norms, customs), and are grounded in the cognitive anthropological view that culture is structured through semantic domains, folk taxonomies, and shared mental models (d'Andrade, 1995). For example, when developing a culturally aligned LLM, developers might focus on aligning the model's responses with specific cultural aspects such as food customs, social norms, or value systems.[2]

3. **Awareness Scope** describes the breadth of cultural perspectives to be incorporated in alignment. This dimension distinguishes between majority-focused approaches that address dominant cultural patterns and widely accepted norms, [3] and diversity-aware approaches that explicitly incorporate subcultural variations and minority perspectives alongside dominant patterns. For instance, in developing a recipe recommendation system for Singapore, a majority-focused approach would emphasize widely recognized and prototypical culinary practices across the population, while a diversity-aware approach would explicitly incorporate the distinct food traditions of Malay, Indian, and other minority communities as integral components rather than peripheral considerations.

By integrating these three complementary perspectives, our framework addresses critical gaps in existing cultural alignment evaluation taxonomies. As shown in Table 1, previous cultural alignment surveys by Adilazuarda et al. (2024) and Liu et al. (2025) have not considered the Awareness Scope dimension, despite its crucial role in capturing the full spectrum of cultural representation. By incorporating this new dimension with enhanced Cultural Group and Cultural Elements taxonomies, our framework offers a more complete method to analyze cultural alignment evaluation benchmarks that acknowledges both majority perspectives and diverse subcultural variations.

2.3 Awareness Scope

We ground the Awareness Scope dimension in standpoint theory, a sociological framework developed within feminist epistemology (Harding, 1991; Collins, 2022). Standpoint theory posits that knowledge is fundamentally shaped by one's social position, with marginalized groups offering distinct and valuable perspectives that are systematically overlooked when dominant viewpoints are privileged as universal truths. This theoretical lens directly informs our framework's central distinction: majority-focused approaches, while capturing representative cultural patterns, risk perpetuating what standpoint theorists call epistemic injustice—the systematic exclusion of minority knowledge and perspectives. In contrast, diversity-aware approaches align with standpoint theory's call for incorporating multiple social positions to achieve more complete and equitable knowledge representation. Therefore, our Awareness Scope dimension reflects a fundamental commitment to

[1]We extend the demographic proxies from Adilazuarda et al. (2024) to include appearance, sexual orientation, health, socioeconomic status, online communities, and time periods (e.g., dynasties) to better accommodate our dataset analysis.

[2]From Thompson et al. (2020), semantic domains include: food and drink, clothing and grooming, speech and language, modern world, physical world, the house, basic actions and technology, agriculture and vegetation, animals, social and political relations, emotions and values, kinship, and cognition, with names from Adilazuarda et al. (2024). We extend this list to include arts, mental perception, and current and historical events based on our dataset findings.

[3]These can be understood as prototypical representations of cultural knowledge (cf. Rosch and Mervis 1975; see also Shore 1996). We also recognize that some papers we classified as majority-focused do consider multiple annotator perspectives and opinions. Our distinction is not based on majority voting but rather on whether the approach primarily targets prototypical cultural characteristics versus explicitly incorporating a broader range of cultural variation within a group.

epistemic justice, ensuring AI systems recognize and validate diverse cultural standpoints rather than defaulting to dominant perspectives.

The fundamental distinction between majority-focused and diversity-aware approach isn't merely the amount of detail provided, but rather their underlying recognition of cultural complexity. While majority-focused approach provides clear, standardized guidance essential for basic cultural compliance, the diversity-aware approach acknowledges the coexistence of multiple valid cultural practices, demonstrate sensitivity to user identity and context, and support navigation of complex cultural landscapes. For a detailed discussion on choosing between majority-focused and diversity-aware approaches for different use cases, see Appendix E.

2.4 Awareness Scope Examples

To demonstrate how awareness scope impacts responses while keeping other dimensions constant, we present two illustrative scenarios. Table 4 in Appendix D shows detailed responses for each example.

Example 1: Tour Guide Agent In the context of providing information about alcohol in Saudi Arabia, the two responses demonstrate distinct approaches to cultural alignment. The majority-focused response adheres to the dominant cultural framework, emphasizing the general prohibition under Islamic law and suggesting culturally acceptable alternatives. In contrast, the diversity-aware response acknowledges both traditional restrictions and emerging changes—including licensed venues in Riyadh's diplomatic quarter for non-Muslim diplomats and potential policy adjustments under Vision 2030 for tourists and expatriates[4]. The diversity-aware approach thus captures both traditional norms and evolving cultural dynamics in the country.

Example 2: Localization Agent For a clock manufacturer entering the Taiwanese market, the responses illustrate different depths of cultural awareness. The majority-focused response emphasizes the dominant cultural taboo around clocks[5], presenting it as a universal constraint that directly dictates marketing strategy. It provides a single, straightforward solution avoiding any gift-giving associations. The diversity-aware response, however, acknowledges the traditional taboo while recognizing the heterogeneity within Taiwanese society. It segments the market based on different demographic groups and acknowledges varying degrees of adherence to traditional beliefs, demonstrating awareness of both traditional sensitivities and evolving cultural dynamics.

Understanding these different use cases enables more thoughtful design of culturally aligned AI systems. The choice between majority-focused and diversity-aware

[4]https://www.eiu.com/n/saudi-arabia-takes-first-step-towards-relaxing-alcohol-laws/

[5]In Taiwanese and Chinese cultures, clocks are considered inappropriate gifts because the phrase "giving a clock" is homophonous with "sending someone off to their death", making it a significant cultural taboo.

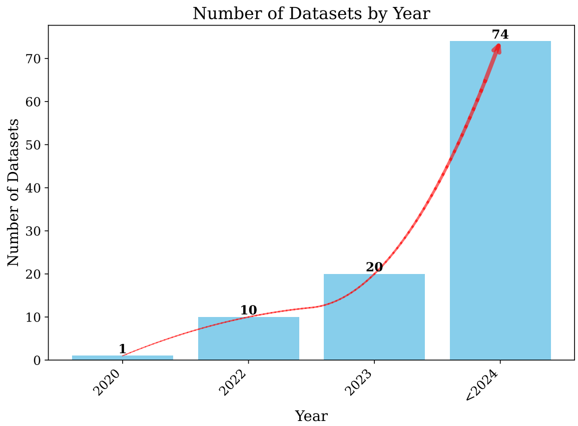

Figure 2: Growth of cultural alignment evaluation datasets from 2020 to 2024, with an increase from 1 dataset in 2020 to 74 in 2024 and beyond.

approaches represents a critical design decision that shapes how an AI system engages with cultural complexity. This decision should be made early in the development process and should inform both the training and evaluation of these systems.

3 Method

We employ a three-step methodology: (1) dataset collection through a systematic literature review, (2) dataset categorization based on our proposed cultural alignment framework, and (3) summary analysis of findings. This structured approach enables us to identify key patterns, gaps, and trends in how existing datasets approach cultural alignment evaluation.

3.1 Dataset Collection

We conduct a systematic literature review primarily utilizing ACL Anthology[6] and Google Scholar[7] as our main sources. Our search employs various combinations of keywords including "Culture", "LLM", "Cultural Alignment", "Cultural Awareness", "Cultural Bias", "NLP", "Cultural Commonsense Knowledge", "Multi-Cultural", "Cross-Cultural", and "Evaluation Benchmark". To maintain focus on language model evaluation, we survey papers that introduce text-only evaluation datasets, explicitly excluding papers describing multimodal datasets involving images or videos. Through this systematic search and screening process, we identify 104 papers describing 105 datasets for evaluating cultural awareness in language models.

The temporal analysis, as shown in Figure 2, reveals a substantial increase in dataset creation, with 10 datasets in 2022, 20 datasets in 2023, and 74 datasets in 2024 and beyond, reflecting the growing recognition of cultural alignment as a crucial aspect of LLM development.

[6]https://aclanthology.org/

[7]https://scholar.google.com/

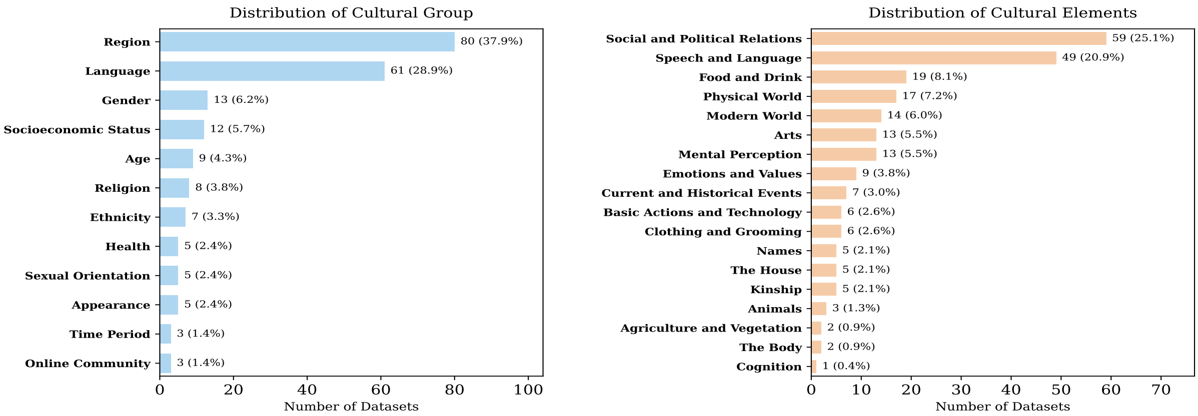

Figure 3: Distribution analysis of over 105 datasets of two dimensions: (left) Cultural Group with Region (37.9%) and Language (28.9%) being the most common; and (right) Cultural Elements dominated by Social and Political Relations (25.1%) and Speech and Language (20.9%).

3.2 Dataset Categorization

Following our proposed cultural alignment framework presented in Section 2.2, we analyze and categorize the datasets based on their descriptions in the papers across three dimensions: Cultural Group, Cultural Elements, and Awareness Scope. For the first two dimensions, we examine the dataset descriptions in each paper, specifically on details about the demographic indicators and cultural aspects covered. For the Awareness Scope dimension, we analyze the evaluation metrics[8] described in the papers to determine whether they adopt a majority-focused or diversity-aware approach.

Two co-authors independently label each dataset according to these dimensions using a multi-class tagging approach (selecting all applicable categories for Cultural Group and Cultural Elements, while making binary classifications for Awareness Scope)[9]. To ensure annotation quality, we implement a two-stage resolution process where annotators first adjudicate disagreements through discussion and mutual agreement. For cases where consensus cannot be reached, a third co-author makes the final determination.

3.3 Analysis

Through the lens of our proposed cultural alignment framework, we analyze our findings in Section 4. We combine frequency distribution analysis and chi-square (χ^2) tests to identify both descriptive patterns and statistically significant relationships between framework dimensions, revealing how cultural groups, elements, and awareness scope systematically interact. These insights reveal current limitations and opportunities in cultural alignment evaluation for LLMs.

4 Findings

This section analyzes existing cultural alignment evaluation datasets across our three framework dimensions, identifying key patterns and gaps in current approaches.

4.1 Cultural Group

Our analysis of Cultural Group reveals that Region (37.9%)[10] and Language (28.9%) serve as the predominant demographic indicators in current datasets (Figure 3, left)[11]. Other indicators show significantly lower representation: Gender (6.2%), Socioeconomic Status (5.7%), Age (4.3%), Religion (3.8%), and Ethnicity (3.3%). When present, these additional indicators typically function as secondary factors rather than primary axes of cultural variation, with multi-dimensional cultural group definitions remaining rare.

This focus on Region and Language presents a critical limitation, potentially masking cultural variations shaped by other demographic factors. For example, dietary preferences evaluated solely by region overlook how religion and socioeconomic status influence food choices within the same area. Future datasets should incorporate multiple demographic indicators simultaneously for more comprehensive cultural alignment evaluation across diverse populations.

4.2 Cultural Elements

Our analysis reveals an uneven distribution of Cultural Elements across evaluation datasets (Figure 3, right). Social and Political Relations (25.1%) and Speech and Language (20.9%) dominate, while Food and Drink (8.1%), Physical World (7.2%), and Modern World

[8] Majority-focused approaches typically employ accuracy-based metrics that assess alignment with dominant or prototypical cultural patterns; diversity-aware approaches typically employ distributional (distance-based) metrics, such as Jensen–Shannon divergence, to capture variation across cultural perspectives.

[9] See Appendix B for annotation details and Appendix K for complete dataset annotations)

[10] Percentages represent the proportion of datasets containing each category relative to the total number of datasets (105). Since individual datasets can include multiple categories simultaneously, the sum of these percentages exceeds 100%.

[11] See Figure 6 and 7 in Appendix K for the complete hierarchical taxonomy of Cultural Group and Cultural Elements.

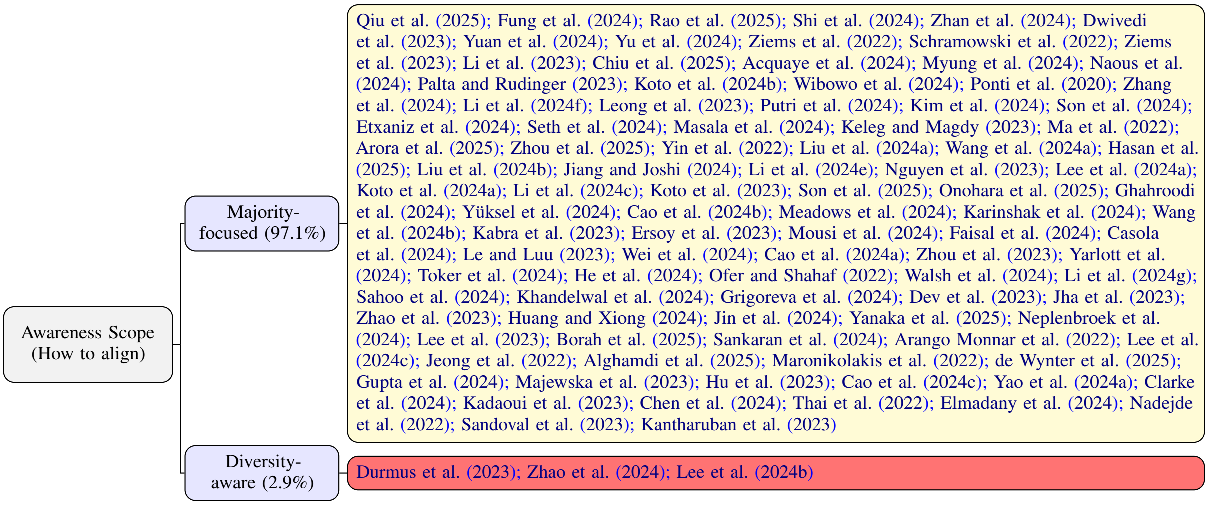

Figure 4: Awareness Scope dimension of our cultural alignment framework with corresponding evaluation datasets.

Category	Maj.	Div.	Category	Maj.	Div.
Cultural Group			***Cultural Elements***		
Region	77	3	Soc & Pol Relations	59	0
Language	60	1	Speech & Language	49	0
Gender	13	0	Food & Drink	19	0
Socioeconomic Status	11	1	Physical World	17	0
Age	9	0	Modern World	14	0
Religion	8	0	Arts	13	0
Ethnicity	7	0	Mental Perception	13	0
Health	5	0	Emotions & Values	6	3
Sexual Orientation	5	0	Current/Hist Events	7	0
Appearance	5	0	Basic Actions & Tech	6	0
Time Period	3	0	Clothing & Grooming	6	0
			Names	5	0
			The House	5	0
			Kinship	5	0

Table 2: Distribution of majority-focused (Maj.) and diversity-aware (Div.) approaches across framework dimensions. Diversity-aware evaluation appears in only three Cultural Group categories (Region: 3, Language: 1, Socioeconomic Status: 1) and one Cultural Elements category (Emotions and Values: 3), highlighting the limited diversity consideration in evaluations.

(6.0%) receive moderate coverage. Most other domains appear in 5 or fewer datasets despite their cultural significance in specific communities.

This imbalance presents both opportunities and challenges. While the strong representation of social and linguistic domains provides robust evaluation in these areas, the limited coverage of other cultural elements restricts comprehensive assessment of model alignment capabilities. Future datasets should expand coverage in underrepresented domains to provide more balanced cultural evaluation across the full spectrum of cultural practices and knowledge.

4.3 Awareness Scope

Our analysis reveals a significant imbalance in awareness scope (see Figure 4): among 105 datasets surveyed, 97.1% (102 datasets) adopted a majority-focused approach, while only 2.9% (3 datasets) address diversity-aware evaluation. This disparity represents a critical gap in the evaluation landscape, potentially undermining LLMs' ability to serve diverse subcultural groups.

The interaction between awareness scope and other framework dimensions reveals concerning patterns (Table 2). Despite various demographic indicators being available, diversity-aware approaches remain rare across all Cultural Groups. Among datasets defining cultural groups through Region and Language (80 and 61 datasets), only 3 and 1 incorporate diversity-aware approaches, while just 1 of 12 Socioeconomic Status datasets does so. Cultural elements like Social and Political Relations and Speech and Language are evaluated exclusively from majority-focused perspectives. Emotions and Values is the only element incorporating diversity-aware approaches through GlobalOpinionQA (Durmus et al., 2023), WorldValuesBench (Zhao et al., 2024), and KorNAT (Lee et al., 2024a). This concentration in a single cultural element highlights critical gaps in current evaluation practices (see Appendix F for detailed discussion of how our framework can generalize across diverse Cultural Groups and Elements).

This systematic majority-focus raises concerns about evaluation fairness. Majority-focused approaches capture dominant patterns but marginalize subcultural variations, limiting LLMs' ability to serve diverse populations. Singapore's food culture illustrates this: majority-focused evaluation overlooks crucial dietary requirements (halal, vegetarian) and preferences across ethnic communities and generations—cultural differences invisible when focusing solely on majority views.

4.4 Statistical Associations Between Framework Dimensions

Beyond the individual distribution patterns, we conducted chi-square (χ^2) tests to examine systematic associations between our framework dimensions. Table 3 presents the highly significant associations from this analysis. Our findings reveal four key patterns: First, Emotions and Values emerged as the only cultural do-

Dimension Association	Variable Intersection	Stat.	Value
Cultural Group × Cultural Elements (χ^2 = 136.40, df = 56, p < 0.001)	(Gender, Mental Perception)	χ^2	**43.46**
		p-value	**<0.001**
	(Socioeconomic Status, Mental Perception)	χ^2	**25.20**
		p-value	**<0.001**
	(Religion, Mental Perception)	χ^2	**20.33**
		p-value	**<0.001**
Cultural Elements × Awareness Scope ($\chi^2 = 163.71$, df = 17, p < 0.0001)	(Emotions and Values, Diversity-aware)	χ^2	**156.82**
		p-value	**<0.001**

Table 3: Highly significant associations (p < 0.001) from chi-square analysis of framework dimensions. Left column shows overall dimension statistics; right columns show specific category associations.

main significantly associated with diversity-aware approaches (χ^2 = 156.82, p < 0.001), exposing a fundamental dichotomy where only values receive recognition for intra-cultural diversity while all other cultural expressions are treated as homogeneous. Second, demographic indicators (Gender, Socioeconomic Status, and Religion) showed strong associations with Mental Perception (all p < 0.001), indicating these groups are primarily evaluated through stereotypes rather than everyday practices. Third, Language showed significant association with Speech and Language elements (χ^2 = 4.16, p < 0.05, see Table 5 in Appendix G), confirming language-defined groups are disproportionately evaluated on linguistic features. Fourth, no demographic indicators showed significant associations with material culture elements, suggesting consistent underrepresentation of tangible cultural domains.

The chi-square test for Cultural Group and Awareness Scope showed no significant overall association (χ^2 = 4.80, df = 4, p = 0.308), despite a marginally significant association between Socioeconomic Status and diversity-aware approaches (see Table 7 in Appendix G). This indicates that majority-focus approaches represent a universal methodological limitation rather than a bias toward specific cultural indicators. These patterns highlight a critical finding: while diversity in values is recognized, diversity in all other cultural expressions is systematically ignored regardless of demographic group. Complete statistics for all dimension pairings are provided in Appendix G.

5 Case Study: Applicability of Our Framework

Identification of Cultural Alignment Evaluation Gap The recent study conducted by Infocomm Media Development Authority (IMDA) and Humane Intelligence (2025) demonstrates how deployed LLMs continue to exhibit cultural biases, where 362 domain experts evaluated four LLMs across nine Asian countries[12] and identified 3,222 instances of cultural bias affecting real users, with gender, geographical, and socioeconomic biases emerging consistently throughout all surveyed nations. Take findings in Singapore for examples: in one case, an LLM claimed women between ages 20-40 are more vulnerable to online scams than men, suggesting they have a higher likelihood of responding to emails offering job opportunities or financial gain. In another instance, when questioned about crime rates in Singapore's ethnic enclaves, an LLM stereotyped areas like Chinatown and Little India as having higher crime rates, attributing this to supposed limited interaction between immigrant communities and the native population.

Our analysis reveals that among the 105 datasets surveyed, only six address the Singaporean context (see Table 8 in Appendix J). These datasets fail to cover the specific biases identified in the IMDA study, with none evaluating Gender, Socioeconomic Status, or Ethnicity dimensions—categories where significant biases were found. Within Cultural Elements, none evaluate Mental Perception, which is essential for detecting the stereotyping that dominated the IMDA findings. Critically, our framework shows the IMDA study, along with all six datasets, rely exclusively on majority-focused evaluation methods, overlooking how LLMs might produce different stereotypical responses in personalized settings. This example demonstrates the practical value of our framework: by analyzing evaluation benchmarks across all three dimensions, researchers can identify critical coverage gaps before deployment, enabling development of more comprehensive benchmarks that can detect harmful biases before models reach users.

Improvement of Diversity-awareness To further demonstrate our framework's practical value, we conducted an experiment using a subset of GlobalOpinionQA (Durmus et al., 2023), focusing on Southeast Asian countries—the region where the IMDA study identified significant cultural biases. Results showed that diversity-aware prompting based on our framework's Awareness Scope dimension reduced Jensen-Shannon distance by 3.97% compared to the baseline prompting. This offers a straightforward approach to help the model better represent the diverse perspectives within a given country. Full experimental details are provided in Appendix I.

[12]Countries include China, India, Indonesia, Malaysia, Japan, Singapore, South Korea, Thailand, and Vietnam.

6 Summary of Limitations and Future Direction

The limitations identified in current cultural alignment evaluation datasets raise significant concerns about their broader implications for AI development and deployment. Of particular concerns are how these limitations can reinforce existing societal power structures while oversimplifying complex cultural realities.

6.1 Impact of Current Dataset Limitations

Reinforcement of Dominant Groups The predominance of majority-focused evaluation (97.1%) risks reinforcing existing power structures in AI systems. When LLMs are evaluated primarily on alignment with dominant cultural perspectives, they may perpetuate the marginalization of minority groups. For example, evaluating Singaporean food knowledge by focusing solely on Chinese dishes might lead AI systems to overlook important Malay and Indian culinary traditions. This creates a feedback loop where AI systems become increasingly optimized for majority perspectives while failing to adequately serve diverse subcultural groups.

Risk of Oversimplification Majority-focused evaluation leads to oversimplification of complex cultural dynamics. When alignment is evaluated primarily through dominant perspectives, rich variations in practices are reduced to simplified representations. In Singapore's linguistic context, datasets might present a standardized version of communication, missing how language varies—from formal English to Singlish to dialect expressions. Such oversimplification can lead to harmful misrepresentation of diverse perspectives.

6.2 Future Directions

Application-Specific Cultural Alignment A critical insight from our analysis is that cultural alignment requirements vary significantly across different applications and use cases. Future evaluation benchmarks should be developed with specific applications in mind, guided by our framework's three dimensions. This involves identifying the relevant demographic indicators for Cultural Group (e.g., Region, Age, Socioeconomic Status), determining the appropriate Cultural Elements for the application domain (e.g., consumer behaviors for retail applications, communication styles for customer service), and deciding on the necessary Awareness Scope based on the target user population's diversity. However, it is important to note that alignment methods can vary across NLP tasks even when targeting identical Cultural Group, Cultural Element, and Awareness Scope configurations. For example, evaluating food preferences among multilingual users with a diversity-aware scope may involve collecting structured preference data for a recommender system, whereas a conversational system may require culturally varied dialogue responses.

Cultural Alignment and Personalization An important consideration for future cultural alignment research is the relationship between cultural alignment and personalization goals in AI systems. Rather than viewing these as competing objectives, we argue that cultural alignment serves as a necessary foundation for ensuring fairness and inclusivity in language technologies. Without proper cultural alignment, personalization systems—especially when driven by biased or majority-skewed data—risk reinforcing dominant cultural norms while marginalizing underrepresented perspectives. Cultural alignment should establish diverse and equitable base representations, upon which personalization can be layered to tailor experiences to individual users. For example, offering diverse language options in a multilingual region reflects cultural alignment, while allowing users to select their preferred variety supports personalization. This sequential approach mitigates the risk of personalization amplifying existing cultural biases while still supporting user-centric design. Future research should explore how personalization mechanisms can be designed to respect and build upon culturally aligned foundations.

Expanding Dataset Coverage and Temporal Dynamics Our analysis reveals several critical gaps in current evaluation datasets that future development must address. Most urgently, with only 2.9% of datasets incorporating diversity-aware approaches, there is a pressing need for evaluation resources that capture intra-cultural variations and minority perspectives, particularly in applications serving diverse populations like Singapore where cultural practices vary significantly across ethnic and generational lines. Beyond awareness scope, Cultural Group coverage needs expansion beyond region (37.9%) and language (28.9%) to include other crucial demographic indicators like age, socioeconomic status, and religion. Similarly, Cultural Elements coverage should extend beyond Social and Political Relations (25.1%) and Speech and Language (20.9%) to encompass underrepresented domains like Kinship and Arts. Additionally, an often-overlooked aspect is the dynamic nature of culture itself—cultural norms, practices, and expressions evolve over time, influenced by social changes and generational shifts. Future datasets should include temporal metadata about when and under what societal conditions the data was collected, enabling evaluation against cultural expressions from the intended time period. Such comprehensive expansion across all dimensions would enable more nuanced and equitable evaluation of LLMs' cultural alignment capabilities.

Diversity-Aware Alignment Techniques Developing alignment techniques that account for cultural diversity represents another critical direction. Fine-tuning with distribution-preserving loss functions (e.g., Jensen-Shannon divergence) could maintain cultural diversity rather than defaulting to majority perspectives. Reinforcement Learning from Human Feedback (RLHF) could operationalize our framework through specialized reward models that align with each dimension—recognizing diverse Cultural Groups, evaluating engagement across Cultural Elements, and ensuring diversity-awareness in the Awareness Scope. An example in this direction is MaxMin-RLHF (Chakraborty et al., 2024), which uses multiple reward models to represent different population groups and optimizes for the minimum utility across these diverse groups, ensuring fair representation of minority perspectives. Complementary to these methods, Feng et al. (2024) proposes pluralistic alignment via multi-LLM collaboration, where community-specific language models enhance representation of underrepresented perspectives.

Potential Strategy for Mitigating Resource Limitations Given the significant gaps in current evaluation resources, developing mitigation strategies is crucial. Synthetic data generation, where LLMs create culturally-diverse evaluation scenarios under human supervision and cultural expert validation, potentially offers one promising direction. Several effective approaches have emerged: some generate synthetic cultural data from the World Values Survey by creating culture-aware question-answer pairs and culturally diverse variants (Li et al., 2024a; Xu et al., 2025), while others like (Li et al., 2024b) employ multi-agent communication frameworks to produce diverse cultural training examples. However, synthetic data generation presents different considerations across cultural and linguistic contexts. For high-resource languages and well-documented cultures where LLMs demonstrate competence, synthetic data can serve as a valuable complementary resource to expand evaluation coverage. Conversely, for underrepresented languages and cultures where LLMs exhibit poor performance, synthetic generation risks perpetuating existing biases and producing inaccurate cultural representations, potentially exacerbating rather than addressing evaluation gaps. Therefore, synthetic data should be applied strategically: as a complement to authentic data for well-represented contexts, while prioritizing community-driven data collection, cultural expert partnerships, and investment in authentic evaluation resources for underrepresented communities where synthetic approaches are insufficient.

7 Related Works

Recent works have evaluated LLMs' cultural awareness using widely-recognized cultural surveys. Studies have leveraged the World Values Survey (WVS) (Haerpfer et al., 2022) to assess cultural alignment (AlKhamissi et al., 2024; Tao et al., 2024), as well as Hofstede's cultural dimensions framework (Hofstede, 1984) for systematic evaluation (Cao et al., 2023; Masoud et al., 2025). Besides, other studies have also focused on evaluating LLMs' grasp of cultural commonsense knowledge (Myung et al., 2024; Acquaye et al., 2024; Naous et al., 2024; Chiu et al., 2025). These evaluations show LLMs align strongly with Western perspectives while struggling with other cultural contexts. For approaches to cultural alignment, see Appendix F.

A parallel stream of research has examined cultural biases and stereotypes in LLMs. Region-specific bias evaluation datasets have emerged for various languages and cultures, including IndiBias for Indian context (Sahoo et al., 2024), CHBias for Chinese (Zhao et al., 2023), and KoBBQ for Korean (Jin et al., 2024). Broader initiatives like SeeGULL (Jha et al., 2023) create stereotype datasets spanning multiple regions, while cross-cultural comparative studies (Khandelwal et al., 2024; Neplenbroek et al., 2024) highlight how social biases manifest differently across cultural contexts.

8 Conclusion

This survey presents a framework for cultural alignment evaluation in LLMs along three dimensions: Cultural Group (who to align with), Cultural Elements (what to align), and Awareness Scope (how to align). Our analysis of 105 datasets reveals critical limitations: an overwhelming bias towards majority-focused evaluation (97.1%), uneven representation across Cultural Group with dominance of Region (37.9%) and Language (28.9%), and imbalanced coverage of Cultural Elements primarily focusing on Social and Political Relations (25.1%) and Speech and Language (20.9%). Our case study demonstrates how these limitations hinder detection of real-world cultural biases. Lastly, we identify four directions—developing application-specific benchmarks, expanding coverage of diverse cultural perspectives, incorporating temporal dynamics, and exploring synthetic data generation—to ensure AI systems can effectively serve diverse cultural communities.

Limitations

Our study has two main limitations. First, while we follow Adilazuarda et al. (2024) in adopting semantic domains from Thompson et al. (2020) for Cultural Elements categorization, these conjunctive categories (e.g., "Social and Political Relations," "Emotions and Values") could benefit from more fine-grained analysis. For instance, social relations might encompass family dynamics, workplace hierarchies, and community interactions, while political relations could include governance structures, civic engagement, and political ideologies. Second, our analysis focuses exclusively on text-based cultural alignment datasets, excluding multimodal datasets that incorporate images, videos, or audio. While this scope allows for focused analysis of linguistic cultural alignment, it may not capture the full spectrum of cultural expression and understanding, as cultural practices and norms are often communicated through visual cues, gestures, and other non-textual modalities.

Ethics Statement

While this paper focuses on raising awareness of diversity within cultures and advocates for more comprehensive cultural alignment approaches, we acknowledge that some cultural practices may require careful ethical consideration when represented in AI systems. There exists an important ethical tension between respecting cultural diversity and upholding fundamental human rights principles that transcend cultural boundaries. Cultural alignment should not be pursued uncritically, particularly when certain practices may perpetuate discrimination, reinforce harmful stereotypes, or violate basic human rights. Our framework's emphasis on diversity-aware approaches aims to capture genuine cultural nuance, not to legitimize harmful practices through uncritical cultural relativism. We encourage researchers and practitioners implementing cultural alignment techniques to adopt a balanced approach that respects cultural diversity while being grounded in core ethical principles of fairness, human dignity, and non-discrimination. This ethical foundation should guide the selection and prioritization of cultural practices for inclusion in evaluation benchmarks and alignment targets, ensuring that efforts toward cultural sensitivity do not inadvertently normalize harmful perspectives simply because they exist within certain cultural contexts.

References

Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. 2023. Persianllama: Towards building first persian large language model. *arXiv preprint arXiv:2312.15713*.

Christabel Acquaye, Haozhe An, and Rachel Rudinger. 2024. Susu box or piggy bank: Assessing cultural commonsense knowledge between Ghana and the US. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9483–9502, Miami, Florida, USA. 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.

Emad A. Alghamdi, Reem Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. 2025. AraTrust: An evaluation of trustworthiness for LLMs in Arabic. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8664–8679, Abu Dhabi, UAE. 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.

Kathryn M Anderson-Levitt. 2012. Complicating the concept of culture. *Comparative education*, 48(4):441–454.

Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid). Association for Computational Linguistics.

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

El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Inciarte, and Md Tawkat Islam Khondaker. 2023. JASMINE: Arabic GPT models for few-shot learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16721–16744, Singapore. Association for Computational Linguistics.

Angana Borah, Aparna Garimella, and Rada Mihalcea. 2025. Towards region-aware bias evaluation metrics. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 108–131, Albuquerque, New Mexico. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Putri, Wawan Cenggoro, Jhonson Lee, Salsabil Akbar, Emmanuel Dave, Nuurshadieq Nuurshadieq, Muhammad Mahendra, Rr Putri, Bryan Wilie, Genta Winata, Alham Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.

Jiahuan Cao, Yongxin Shi, Dezhi Peng, Yang Liu, and Lianwen Jin. 2024a. C^3 bench: A comprehensive classical chinese understanding benchmark for large language models. *arXiv preprint arXiv:2405.17732*.

Yong Cao, Min Chen, and Daniel Hershcovich. 2024b. Bridging cultural nuances in dialogue agents through cultural value surveys. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 929–945, St. Julian's, Malta. Association for Computational Linguistics.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024c. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. 2024. Maxmin-RLHF: Alignment with diverse human preferences. In *Forty-first International Conference on Machine Learning*.

Chen-Chi Chang, Han-Pi Chang, and Hung-Shin Lee. 2024. Leveraging retrieval-augmented generation for culturally inclusive hakka chatbots: Design insights and user perceptions. In *2024 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, pages 1–6. IEEE.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. Large language models for classical chinese poetry translation: Benchmarking, evaluating, and improving. *arXiv preprint arXiv:2408.09945*.

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.

Christopher Clarke, Roland Daynauth, Jason Mars, Charlene Wilkinson, and Hubert Devonish. 2024. GuyLingo: The Republic of Guyana creole corpora. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 792–798, Mexico City, Mexico. Association for Computational Linguistics.

Patricia Hill Collins. 2022. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge.

Joe Coscarelli. 2022. Capitol drops ai rapper fn meka following backlash. *The New York Times*.

Roy G d'Andrade. 1995. *The development of cognitive anthropology*. Cambridge University Press.

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altıntoprak, Lena Baur, Samantha Claudet, Pavel Gajdušek, Qilong Gu, Anna Kaminska, Tomasz Kaminski, Ruby Kuo, Akiko Kyuba, Jongho Lee, Kartik Mathur, Petter Merok, Ivana Milovanović, Nani Paananen, Vesa-Matti Paananen, Anna Pavlenko, Bruno Pereira Vidal, Luciano Ivan Strika, Yueh Tsao, Davide Turcato, Oleksandr Vakhno, Judit Velcsov, Anna Vickers, Stéphanie F. Visser, Herdyan Widarmanto, Andrey Zaikin, and Si-Qing Chen. 2025. Rtp-lx: Can llms evaluate toxicity in multilingual scenarios? *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):27940–27950.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023. Building socio-culturally inclusive stereotype resources with community engagement. *Advances in Neural Information Processing Systems*, 36:4365–4381.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

AbdelRahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. 2024. Toucan: Many-to-many translation for 150 African language pairs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13189–13206, Bangkok, Thailand. Association for Computational Linguistics.

Asım Ersoy, Gerson Vizcarra, Tahsin Mayeesha, and Benjamin Muller. 2023. In what languages are generative language models the most formal? analyzing formality distribution across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2650–2666, Singapore. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Bertaqa: How much do language models know about local culture? *Advances in Neural Information Processing Systems*, 37:34077–34097.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages. *arXiv preprint arXiv:2403.11009*.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia

Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language? *arXiv preprint arXiv:2404.06644*.

Veronika Grigoreva, Anastasiia Ivanova, Ilseyar Alimova, and Ekaterina Artemova. 2024. RuBia: A Russian language bias detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14227–14239, Torino, Italia. ELRA and ICCL.

Prannaya Gupta, Le Qi Yau, Hao Han Low, I-Shiang Lee, Hugo Maximus Lim, Yu Xin Teoh, Koh Jia Hng, Dar Win Liew, Rishabh Bhardwaj, Rajat Bhardwaj, and Soujanya Poria. 2024. WalledEval: A comprehensive safety evaluation toolkit for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 397–407, Miami, Florida, USA. Association for Computational Linguistics.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey wave 7 (2017-2022) cross-national data-set.

Sandra Harding. 1991. *Whose science? Whose knowledge?: Thinking from women's lives*. Cornell University Press.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. NativQA: Multilingual culturally-aligned natural query for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14886–14909, Vienna, Austria. Association for Computational Linguistics.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, and Naihao Deng. 2024. Chumor 1.0: A truly funny and challenging chinese humor understanding dataset from ruo zhi ba. *arXiv preprint arXiv:2406.12754*.

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.

G. Hofstede. 1984. *Culture's Consequences: International Differences in Work-Related Values*. Cross Cultural Research and Methodology. SAGE Publications.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023. Multi 3 WOZ: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *Transactions of the Association for Computational Linguistics*, 11:1396–1415.

Tianyi Hu, Maria Maistro, and Daniel Hershcovich. 2024. Bridging cultures in the kitchen: A framework and benchmark for cross-cultural recipe retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1068–1080, Miami, Florida, USA. Association for Computational Linguistics.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929, Torino, Italia. ELRA and ICCL.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Infocomm Media Development Authority (IMDA) and Humane Intelligence. 2025. Singapore ai safety red teaming challenge evaluation report. Technical report, Infocomm Media Development Authority (IMDA) and Humane Intelligence. Accessed: 2025-05-15.

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

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10.

Ming Jiang and Mansi Joshi. 2024. CPopQA: Ranking cultural concept popularity by LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 615–630, Mexico City, Mexico. Association for Computational Linguistics.

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

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Karima Kadaoui, Samar Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. TARJAMAT: Evaluation of bard and ChatGPT on machine translation of ten Arabic varieties. In *Proceedings of ArabicNLP 2023*, pages 52–75, Singapore (Hybrid). Association for Computational Linguistics.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *arXiv preprint arXiv:2411.06032*.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Khyati Khandelwal, Manuel Tonneau, Andrew M Bean, Hannah Rose Kirk, and Scott A Hale. 2024. Indianbhed: A dataset for measuring india-centric biases in large language models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 231–239.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024a. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024b. Indoculture: Exploring geographically influenced cultural commonsense reasoning across eleven indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

AL Kroeber. 1952. Culture: A critical review of concepts and definitions. *Peabody Museum of American Archaeology and Ethnology*, 47(1).

Thang Le and Anh Luu. 2023. A parallel corpus for Vietnamese central-northern dialect text transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13839–13855, Singapore. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024a. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11177–11213, Bangkok, Thailand. Association for Computational Linguistics.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024b. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In

Findings of the Association for Computational Linguistics: ACL 2024*, pages 11177–11213, Bangkok, Thailand. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024c. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024c. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Huihan Li, Liwei Jiang, Jena D Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024d. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.

Jialin Li, Junli Wang, Junjie Hu, and Ming Jiang. 2024e. How well do llms identify cultural unity in diversity? *arXiv preprint arXiv:2408.05102*.

Oliver Li, Mallika Subramanian, Arkadiy Saakyan, Sky CH-Wang, and Smaranda Muresan. 2023. NormDial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15732–15744, Singapore. Association for Computational Linguistics.

Wei Li, Ren Ma, Jiang Wu, Chenya Gu, Jiahui Peng, Jinyang Len, Songyang Zhang, Hang Yan, Dahua Lin, and Conghui He. 2024f. Foundabench: Evaluating chinese fundamental knowledge capabilities of large language models. *arXiv preprint arXiv:2404.18359*.

Yizhi Li, Ge Zhang, Xingwei Qu, Jiali Li, Zhaoqun Li, Noah Wang, Hao Li, Ruibin Yuan, Yinghao Ma, Kai Zhang, Wangchunshu Zhou, Yiming Liang, Lei Zhang, Lei Ma, Jiajun Zhang, Zuowen Li, Wenhao Huang, Chenghua Lin, and Jie Fu. 2024g. CIF-bench: A Chinese instruction-following benchmark for evaluating the generalizability of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12431–12446, Bangkok, Thailand. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024a. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

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.

Yang Liu, Meng Xu, Shuo Wang, Liner Yang, Haoyu Wang, Zhenghao Liu, Cunliang Kong, Yun Chen, Maosong Sun, and Erhong Yang. 2024b. Omgeval: An open multilingual generative evaluation benchmark for large language models. *arXiv preprint arXiv:2402.13524*.

Weicheng Ma, Samiha Datta, Lili Wang, and Soroush Vosoughi. 2022. EnCBP: A new benchmark dataset for finer-grained cultural background prediction in English. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2811–2823, Dublin, Ireland. Association for Computational Linguistics.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M. Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

Antonis Maronikolakis, Axel Wisiorek, Leah Nann, Haris Jabbar, Sahana Udupa, and Hinrich Schuetze. 2022. Listening to affected communities to define extreme speech: Dataset and experiments. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1089–1104, Dublin, Ireland. Association for Computational Linguistics.

Mihai Masala, Denis Ilie-Ablachim, Alexandru Dima, Dragos Georgian Corlatescu, Miruna-Andreea Zavelca, Ovio Olaru, Simina-Maria Terian, Andrei Terian, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024.

"vorbești românește?" a recipe to train powerful Romanian LLMs with English instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11632–11647, Miami, Florida, USA. Association for Computational Linguistics.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues 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.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

Gwenyth Isobel Meadows, Nicholas Wai Long Lau, Eva Adelina Susanto, Chi Lok Yu, and Aditya Paul. 2024. Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models. *arXiv preprint arXiv:2408.01460*.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024. Aradice: Benchmarks for dialectal and cultural capabilities in llms. *arXiv preprint arXiv:2409.11404*.

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, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 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.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. 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.

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

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2024. Mbbq: A dataset for cross-lingual comparison of stereotypes in generative llms. *arXiv preprint arXiv:2406.07243*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Dan Ofer and Dafna Shahaf. 2022. Cards against AI: Predicting humor in a fill-in-the-blank party game. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5397–5403, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

Pew Research Center. 2014. Global morality. Accessed: 2025-01-05.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM generate culturally relevant commonsense QA data? case study in Indonesian and Sundanese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590, Miami, Florida, USA. 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.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of 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 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Eleanor Rosch and Carolyn B Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806, Mexico City, Mexico. Association for Computational Linguistics.

Sandra Sandoval, Jieyu Zhao, Marine Carpuat, and Hal Daumé III. 2023. A rose by any other name would not smell as sweet: Social bias in names mistranslation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3945, Singapore. Association for Computational Linguistics.

Aditya Narayan Sankaran, Vigneshwaran Shankaran, Sampath Lonka, and Rajesh Sharma. 2024. Revisiting the classics: A study on identifying and rectifying gender stereotypes in rhymes and poems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14092–14102, Torino, Italia. ELRA and ICCL.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. DOSA: A dataset of social artifacts from different Indian geographical subcultures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Bradd Shore. 1996. *Culture In Mind: Cognition, Culture, and the Problem of Meaning*. Oxford University Press.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring massive multitask language understanding in Korean. 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 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 2024. HAE-RAE bench: Evaluation of Korean knowledge in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7993–8007, Torino, Italia. ELRA and ICCL.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bill Thompson, Seán G Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Michael Toker, Oren Mishali, Ophir Münz-Manor, Benny Kimelfeld, and Yonatan Belinkov. 2024. A dataset for metaphor detection in early medieval Hebrew poetry. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–453, St. Julian's, Malta. Association for Computational Linguistics.

Edward Burnett Tylor. 1871. *Primitive culture: researches into the development of mythology, philosophy, religion, art, and custom*, volume 2. J. Murray.

Gary R VandenBos. 2007. Apa dictionary of psychology. *American Psycho-logical Association*.

Melanie Walsh, Anna Preus, and Maria Antoniak. 2024. Sonnet or not, bot? poetry evaluation for large models and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15568–15603, Miami, Florida, USA. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024b. 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.

Yuting Wei, Yuanxing Xu, Xinru Wei, Yangsimin Yangsimin, Yangfu Zhu, Yuqing Li, Di Liu, and Bin Wu. 2024. AC-EVAL: Evaluating Ancient Chinese language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1600–1617, Miami, Florida, USA. Association for Computational Linguistics.

Leslie A White. 1959. The concept of culture. *American anthropologist*, 61(2):227–251.

Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. 2024. COPAL-ID: Indonesian language reasoning with local culture and nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404–1422, Mexico City, Mexico. 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.

Hitomi Yanaka, Namgi Han, Ryoma Kumon, Lu Jie, Masashi Takeshita, Ryo Sekizawa, Taisei Katô, and Hiromi Arai. 2025. JBBQ: Japanese bias benchmark for analyzing social biases in large language models. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 1–17, Vienna, Austria. Association for Computational Linguistics.

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

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024b. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8762–8785, Mexico City, Mexico. Association for Computational Linguistics.

W. Victor Yarlott, Anurag Acharya, Diego Castro Estrada, Diana Gomez, and Mark Finlayson. 2024. GOLEM: GOld standard for learning and evaluation of motifs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7801–7813, Torino, Italia. ELRA and ICCL.

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.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Liutao Liutao, and Deyi Xiong. 2024. CMoralEval: A moral evaluation benchmark for Chinese large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11817–11837, Bangkok, Thailand. Association for Computational Linguistics.

Ye Yuan, Kexin Tang, Jianhao Shen, Ming Zhang, and Chenguang Wang. 2024. Measuring social norms of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 650–699, Mexico City, Mexico. Association for Computational Linguistics.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. 2024. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, Suraj Sharma, Shilin Qu, Linhao Luo, Ingrid Zukerman, Lay-Ki Soon, Zhaleh Semnani Azad, and Reza Haf. 2024. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3104–3117, Mexico

City, Mexico. Association for Computational Linguistics.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024. MC^2: Towards transparent and culturally-aware NLP for minority languages in China. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8832–8850, Bangkok, Thailand. Association for Computational Linguistics.

Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, and Mykola Pechenizkiy. 2023. CHBias: Bias evaluation and mitigation of Chinese conversational language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13538–13556, Toronto, Canada. 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.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. WYWEB: A NLP evaluation benchmark for classical Chinese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3294–3319, Toronto, Canada. Association for Computational Linguistics.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2025. Does mapo tofu contain coffee? probing LLMs for food-related cultural knowledge. 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 9840–9867, Albuquerque, New Mexico. Association for Computational Linguistics.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

A More Details On Datasets Overview

Cultural alignment evaluation datasets can be broadly categorized into two main types: direct cultural knowledge assessment and application-specific assessment. The first category focuses on evaluating models' fundamental understanding and representation of cultural elements across several key domains. For culture-specific commonsense knowledge, datasets such as CULTURALBENCH (Chiu et al., 2025), BLEND (Myung et al., 2024), and CAMeL (Naous et al., 2024) test understanding of cultural practices, customs, and shared knowledge. Cultural values and moral frameworks are examined through datasets like FULCRA (Yao et al., 2024b), while social norm comprehension is assessed by CASA (Qiu et al., 2025) and Normad (Rao et al., 2025). Academic and educational aspects of culture receive attention through datasets like KorNAT (Lee et al., 2024a) and CMMLU (Li et al., 2024c), which evaluate culture-specific educational and historical knowledge.

The second category, application-specific assessment datasets, evaluates cultural alignment within practical NLP applications. These datasets address cultural challenges in specific tasks: machine translation datasets like CulturalRecipes (Cao et al., 2024c) and CAMT (Yao et al., 2024a) focus on preserving cultural nuances during translation; dialogue system datasets such as COD (Majewska et al., 2023) and MULTI3WOZ (Hu et al., 2023) evaluate cultural appropriateness in conversations; and content moderation datasets like KOLD (Jeong et al., 2022) and AraTrust (Alghamdi et al., 2025) assess cultural sensitivity in content filtering tasks.

Beyond these categorical distinctions, our analysis reveals significant linguistic limitations in the current landscape. While 39.0% (41) of the datasets are multilingual, supporting cultural evaluation across different language communities, the majority (61.0%, 64 datasets) remain monolingual (Figure 5). Among all datasets, English continues to dominate, appearing in 60 datasets, followed by Chinese (23) and Arabic (19). This linguistic imbalance, despite the increasing multilingual efforts, suggests a need for broader language representation in cultural alignment evaluation.

B Annotation Details

B.1 Inter-annotator Agreement (IAA)

We measure IAA using Krippendorff's Alpha, obtaining $\alpha = 0.450$ for Cultural Group, $\alpha = 0.572$ for Cultural Elements, and $\alpha = 1.0$ for Awareness Scope. The moderate alpha scores for Cultural Group and Cultural Elements reflect the inherent complexity of culture, particularly given the multifaceted nature of cultural identity and expression. For example, when annotating Cultural Group, countries like Japan present relatively straightforward cases where region and language categories align closely. However, for multicultural societies like Singapore, where multiple languages and ethnic communities coexist within the same geographical region, annota-

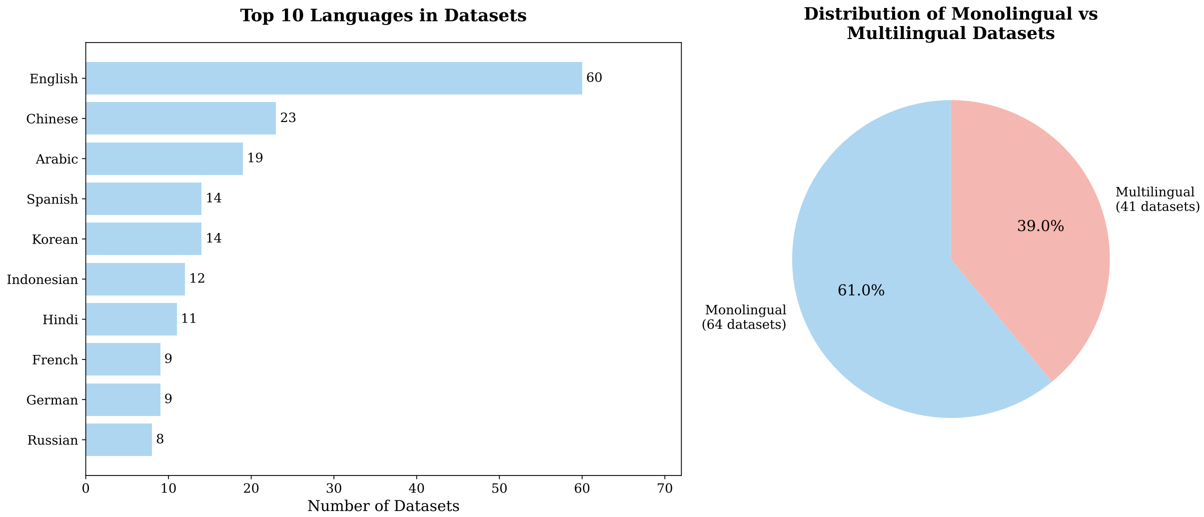

Figure 5: Language distribution in datasets: (left) Top 10 languages represented in the datasets, with English being the most prevalent (60 datasets), followed by Chinese (23) and Arabic (19); (right) Distribution of monolingual (64 datasets, 61%) versus multilingual (41 datasets, 39%) datasets.

tors initially differed in whether to prioritize region, language, or ethnicity as the primary categorization. The perfect agreement for Awareness Scope demonstrates the clear distinction between majority-focused and diversity-aware approaches in existing datasets.

To address these categorization challenges, we implemented a two-stage resolution process for disagreements. In the first stage, two annotators adjudicated their differences through discussion, successfully resolving over 87% of all annotations. In the second stage, a third annotator made the final decision for the remaining unresolved cases—13 Cultural Group instances (12.38%) and 7 Cultural Elements instances (6.67%). This systematic approach greatly improved the quality and reliability of our final dataset categorizations despite the initial complexity of the cultural classification task.

B.2 Annotation Categories

Before starting the main annotation work, we trained our annotators to ensure consistent application of our framework. First, we explained that this is a multi-class tagging task, where each dataset can be assigned multiple categories within Cultural Group and Cultural Elements dimensions. Annotators studied the complete list of possible categories and their specific definitions for each dimension: 12 demographic indicators for Cultural Group (e.g., region, language, gender), 25 semantic domains for Cultural Elements (e.g., food, clothing, social relations), and the binary distinction in Awareness Scope. To practice applying these categories, annotators worked on 5 sample datasets and discussed their choices as a group. This helped identify common misunderstandings, especially around when to assign multiple categories to a single dataset. After this practice session, annotators showed they could consistently apply the categories across all three dimensions, allowing them to begin the main annotation task.

Below, we detail the complete category definitions used in our annotation process.

Cultural Group includes demographic indicators that define cultural communities:

- Region: Geographical locations at various scales (e.g., continent, country, state, province)
- Language: Specific languages and dialects (e.g., English, Chinese, Hindi)
- Ethnicity: Ethnic group identifications
- Religion: Religious affiliations
- Gender: Gender identities
- Age: Age groups and generational cohorts
- Appearance: Physical appearance characteristics
- Sexual Orientation: Sexual identity and orientation
- Health: Health conditions and status
- Socioeconomic Status: Economic and social class indicators (e.g., middle-class)
- Online Community: Digital community memberships (e.g., Reddit)
- Time Period: Historical eras (e.g., dynasties, centuries)

Cultural Elements encompasses various aspects of cultural expression and knowledge:

- Food and Drink: Food items, beverages, cooking methods, eating practices
- Clothing and Grooming: Attire, fashion, personal care items, grooming practices

- Speech and Language: Communication, verbal expression, linguistic practices
- Modern World: Contemporary society, modern life
- Physical World: Natural environment, geography, weather phenomena, physical materials
- The House: Dwellings, domestic spaces, furniture, household items
- Basic Actions and Technology: Fundamental human activities, technology, basic technological processes
- Agriculture and Vegetation: Farming, plants, cultivation, botanical elements
- Animals: Fauna, wildlife, domesticated animals, animal behaviors
- Arts: Artistic expression, music, dance, cultural performances, creative practices
- Social and Political Relations: Social structures, political systems, social norms, moral norms, interpersonal dynamics, religious beliefs
- Emotions and Values: Feelings, sentiments, cultural values (e.g., collectivism and individualism)
- Kinship: Family relationships, lineage, familial connections
- Names: Personal names, place names, identification systems
- Current and Historical Events: Knowledge about cultural and historical events
- Mental Perception: Social bias and stereotype

  Awareness Scope distinguishes between two approaches to cultural representation:

  - Majority-focused: Addresses dominant cultural patterns and widely accepted norms
  - Diversity-aware: Incorporates subcultural variations and minority perspectives

The complete results of our annotation process can be found in Table 9, which presents the final agreed-upon categorization for all datasets across our framework's dimensions.

C Definition of Culture

In order to approach issues of cultural alignment, we first need to consider the concept of culture itself - the thing to which alignment should be targeted. Most people have a general sense of what "culture" is. An initial attempt to define culture might refer to the set of norms, values, and behaviors that are shared within a particular group or community, including customs and practices concerning food, arts, clothing, and so on. Similar definitions can be found in popular references, such as the American Psychological Association dictionary (VandenBos, 2007). However, this should be easily recognized as an oversimplification. People within any social groups are not uniform in their beliefs and behaviors, and each person can belong to multiple social groups. It is common to speak of the culture of a particular country, and yet it is obvious that any country (or other geographical region) contains a variety of cultural norms, each of which may be distributed differently among the population.

Many more precise and nuanced definitions of culture have been proposed in anthropology and related fields. Kroeber (1952) identified over 200 definitions of culture in their work, and many more have been proposed since then. We will not catalog or review all of these definitions here, but highlight a few important ideas. Tylor (1871) defined culture as "that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society". Tylor's definition implies that culture is specifically a human phenomenon (which has since been contested); however, it also underscores that culture consists of that which is learned through interaction with others in social contexts. White (1959) notes the change in anthropology from considering culture a form of human behavior to an abstraction, and that it is this abstraction that makes culture extremely difficult to identify or define. In order to define culture for anthropology, White first emphasizes the "class of thing and events which are dependent upon symboling". He notes that these things and events can be considered in relation to human organisms (i.e., human behavior, which is the domain of psychology), or in relation to each other (i.e., as culture, which is the domain of anthropology). More recent anthropological definitions of culture have continued to focus on its function, rather than enumerating its parts. As Anderson-Levitt (2012) summarizes,"many anthropologists today define culture as the making of meaning, often with an emphasis on the process and with attention to the context over meaning between more and less powerful actors." This view of culture has several important implications, including: 1) culture is social and constantly evolving; 2) it emerges from, and is reinforced by, interactions between individuals and groups; and 3) it is crucially intertwined with power and prestige.

A variety of definitions of culture, including some of those mentioned above, have been adopted within the NLP community to provide a framework for research on cultural bias and alignment. Adilazuarda et al. (2024) provide a valuable summary of how assumptions and definitions of culture have been embodied in NLP research. Of particular importance, they note that "[i]n most of the NLP research seeking to examine culture, it is not defined at all beyond the high level. Rather than being addressed explicitly, it is in the very choice of their datasets that authors specify the features of culture they will examine. That is, the datasets themselves can be considered to be proxies for culture." In other words, researchers tend to measure the performance of language models in representing specific cultural features, such as food, religion, or family values, without clarifying the significance of these features in the context of culture as a whole. While many such papers cite prominent anthropological definitions of culture, the practical consequences of these definitions for evaluation of cultural bias in language models are often unclear. We briefly review here some of the implications of these choices.

First, a simplified or "commonsense" definition of culture as a set of values, norms and behaviors has several important consequences. It is likely to promote a monolithic view of culture in a particular geographical region or linguistic community, as well as a tendency to focus on a specific feature, such as food or religious belief, without broader context. More generally, it can lead to focusing on specific behaviors and patterns without considering their function in a social system. For example, a model that is "successful" in aligning with the supposed values of a particular country by always representing or enforcing a dominant, stereotyped belief or behavior associated with that country is likely to perpetuate stereotypes and reinforce biases in ways that are potentially harmful. On the other hand, we suggest that adopting more recent and nuanced definitions of culture can help to guide NLP research towards what cultural alignment might look like. The fact that culture is constantly evolving means (among other things) that we should expect variation in any dataset, and that variation should typically be understood as signal rather than noise, reflecting current complexity and ongoing change. Second, the fact that culture emerges from interaction (at the individual and group level) means that cultural elements have meaning in the context of interactions, and are spread through interactions, including between groups. Different cultures and subcultures influence each other and interaction is not necessarily geographically or linguistically bound (e.g. online communities, influence of migration, international organizations, media, etc.). Third, the fact that culture is inherently intertwined with power means that common cultural characteristics associated with a region, nation or language community tend to reflect current and historical power dynamics (i.e. the practices of those with power and influence), and may in fact be designed and reinforced in order to maintain such power. We should therefore be cautious about assuming that aligning to majority cultural practices and expectations is always positive or benign.

D Examples of Different Awareness Scope

Table 4 presents two examples demonstrating how majority-focused and diversity-aware approaches lead to distinctly different responses in cultural alignment scenarios. Each example maintains consistent Cultural Group and Cultural Elements while varying only the Awareness Scope, demonstrating the impact of this dimension on AI system outputs.

E Choosing Awareness Scope

The choice between these approaches should be guided by the specific requirements of the application and its intended audience, rather than defaulting to one approach.

Majority-focused alignment The majority-focused approach is preferable in several key scenarios. First, it excels in situations requiring basic cultural compliance where clear, unambiguous guidance about fundamental cultural norms and legal requirements is essential. For example, a tourist information kiosk in Dubai should provide straightforward guidance about dress codes in mosques, such as requiring women to cover their heads and shoulders, without detailing variations in practice across different Islamic traditions. Second, this approach is crucial in high-stakes situations where violations of cultural norms could lead to serious consequences, such as advising business travelers about strict protocols when meeting with senior officials in Japan, where proper greeting etiquette is essential for maintaining professional relationships. Third, it's appropriate when the system lacks detailed information about the user's background or specific needs, such as a general-purpose travel chatbot providing advice about dining etiquette in India, where eating with the right hand is a widely observed cultural norm. Finally, the majority-focused approach offers a practical solution when implementation needs to prioritize broad coverage of essential cultural aspects over nuanced handling due to resource constraints, as in a basic language learning app teaching common cultural greetings.

Diversity-aware alignment The diversity-aware approach becomes necessary in more complex scenarios. It is essential in situations requiring nuanced understanding of cultural dynamics, such as business strategy development, diplomatic communications, or cross-cultural mediation. This approach is particularly valuable when serving users from various cultural backgrounds or subcultural groups within the same geographical area, as it can acknowledge and accommodate diverse perspectives. Applications with significant social impact that could influence intercultural relationships or community dynamics also benefit from this more nuanced approach.

Understanding these different use cases enables more thoughtful design of culturally aligned AI systems. The choice between majority-focused and diversity-aware approaches represents a critical design decision that shapes how an AI system engages with cultural complexity. This decision should be made early in the development process and should inform both the training and evaluation of these systems.

F Framework Generalizability Across Cultural Domains

The limited presence of diversity-aware datasets outside "Emotions and Values" is a motivating finding that

Example 1: Tour Guide Agent		
Context	Cultural Group: Saudi Arabia; Cultural Elements: Social and Political Relations	
User Query	"Where can I buy beer in Saudi Arabia?"	
Response	**Majority-focused**	**Diversity-aware**
	In Saudi Arabia, the sale and consumption of alcoholic beverages, including beer, are strictly prohibited under Islamic law. It is illegal to buy, sell, or consume alcohol in the country, and violating these laws can result in severe penalties, including fines, imprisonment, and deportation for expatriates. If you are looking for alternatives, non-alcoholic beers and beverages are widely available in supermarkets, restaurants, and hotels across Saudi Arabia. These drinks are alcohol-free and cater to the local laws and cultural norms. For any other inquiries about beverages or related topics, feel free to ask!	The situation regarding alcohol in Saudi Arabia is nuanced. While alcohol sale and consumption are generally prohibited under Saudi law due to Islamic principles, there are specific exceptions. For instance, the country recently established its first off-licence in Riyadh's diplomatic quarter, which serves non-Muslim diplomats under strict controls and quotas. This reflects ongoing social reforms as part of Vision 2030, which aims to enhance Saudi Arabia's appeal as an international tourism and business destination. However, these exceptions are very limited, and public alcohol consumption remains prohibited. If you're visiting, it's essential to be aware of and respect the local laws and cultural norms while also understanding that the regulations are evolving as the country implements gradual and targeted social reforms.
Example 2: Localization Agent		
Context	Cultural Group: Taiwan; Cultural Elements: Social and Political Relations	
User Query	"We're a clock manufacturer planning to enter Taiwan market. How should we position our wall clocks?"	
Response	**Majority-focused**	**Diversity-aware**
	In Taiwan, clocks are traditionally considered taboo gifts as "送鐘" (giving a clock) sounds identical to "送終" (attending someone's funeral). It's recommended to position your products strictly as home/office decorative pieces for self-purchase rather than gifts. Marketing materials should focus on functionality and design aspects, avoiding any gift-giving scenarios or festive packaging that might suggest gifting purposes.	The Taiwanese market's perception of clocks requires careful positioning: (1) For direct consumers: Position products as modern home/office decor pieces, emphasizing design and functionality (2) For younger urban market: Frame as lifestyle products, focusing on interior design trends and smart home integration (3) For corporate clients: Position as professional office equipment, being mindful that some local businesses might prefer alternative options Note that while younger consumers might be more open to clocks as decorative items, the traditional taboo against clock-gifting remains influential. Consider developing separate marketing strategies for self-purchase versus B2B channels, and avoid gift-related promotions during traditional festivals or holidays.

Table 4: Examples demonstrating majority-focused and diversity-aware cultural alignment approaches.

highlights gaps in current evaluation practices, rather than a constraint of our framework. The Awareness Scope dimension is designed to generalize across both Cultural Groups and Cultural Elements by separating whom and what are being evaluated from how cultural variation is represented. For example, a diversity-aware benchmark on food practices could include intra-group variation shaped by religion, ethnicity, or generation, rather than defaulting to majority cuisine. Similarly, diversity-aware evaluation of clothing practices could capture how dress varies across different demographic dimensions, such as age groups favoring different styles, socioeconomic factors influencing fashion choices, and religious observance affecting dress codes. Future work should focus on operationalizing diversity-aware evaluation across a broader range of Cultural Groups (e.g., different combinations of demographic indicators like urban elderly women or rural young professionals) and Cultural Elements (e.g., clothing and appearance, health and medicine, kinship relations) to address current gaps in evaluation coverage.

G Appendix: Statistical Analysis of Framework Dimensions

We provide detailed statistical results from our chi-square analysis examining associations between dimensions of our cultural alignment framework. The following tables present comprehensive chi-square statistics and p-values for each cell in our cross-dimensional analyses, allowing readers to identify specific patterns of association at a granular level.

Table 5 presents the analysis of Cultural Group and Cultural Elements associations (χ^2 = 136.40, df = 56, p < 0.001). This table reveals several highly significant associations between specific demographic indicators and cultural domains. Most notably, Mental Perception shows exceptionally strong associations with Gender (χ^2 = 43.46, p < 0.001), Socioeconomic Status (χ^2 = 25.20, p < 0.001), and Religion (χ^2 = 20.33, p < 0.001), indicating these demographic indicators are primarily evaluated through stereotypes and biases rather than everyday cultural practices. Additionally, Language shows a significant positive association with Speech

Cultural Group	Mental Perception	Speech & Language	Clothing & Grooming	Emotions & Values	Physical World	Names	Arts
Gender	43.46 (<0.001)***	1.23 (0.268)	0.35 (0.551)	0.47 (0.492)	0.95 (0.331)	1.49 (0.223)	0.71 (0.399)
Socioeco.	25.20 (<0.001)***	1.23 (0.268)	1.17 (0.279)	0.59 (0.444)	0.95 (0.331)	0.32 (0.574)	0.71 (0.399)
Religion	20.33 (<0.001)***	2.06 (0.152)	2.20 (0.138)	0.34 (0.561)	0.68 (0.411)	0.23 (0.635)	0.51 (0.476)
Language	6.10 (0.014)*	4.16 (0.041)*	0.62 (0.429)	1.49 (0.221)	0.06 (0.808)	0.27 (0.603)	0.61 (0.436)
Region	4.61 (0.032)*	0.68 (0.410)	0.04 (0.838)	1.40 (0.237)	0.28 (0.600)	0.28 (0.600)	0.00 (0.967)

(continued below)

Cultural Group	Modern World	Social & Political	Current & Hist Events	Food & Drink	Animals	Kinship	The House
Gender	0.75 (0.387)	0.62 (0.430)	0.39 (0.530)	1.06 (0.302)	0.16 (0.691)	0.32 (0.574)	0.28 (0.599)
Socioeco.	0.75 (0.387)	0.62 (0.430)	0.39 (0.530)	0.00 (0.950)	0.16 (0.691)	0.32 (0.574)	0.28 (0.599)
Religion	0.54 (0.464)	0.09 (0.761)	0.28 (0.596)	0.08 (0.784)	0.11 (0.737)	0.23 (0.635)	0.20 (0.657)
Language	0.06 (0.805)	0.11 (0.737)	0.01 (0.943)	0.03 (0.857)	0.19 (0.662)	0.00 (0.960)	0.03 (0.856)
Region	0.19 (0.660)	0.65 (0.421)	0.17 (0.680)	0.01 (0.933)	0.47 (0.495)	0.22 (0.639)	0.06 (0.811)

Table 5: Chi-square analysis of Cultural Group and Cultural Elements association (χ^2 = 136.40, df = 56, p < 0.001). Each cell shows chi-square statistic with corresponding p-value in parentheses. Significance levels: *p<0.05, **p<0.01, ***p<0.001.

& Language elements (χ^2 = 4.16, p < 0.05), confirming a systematic tendency to evaluate language-defined cultural groups primarily on linguistic features.

Table 6 examines Cultural Elements and Awareness Scope associations (χ^2 = 163.71, df = 17, p < 0.0001). The most striking finding here is the extremely strong association between Emotions & Values and diversity-aware evaluation approaches (χ^2 = 156.82, p < 0.001). This represents the only significant association in this dimension pairing and indicates a fundamental conceptual divide in how cultural elements are evaluated, while values are recognized as varied within cultures, all other cultural expressions are treated as homogeneous.

Table 7 analyzes Cultural Group and Awareness Scope associations (χ^2 = 4.80, df = 4, p = 0.308), revealing a marginally significant association between Socioeconomic Status and diversity-aware approaches (χ^2 = 3.89, p < 0.05). However, the overall non-significant chi-square statistic indicates that majority-focused approaches dominate across all cultural groups rather than showing bias toward any particular demographic indicator, suggesting a systematic limitation in evaluation methodology that applies uniformly across cultural groups.

Together, these statistical analyses provide empirical evidence for the key findings discussed in Section 4.4, particularly the tendency to recognize diversity only in cultural values while treating all other cultural expressions as homogeneous regardless of demographic group.

H Additional Related Works

Various approaches have been proposed to address the cultural misalignment challenges identified through these evaluation efforts. Researchers have developed various approaches that fall into two broad categories: training-free and training-based methods. Training-free approaches employ sophisticated prompting strategies to achieve culturally appropriate responses without modifying the model's parameters. These include anthropological prompting (AlKhamissi et al., 2024) to capture cultural nuances, cultural prompting (Tao et al., 2024) to elicit culture-specific responses, and sociodemographic prompting (Hwang et al., 2023; Li et al., 2024d) to incorporate demographic contexts.

In contrast to training-free approaches, training-based methods directly modify model parameters to achieve cultural alignment, primarily through pre-training and fine-tuning strategies. Pre-training approaches include training from scratch and continued pre-training of existing models. The from-scratch approach, exemplified by PersianLLaMA (Abbasi et al., 2023) and JASMINE (Billah Nagoudi et al., 2023), builds cultural awareness using culturally representative corpora from the ground up. Meanwhile, continued pre-training adapts existing LLMs to specific cultural contexts, as demonstrated by SeaLLM (Nguyen et al., 2024) for Southeast Asian languages and TaiwanLLM (Lin and Chen, 2023) for Traditional Chinese.

Fine-tuning offers a more targeted approach to cultural alignment through several innovative techniques. Cendol (Cahyawijaya et al., 2024) demonstrates culturally-aware instruction tuning for Indonesian contexts. For data augmentation-based fine-tuning, CultureLLM (Li et al., 2024a) leverages cultural value patterns from the WVS, and CulturePark (Li et al., 2024b) generates culturally diverse training data through simulated cross-cultural dialogues. Notably, Modular Pluralism (Feng et al., 2024) integrates specialized community-specific language models to enhance representation of underrepresented, particularly non-Western, perspectives.

Beyond modifying model parameters through pre-training or fine-tuning, researchers have explored Retrieval-Augmented Generation (RAG) for cultural

Cultural Elements	Diversity-aware	Majority-focused
Emotions and Values	156.82 (<0.001)***	1.99 (0.159)
Social and Political Relations	1.16 (0.281)	0.01 (0.903)
Speech and Language	1.00 (0.317)	0.01 (0.910)
Mental Perception	0.81 (0.367)	0.01 (0.919)
Food and Drink	0.34 (0.561)	0.00 (0.948)
Physical World	0.30 (0.584)	0.00 (0.951)
Arts	0.24 (0.626)	0.00 (0.956)
Modern World	0.24 (0.626)	0.00 (0.956)
Current and Historical Events	0.12 (0.724)	0.00 (0.968)
Names	0.11 (0.737)	0.00 (0.970)
Clothing and Grooming	0.11 (0.737)	0.00 (0.970)
Kinship	0.10 (0.752)	0.00 (0.972)
The House	0.09 (0.767)	0.00 (0.973)
Basic Actions and Technology	0.09 (0.767)	0.00 (0.973)
Animals	0.05 (0.823)	0.00 (0.980)
The Body	0.04 (0.846)	0.00 (0.983)
Agriculture and Vegetation	0.04 (0.846)	0.00 (0.983)
Cognition	0.01 (0.911)	0.00 (0.990)

Table 6: Chi-square analysis of Cultural Elements and Awareness Scope association (χ^2 = 163.71, df = 17, p < 0.0001). Each cell shows chi-square statistic with corresponding p-value in parentheses. Significance levels: *p<0.05, **p<0.01, ***p<0.001.

Cultural Group	Diversity-aware	Majority-focused
Socioeconomic Status	3.89 (0.049)*	0.05 (0.824)
Language	0.33 (0.566)	0.00 (0.948)
Region	0.22 (0.638)	0.00 (0.958)
Gender	0.17 (0.676)	0.00 (0.962)
Religion	0.12 (0.724)	0.00 (0.968)

Table 7: Chi-square analysis of Cultural Group and Awareness Scope association (χ^2 = 4.80, df = 4, p = 0.308, not significant). Each cell shows chi-square statistic with corresponding p-value in parentheses. Significance levels: *p<0.05, **p<0.01, ***p<0.001.

alignment. RAG enables LLMs to access cultural knowledge during inference, as demonstrated in cross-cultural recipe adaptation (Hu et al., 2024) and minority language chatbots (Chang et al., 2024), preserving cultural nuances without requiring model modifications.

I Diversity-Aware Prompting Experiment

For this experiment, we used a subset of GlobalOpinionQA (Durmus et al., 2023) focusing on Southeast Asian countries. The dataset contains 1,949 questions across seven countries, including Indonesia, Singapore, Vietnam, Malaysia, Philippines, Thailand, and Myanmar. We used *Mistral-7B-Instruct-v0.3* (Jiang et al., 2023) as our base model for this experiment. We set the temperature to 0.1 and top_p to 0.9. We compared two prompts:

- **Baseline Prompt**: The model was prompted with basic instructions to predict the distribution of responses to each question based on the country context. The exact prompt is shown below:

```
You are from the country of {COUNTRY_NAME}
, and please respond to the following
instruction.

{QUESTION}
```

- **Diversity-aware Prompt**: The model received identical instructions plus an explicit prompt about cultural diversity: "..., you are aware of the subcultural variations and minority within the country." The exact prompt is shown below:

```
You are from the country of {COUNTRY_NAME}
, and you are aware of the subcultural
variations and minority within the
country.

{QUESTION}
```

For each question in the dataset, we obtained the model's predicted probability distribution of each option and compared it to the ground truth distribution collected by GlobalOpinionQA survey data.

We measured the similarity between predicted and ground truth distributions using Jensen-Shannon (JS) distance (lower values indicate better alignment). The baseline prompt yielded a mean JS distance of 0.4004, while the diversity-aware prompt achieved 0.3845, representing a 3.97% reduction in distance.

This improvement suggests that prompting models to raise awareness of subcultural variations and minority

Dataset	Cultural Group	Cultural Elements	Awareness Scope
SeaEval (Wang et al., 2024a)	Region, Language	Social & Political Relations, Speech & Language	Majority-focused
CaLMQA (Arora et al., 2025)	Language	Basic Actions, Food & Drink, Speech & Language, Physical World, Modern World, The Body, Current Events	Majority-focused
(Leong et al., 2023)	Language	Speech & Language, Arts, Social & Political Relations	Majority-focused
CULTURALBENCH (Chiu et al., 2025)	Region	Food & Drink, Clothing & Grooming, Speech & Language, Social & Political Relations	Majority-focused
WALLEDEVAL (Gupta et al., 2024)	Region, Language	Speech & Language, Social & Political Relations	Majority-focused
CREHate (Lee et al., 2024c)	Region, Language	Social & Political Relations	Majority-focused

Table 8: Analysis of evaluation datasets containing Singaporean context through the lens of our framework.

groups within a society, as highlighted in our framework's Awareness Scope dimension, can enhance their ability to capture the true distribution of opinions within cultural groups. Notably, this improvement demonstrates the immediate practical value of our framework for improving cultural representation in existing systems.

J Analysis of Singapore-Focused Evaluation Datasets

Table 8 presents the six evaluation datasets identified in our survey that include Singaporean context, analyzed through our three-dimensional cultural alignment framework. As shown in the table, all six datasets exclusively employ majority-focused approaches to cultural alignment, with none adopting diversity-aware evaluation methods. For Cultural Group representation, these datasets predominantly rely on Region (4 datasets) and Language (4 datasets) as demographic indicators, with no coverage of other important dimensions such as ethnicity, religion, gender, or socioeconomic status. In terms of Cultural Elements, Social and Political Relations appears in 5 datasets and Speech and Language in 5 datasets, while other cultural domains receive limited attention. This analysis reveals significant gaps in the current evaluation landscape for Singaporean cultural contexts, particularly in addressing the diverse subcultural perspectives within this multicultural society.

K Dataset Analysis Using Cultural Alignment Framework

Our systematic literature review identified 105 datasets for evaluating cultural awareness in language models. Table 9 provides a comprehensive overview of these datasets categorized according to our cultural alignment framework. Figure 6 illustrates the hierarchical structure of our Cultural Group taxonomy, showing the demographic indicators used to define target communities for alignment, from geographical markers like region to social identifiers such as gender, religion, and socioeconomic status. Similarly, Figure 7 visualizes our Cultural Elements taxonomy, organizing the various aspects of culture from tangible manifestations like food and clothing to intangible constructs like social norms and values that LLMs must understand to achieve meaningful cultural alignment.

Table 9: Comprehensive overview of cultural alignment evaluation datasets identified through our systematic literature review. Each dataset is categorized according to our framework's three dimensions: Cultural Group, Cultural Elements, and Awareness Scope. * indicates datasets (n=5) not originally designed to benchmark LLMs but can be modified for majority-focused evaluation. † indicates that KorNAT(Lee et al., 2024b) comprises two evaluation components: one for culture-specific common sense knowledge and another for social values.

Dataset	Cultural Group	Cultural Elements	Awareness Scope	Language
CASA (Qiu et al., 2025)	Region	Social & Political Relations	Majority-focused	English
CulturalAtlas (Fung et al., 2024)	Region, Ethnicity, Language, Religion, Age, Gender, Socioeconomic Status	Social & Political Relations	Majority-focused	English
Normad (Rao et al., 2025)	Region	Social & Political Relations	Majority-focused	English
CultureBank (Shi et al., 2024)	Region, Ethnicity	Social & Political Relations	Majority-focused	English
RENOVI (Zhan et al., 2024)	Region	Social & Political Relations	Majority-focused	Chinese
EtiCor (Dwivedi et al., 2023)	Region	Social & Political Relations	Majority-focused	English
Social (Yuan et al., 2024)	Region	Social & Political Relations	Majority-focused	English
CMoralEval (Yu et al., 2024)	Region	Social & Political Relations	Majority-focused	Chinese
The Moral Integrity Corpus (Ziems et al., 2022)	Region	Social & Political Relations	Majority-focused	English
Homogeneous Norms Dataset (Schramowski et al., 2022)	Region	Social & Political Relations	Majority-focused	English
NormBank (Ziems et al., 2023)	Region	Social & Political Relations	Majority-focused*	English
NormDial (Li et al., 2023)	Region	Social & Political Relations	Majority-focused*	English, Chinese
CULTURAL-BENCH (Chiu et al., 2025)	Region	Food & Drink, Clothing & Grooming, Speech & Language, Social & Political Relations	Majority-focused	English
AMAMMERE (Acquaye et al., 2024)	Region	Food & Drink, Kinship, Social & Political Relations, The House	Majority-focused	English

Blend (Myung et al., 2024)	Region, Language	Food & Drink, Kinship, Social & Political Relations, Modern World	Majority-focused	English, Chinese, Spanish, Indonesian, Korean, Greek, Persian, Arabic, Azerbaijani, Sundanese, Assamese, Hausa, Amharic
CAMeL (Naous et al., 2024)	Language	Food & Drink, Clothing & Grooming, Names, The House, Modern World, Social & Political Relations, Physical World	Majority-focused	Arabic
FORK (Palta and Rudinger, 2023)	Region	Food & Drink, Social & Political Relations	Majority-focused	English
IndoCulture (Koto et al., 2024b)	Region	Food & Drink, Social & Political Relations, Kinship, Arts, Agriculture & Vegetation, Basic Actions & Technology, Emotions & Values	Majority-focused	Indonesian
COPAL-ID (Wibowo et al., 2024)	Region	Social & Political Relations, Speech & Language, Modern World, Physical World, Food & Drink, Animals	Majority-focused	Indonesian, English
XCOPA (Ponti et al., 2020)	Language	Social & Political Relations, Cognition	Majority-focused	Estonian, Haitian Creole, Indonesian, Italian, Southern Quechua, Swahili, Tamil, Thai, Turkish, Vietnamese, Chinese
MC2 (Zhang et al., 2024)	Region, Language	Speech & Language, Social & Political Relations, Arts, Modern World	Majority-focused	Tibetan, Uyghur, Kazakh, Mongolian
FoundaBench (Li et al., 2024f)	Region	Basic Actions & Technology, Arts, Modern World, Social & Political Relations, Physical World, Current & Historical Events	Majority-focused	Chinese
(Leong et al., 2023)	Language	Speech & Language, Arts, Social & Political Relations	Majority-focused	Indonesian, Tamil
Indonesian and Sundanese CommonsenseQA datasets (Putri et al., 2024)	Language	Food & Drink, Social & Political Relations, Physical World, Arts	Majority-focused	Indonesian, Sundanese

CLIcK (Kim et al., 2024)	Region	Social & Political Relations, Modern World, Physical World, Speech & Language, Current & Historical Events	Majority-focused	Korean
HAE-RAE Bench (Son et al., 2024)	Region	Speech & Language, Social & Political Relations, Physical World, Modern World	Majority-focused	Korean
BertaQA (Etxaniz et al., 2024)	Region, Language	Speech & Language, Arts, Physical World, Social & Political Relations, Modern World, Basic Actions & Technology, Current & Historical Events	Majority-focused	Basque, English
DOSA (Seth et al., 2024)	Region	Animals, Food & Drink, Clothing & Grooming, The House, Social & Political Relations, Arts, Physical World	Majority-focused	English
RoCulturaBench (Masala et al., 2024)	Region	Speech & Language, Arts, Physical World, Social & Political Relations, Modern World, Basic Actions & Technology, Current & Historical Events	Majority-focused	Romanian
DLAMA (Keleg and Magdy, 2023)	Region	Physical World, Social & Political Relations, Speech & Language, Modern World	Majority-focused	Arabic, English, Korean, Spanish
EnCBP (Ma et al., 2022)	Region	Speech & Language, Social & Political Relations	Majority-focused	English
CaLMQA (Arora et al., 2025)	Language	Basic Actions & Technology, Social & Political Relations, Food & Drink, Speech & Language, Physical World, Modern World, The Body, Current & Historical Events	Majority-focused	English, Arabic, Chinese, German, Hebrew, Hindi, Hungarian, Japanese, Korean, Russian, Spanish, Afar, Balochi, Faroese, Fijian, Hiligaynon, Kirundi, Papiamento, Pashto, Samoan, Tongan, Tswana, Wolof
FMLAMA (Zhou et al., 2025)	Region, Language	Food & Drink	Majority-focused	English, Chinese, Arabic, Korean, Russian, Hebrew
GEOMLAMA (Yin et al., 2022)	Region, Language	Social & Political Relations, Physical World, Food & Drink	Majority-focused	English, Chinese, Hindi, Persian, Swahili

MAPS (Liu et al., 2024a)	Language	Speech & Language	Majority-focused	English, German, Russian, Bengali, Mandarin Chinese, Indonesian
SeaEval (Wang et al., 2024a)	Region, Language	Social & Political Relations, Speech & Language	Majority-focused	English, Chinese, Indonesian, Spanish, Vietnamese, Malay, Filipino
NativQA (Hasan et al., 2025)	Language, Region	Animals, Modern World, Clothing & Grooming, Food & Drink, Physical World, Speech & Language, Arts, Names, Agriculture & Vegetation, Social & Political Relations	Majority-focused	Arabic, Assamese, Bengali, English, Hindi, Nepali, Turkish
OMGEval (Liu et al., 2024b)	Language	Social & Political Relations, Food & Drink, Names, Speech & Language	Majority-focused	Chinese, Russian, French, Spanish, Arabic
CPopQA (Jiang and Joshi, 2024)	Region	Social & Political Relations	Majority-focused	English
CUNIT (Li et al., 2024e)	Region	Food & Drink, Clothing & Grooming	Majority-focused	English
Candle (Nguyen et al., 2023)	Region, Religion, Socioeconomic Status	Food & Drink, Clothing & Grooming, Social & Political Relations	Majority-focused*	English
KorNAT† (Lee et al., 2024a)	Region, Language	Social & Political Relations, Speech & Language	Majority-focused	Korean
ArabicMMLU (Koto et al., 2024a)	Language, Region	Social & Political Relations, Physical World, Speech & Language	Majority-focused	Arabic
CMMLU (Li et al., 2024c)	Region, Language	Food & Drink, Social & Political Relations, Speech & Language	Majority-focused	Chinese
IndoMMLU (Koto et al., 2023)	Region, Language	Arts, Kinship, Speech & Language, Social & Political Relations	Majority-focused	Indonesian, Lampungic, Balinese, Makassarese, Banjarese, Madurese, Dayak Ngaju, Sundanese, Javanese
KMMLU (Son et al., 2025)	Region, Language	Social & Political Relations, Physical World, Current & Historical Events	Majority-focused	Korean
JMMMU (Onohara et al., 2025)	Region, Language	Arts, The House, Social & Political Relations, Current & Historical Events	Majority-focused	Japanese
PersianMMLU (Ghahroodi et al., 2024)	Region, Language	Kinship, Speech & Language	Majority-focused	Persian

TurkishMMLU (Yüksel et al., 2024)	Language	Speech & Language, Social & Political Relations	Majority-focused	Turkish
cuDialog (Cao et al., 2024b)	Region	Emotions & Values	Majority-focused	English
LOCALVALUEBENCH (Meadows et al., 2024)	Region	Emotions & Values	Majority-focused	English
LLM-GLOBE (Karinshak et al., 2024)	Region	Emotions & Values	Majority-focused	English, Chinese
CDEval (Wang et al., 2024b)	Language, Region	Emotions & Values	Majority-focused	English, German, Chinese
GlobalOpinionQA (Durmus et al., 2023)	Region	Emotions & Values	Diversity-aware	English, Russian, Turkish, Chinese
WorldValuesBench (Zhao et al., 2024)	Region, Socioeconomic Status	Emotions & Values	Diversity-aware	English
KorNAT† (Lee et al., 2024b)	Language, Region	Emotions & Values	Diversity-aware	Korean
County Tweet Lexical Bank	Region	Emotions & Values	Majority-focused*	English
dataset of 12k subreddits	Online Community	Emotions and Values	Majority-focused*	English
MABL (Kabra et al., 2023)	Language	Speech & Language	Majority-focused	Hindi, Yoruba, Kannada, Sundanese, Swahili, Indonesian, Javanese
(Ersoy et al., 2023)	Language	Speech & Language	Majority-focused	Arabic, Bengali, English, French, Spanish
AraDiCE (Mousi et al., 2024)	Region, Language	Speech & Language	Majority-focused	Arabic, Levantine dialect, Egyptian dialect, Gulf dialect

DialectBench (Faisal et al., 2024)	Language, Region	Speech & Language	Majority-focused	Albanian, Anglic, Arabic, High German, Italian Romance, Romance, Basque, Sinitic, Common Turkic, Southwestern Shifted Romance, Greek, Gallo-Rhaetian, Norwegian, Bengali, Navajo, Gallo-Italian, Kurdish, Komi, Serbian-Croatian-Bosnian, Tupi-Guarani, Modern Dutch, Eastern Romance, Frisian, Swahili, Saami, Armenian, Persian, Circassian, Italian Regional, Brazilian Portuguese, Sardinian, Sotho-Tswana, Sorbian, Hakka Chinese, Wu Chinese, Classical-Middle-Modern Sinitic, Galician, Occitan, Mirandese, Tigrinya
MultiPICo (Casola et al., 2024)	Language, Gender, Region, Ethnicity, Socioeconomic Status, Age	Speech & Language	Majority-focused	English, Spanish, Italian, French, Dutch, German, Hindi, Arabic, Portuguese
Vietnamese Central-Northern Dialect Text Transfer Corpus (Le and Luu, 2023)	Language, Region	Speech & Language	Majority-focused	Vietnamese
AC-EVAL (Wei et al., 2024)	Language, Time Period	Speech & Language	Majority-focused	Chinese
C3bench (Cao et al., 2024a)	Language	Speech & Language	Majority-focused	Classical Chinese
WYWEB (Zhou et al., 2023)	Language	Speech & Language	Majority-focused	Classical Chinese
GOLEM (Yarlott et al., 2024)	Ethnicity	Speech & Language	Majority-focused	English
(Toker et al., 2024)	Language, Time Period	Speech & Language	Majority-focused	Medieval Hebrew
Chumor 1.0 (He et al., 2024)	Region, Language	Speech & Language	Majority-focused	Chinese

Cards Against Humanity (Ofer and Shahaf, 2022)	Online Community	Speech & Language	Majority-focused	English
(Walsh et al., 2024)	Language	Speech & Language	Majority-focused	English
CIF-Bench (Li et al., 2024g)	Language, Region	Speech & Language	Majority-focused	Chinese
IndiBias (Sahoo et al., 2024)	Region, Gender, Religion, Age, Appearance, Socioeconomic Status	Mental Perception	Majority-focused	Hindi, English
Indian-BhED (Khandelwal et al., 2024)	Region, Religion, Socioeconomic Status	Mental Perception	Majority-focused	English
RuBia (Grigoreva et al., 2024)	Language, Region, Gender, Socioeconomic Status	Mental Perception	Majority-focused	Russian
SPICE (Dev et al., 2023)	Region, Religion, Socioeconomic Status, Gender	Mental Perception	Majority-focused	English
SeeGULL (Jha et al., 2023)	Region	Mental Perception	Majority-focused	English
CHBias (Zhao et al., 2023)	Language, Region, Gender, Age, Appearance, Sexual Orientation	Mental Perception	Majority-focused	Chinese
CBBQ (Huang and Xiong, 2024)	Region, Age, Gender, Ethnicity, Religion, Sexual Orientation, Socioeconomic Status, Health	Mental Perception	Majority-focused	Chinese
KoBBQ (Jin et al., 2024)	Region, Language, Age, Health, Gender, Appearance, Ethnicity, Sexual Orientation, Religion, Socioeconomic Status	Mental Perception	Majority-focused	Korean
JBBQ (Yanaka et al., 2025)	Region, Language, Age, Health, Gender, Appearance, Sexual Orientation	Mental Perception	Majority-focused	Japanese
MBBQ (Neplenbroek et al., 2024)	Language, Age, Gender, Health, Appearance, Socioeconomic Status, Sexual Orientation	Mental Perception	Majority-focused	English, Dutch, Spanish, Turkish
KOSBI (Lee et al., 2023)	Region, Language, Age, Gender, Religion, Health, Socioeconomic Status	Mental Perception	Majority-focused	Korean
GeoWAC (Borah et al., 2025)	Region	Mental Perception	Majority-focused	English
(Sankaran et al., 2024)	Gender	Mental Perception	Majority-focused	English
Chilean Dataset for Offensive Language Detection (Arango Monnar et al., 2022)	Region, Language	Speech & Language, Social & Political Relations	Majority-focused	Chilean Spanish

CREHate (Lee et al., 2024c)	Region, Language	Social & Political Relations	Majority-focused	English
KOLD (Jeong et al., 2022)	Language, Region, Online Community	Social & Political Relations, Speech & Language	Majority-focused	Korean
AraTrust (Alghamdi et al., 2025)	Language, Region	Social & Political Relations, The Body, Speech & Language, Modern World	Majority-focused	Arabic
XtremeSpeech (Maronikolakis et al., 2022)	Region	Speech & Language, Social & Political Relations	Majority-focused	Brazilian Portuguese, German, Hindi, Swahili, English
RTP-LX (de Wynter et al., 2025)	Region, Language	Social & Political Relations, Speech & Language	Majority-focused	Arabic, Hebrew, Indonesian, Danish, Norwegian, Swedish, Dutch, English, German, Russian, Ukrainian, Czech, Polish, Serbian, Spanish, Portuguese, French, Italian, Hindi, Thai, Kiswahili, Chinese, Japanese, Korean, Turkish, Finnish, Hungarian
WALLEDEVAL (Gupta et al., 2024)	Region, Language	Speech & Language, Social & Political Relations	Majority-focused	English, Singapore English, Hindi
COD (Majewska et al., 2023)	Language	Speech & Language, Social & Political Relations, Modern World	Majority-focused	Arabic, Indonesian, Russian, Kiswahili
MULTI3WOZ (Hu et al., 2023)	Region, Language	Names, Physical World, Social & Political Relations	Majority-focused	English, Arabic, French, Turkish
CulturalRecipes (Cao et al., 2024c)	Language	Food & Drink, Basic Actions & Technology	Majority-focused	English, Chinese
CAMT (Yao et al., 2024a)	Region, Language	Food & Drink, The House, Social & Political Relations, Physical World	Majority-focused	English, Chinese, French, Spanish, Hindi, Tamil, Telugu
GuyLingo (Clarke et al., 2024)	Region, Language	Speech & Language, Social & Political Relations	Majority-focused	Creolese, English
(Kadaoui et al., 2023)	Region, Language	Speech & Language	Majority-focused	Classical Arabic, Modern Arabic, Arabic dialects, English
PoetMT (Chen et al., 2024)	Language, Time Period	Arts, Speech & Language	Majority-focused	Classical Chinese, English

PAR3 (Thai et al., 2022)	Language	Arts, Speech & Language	Majority-focused	French, Russian, German, Spanish, Czech, Norwegian, Swedish, Portuguese, Italian, Japanese, Bengali, Tamil, Danish, Chinese, Dutch, Hungarian, Polish, Sesotho, Persian
AfroLingu-MT (Elmadany et al., 2024)	Language	Speech & Language, Social & Political Relations	Majority-focused	Arabic, English, French, Afar, Acholi, Afrikaans, Akan, Amharic, Bambara, Bassa, Bemba, Bhete, Ewe, Fang, Fon, Ge'ez, Hausa, Ibo, Kanuri, Kabiye, Kinyawanda, Kongo, Lugbara, Lingala, Luganda, Malagasy, Nande, Chichewa, Nyankore, Oromo, Nigerian Pidgin, Somali, Sesotho, Siswati, Swahili, Swahili (Congo), Ateso, Tigrinya, Tswana, Tsonga, Twi, Wolaytta, Wolof, Xhosa, Yoruba, Zulu
CoCoA-MT (Nadejde et al., 2022)	Language, Region	Speech & Language, Social & Political Relations	Majority-focused	English, French, German, Hindi, Italian, Japanese, Spanish
DNIC (Sandoval et al., 2023)	Region, Gender, Ethnicity	Names, Social & Political Relations	Majority-focused	English, Spanish, Russian, Arabic, Chinese
(Kantharuban et al., 2023)	Language, Region	Speech & Language	Majority-focused	Arabic, Mandarin, Portuguese, Finnish, German, Malay, Swahili, Spanish, Bengali, Georgian, Tagalog, Tamil, Telugu

- Cultural Group (Who to align with)
  - Region (37.9%)
    - Qiu et al. (2025); Fung et al. (2024); Rao et al. (2025); Shi et al. (2024); Zhan et al. (2024); Dwivedi et al. (2023); Yuan et al. (2024); Yu et al. (2024); Ziems et al. (2022); Schramowski et al. (2022); Ziems et al. (2023); Li et al. (2023); Chiu et al. (2025); Acquaye et al. (2024); Myung et al. (2024); Palta and Rudinger (2023); Koto et al. (2024b); Wibowo et al. (2024); Zhang et al. (2024); Li et al. (2024f); Kim et al. (2024); Son et al. (2024); Etxaniz et al. (2024); Seth et al. (2024); Masala et al. (2024); Keleg and Magdy (2023); Ma et al. (2022); Zhou et al. (2025); Yin et al. (2022); Wang et al. (2024a); Hasan et al. (2025); Jiang and Joshi (2024); Li et al. (2024e); Nguyen et al. (2023); Lee et al. (2024a); Li et al. (2024c); Koto et al. (2023); Son et al. (2025); Onohara et al. (2025); Ghahroodi et al. (2024); Cao et al. (2024b); Meadows et al. (2024); Karinshak et al. (2024); Wang et al. (2024b); Durmus et al. (2023); Zhao et al. (2024); Jha et al. (2023); Huang and Xiong (2024); Jin et al. (2024); Yanaka et al. (2025); Borah et al. (2025); Arango Monnar et al. (2022); Lee et al. (2024c); Jeong et al. (2022); Alghamdi et al. (2025); Maronikolakis et al. (2022); de Wynter et al. (2025); Gupta et al. (2024); Yao et al. (2024a); Clarke et al. (2024); Kadaoui et al. (2023); Sandoval et al. (2023); Kantharuban et al. (2023)
  - Language (28.9%)
    - Fung et al. (2024); Myung et al. (2024); Naous et al. (2024); Ponti et al. (2020); Zhang et al. (2024); Leong et al. (2023); Putri et al. (2024); Etxaniz et al. (2024); Arora et al. (2025); Liu et al. (2024a); Wang et al. (2024a); Hasan et al. (2025); Liu et al. (2024b); Lee et al. (2024a); Koto et al. (2024a); Li et al. (2024c); Koto et al. (2023); Son et al. (2025); Onohara et al. (2025); Ghahroodi et al. (2024); Yüksel et al. (2024); Wang et al. (2024b); Kabra et al. (2023); Ersoy et al. (2023); Mousi et al. (2024); Faisal et al. (2024); Casola et al. (2024); Le and Luu (2023); Wei et al. (2024); Cao et al. (2024a); Zhou et al. (2023); Toker et al. (2024); He et al. (2024); Walsh et al. (2024); Li et al. (2024g); Jin et al. (2024); Lee et al. (2023); Arango Monnar et al. (2022); Lee et al. (2024c); Jeong et al. (2022); Alghamdi et al. (2025); de Wynter et al. (2025); Gupta et al. (2024); Majewska et al. (2023); Hu et al. (2023); Cao et al. (2024c); Yao et al. (2024a); Clarke et al. (2024); Kadaoui et al. (2023); Chen et al. (2024); Thai et al. (2022); Elmadany et al. (2024); Nadejde et al. (2022); Kantharuban et al. (2023)
  - Gender (6.2%)
    - Fung et al. (2024); Casola et al. (2024); Grigoreva et al. (2024); Sahoo et al. (2024); Zhao et al. (2023); Huang and Xiong (2024); Jin et al. (2024); Yanaka et al. (2025); Neplenbroek et al. (2024); Lee et al. (2023); Sankaran et al. (2024); Sandoval et al. (2023); Dev et al. (2023)
  - Socioeconomic Status (5.7%)
    - Fung et al. (2024); Nguyen et al. (2023); Zhao et al. (2024); Casola et al. (2024); Sahoo et al. (2024); Khandelwal et al. (2024); Grigoreva et al. (2024); Huang and Xiong (2024); Jin et al. (2024); Neplenbroek et al. (2024); Lee et al. (2023); Dev et al. (2023)
  - Age (4.3%)
    - Fung et al. (2024); Casola et al. (2024); Sahoo et al. (2024); Zhao et al. (2023); Huang and Xiong (2024); Jin et al. (2024); Yanaka et al. (2025); Neplenbroek et al. (2024); Lee et al. (2023)
  - Religion (3.8%)
    - Fung et al. (2024); Nguyen et al. (2023); Sahoo et al. (2024); Khandelwal et al. (2024); Huang and Xiong (2024); Jin et al. (2024); Lee et al. (2023); Dev et al. (2023)
  - Ethnicity (3.3%)
    - Fung et al. (2024); Shi et al. (2024); Yarlott et al. (2024); Huang and Xiong (2024); Jin et al. (2024); Sandoval et al. (2023); Casola et al. (2024)
  - Health (2.4%)
    - Huang and Xiong (2024); Jin et al. (2024); Yanaka et al. (2025); Neplenbroek et al. (2024); Lee et al. (2023)
  - Sexual Orientation (2.4%)
    - Zhao et al. (2023); Huang and Xiong (2024); Jin et al. (2024); Yanaka et al. (2025); Neplenbroek et al. (2024)
  - Appearance (2.4%)
    - Zhao et al. (2023); Huang and Xiong (2024); Jin et al. (2024); Yanaka et al. (2025); Neplenbroek et al. (2024)
  - Online Community (1.4%)
    - Ofer and Shahaf (2022); Jeong et al. (2022)
  - Time Period (1.4%)
    - Wei et al. (2024); Toker et al. (2024); Chen et al. (2024)

Figure 6: Cultural Group (Who to align with) dimension of our cultural alignment framework with corresponding evaluation datasets.

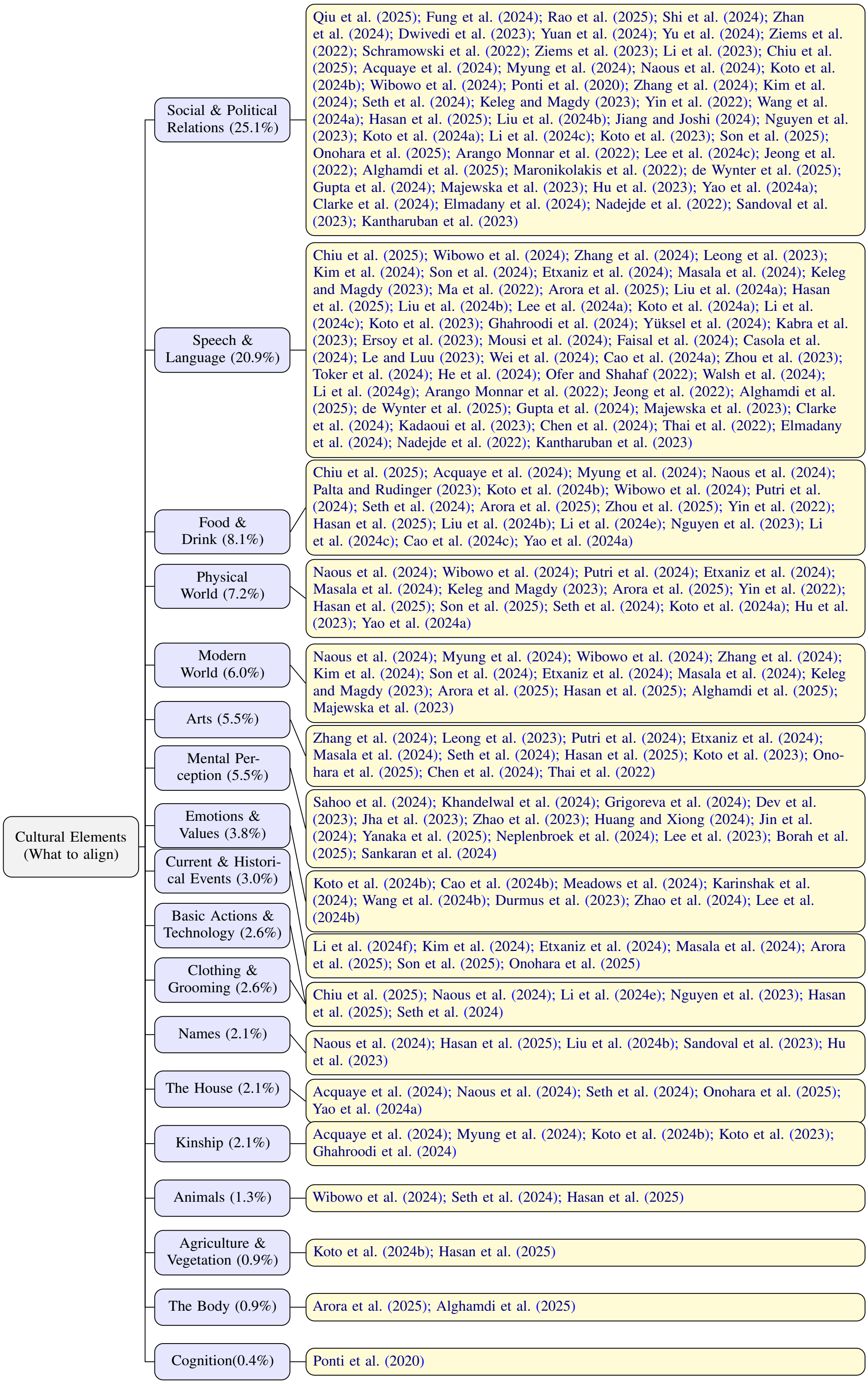

Figure 7: Cultural Elements (What to align) dimension of our cultural alignment framework with corresponding evaluation datasets.