Datasets and Methods for Improving the Cultural Capabilities of NLP Systems: A Survey

Tania Chakraborty[♠*] **Eylon Caplan**[♠*] **Zhaoqing Wu**[♠*] **Kevin Cushing**[♠] **Han Qin**[♠]
Shreya Havaldar[♣] **Dan Goldwasser**[♠]
[♠]Purdue University, [♣]University of Pennsylvania
{tchakrab,ecaplan,wu1828}@purdue.edu

Abstract

In recent years, there has been a surge of interest in Cultural NLP, with substantial efforts to create globally inclusive NLP systems. The rapid growth of literature in this field makes it difficult to track trends in methods and data resources. To address this, we survey over 375 papers to answer three complementary questions: (1) What *Cultural Capabilities (CCs)* are being targeted in NLP systems? (2) How are *cultural data resources* being created? and (3) What *methods* are being used to improve the CCs of those systems? We discuss trends observed across the three questions, and identify relevant research gaps. To facilitate further research in this field, we release our full list of surveyed papers, in the form of an interactive web interface, CULTUREMINE[1], which includes a feature to allow researchers to add their work; we hope this facilitates future research and proves to be a valuable resource for the Cultural NLP community.

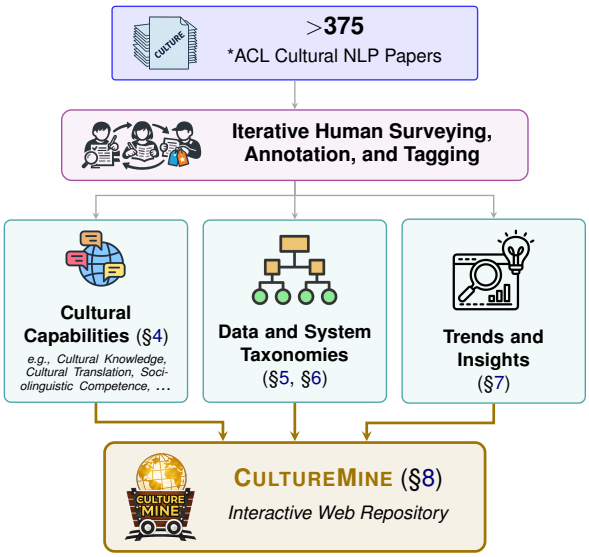

Figure 1: An overview of our surveying pipeline. We conduct iterative reading and manual annotation of over 375 papers, resulting in (1) our categorization of Cultural Capabilities, (2) new cultural Data and System taxonomies, and (3) key insights for the field. All annotated papers and findings are synthesized into CULTUREMINE, an interactive web repository for the community.

1 Introduction

"No people come into possession of a culture without having paid a heavy price for it."

— James Baldwin

Language and culture are fundamentally interdependent; language both reflects, and is shaped by culture (Sapir, 1929). Cultural NLP, which studies this intersection, has seen rapid growth in recent years; As of March 2026, searching the ACL Anthology for the word "culture" or "cultural" yields more than 700 papers from the last two years alone! This volume of emergent work makes it difficult for researchers to keep abreast of novel methodologies and data resources for Cultural NLP. To address this, we survey over 375 papers and provide an overview of (1) what Cultural Capabilities (CCs) are being targeted for improvement (§4), (2) methods utilized for **creating** cultural datasets (§5), and (3) methods proposed to **improve** the CCs of NLP systems (§6).

To help navigate the volume of papers, we first group them by contribution type—*datasets*, and *systems*, and then each contribution is also mapped to a CC, based on what cultural competency it targets. Within each contribution type, we propose a taxonomy to organize papers, and answer the questions: (1) For dataset contributions, how are cultural datasets created? and (2) For system contributions, what methods are adopted to improve the CCs of NLP systems? In contrast to other surveys in this field (discussed in §2), we offer a technical overview of the field and propose an organization scheme to navigate a voluminous and rapidly grow-

[*]Equal contribution.
[1]Accessible now at https://culture-in-nlp.pages.dev/.

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 198–248

ing body of literature.

Furthermore, we release our full list of surveyed papers through an interactive web interface, CULTUREMINE; each paper is manually annotated with our taxonomy—in addition to metadata such as cultural proxy (e.g., languages, countries). CULTUREMINE also includes a submission form that researchers can use to add their work to the collection[2]. We hope this will be a valuable resource for the cultural NLP community.

Our contributions are as follows:

- We survey over 375 papers, organize them based on our proposed taxonomy, and additionally tag them with rich cultural metadata.
- We identify trends in methodologies to provide the reader with an overview of the field, and propose future directions for cultural NLP.
- We release CULTUREMINE as a resource for the community. We welcome additions to CULTUREMINE to foster collaborative research and streamline the discovery of relevant work.

2 Related Surveys

There are some highly relevant surveys for cultural NLP, offering insights into definitions of culture and how culture is operationalized in NLP. Liu et al. (2025a); Adilazuarda et al. (2024) discuss how culture is defined, and survey commonly adopted proxies of culture. Zhou et al. (2025c) offer a nuanced perspective on how cultural NLP systems should be designed, based on theories from sociocultural linguistics. In this paper, we don't redefine culture, and refer readers to the surveys mentioned above for this. To select papers we survey, we deferred to the authors to determine whether their work is cultural NLP or not (§3 for details). Our survey adopts a complementary, technical perspective. Our main goal is to present an overview of recent **technical advances** for improving the cultural capabilities of NLP systems.

A closely related survey is Pawar et al. (2025), which surveys cultural awareness in language models, with a focus on data resources, and provides a comprehensive overview of resources for the community. In contrast, we survey NLP systems broadly without limiting the survey to language models. Additionally, our focus is on the methods, both for creating data resources as well as improving NLP systems. Finally, more than half the papers in our survey are from 2025, and thus not included in Pawar et al. (2025).

[2]Submissions to CULTUREMINE are reviewed by the authors to ensure tag consistency.

3 Literature Collection and Annotation

In this section we provide a brief overview of our literature collection and paper annotation strategy. Papers were added manually by authors as well as automatically via a keyword scrape of the ACL Anthology covering the past five years (up to and including 2025). After the initial keyword scrape, we utilized an LLM-assisted filtering pipeline to retain papers that met our criteria; they proposed either a novel methodology or dataset for cultural NLP.

All remaining papers were manually filtered by the authors to filter out works whose primary contributions were sociolinguistic or sociological rather than computational (see §4.2 for survey scope). This resulted in a final corpus of over 375 papers, which were read and annotated by the authors. We held regular meetings to refine the definitions of our taxonomies; All papers were reviewed to ensure strict adherence to definitions in §4.1, §5, and §6. Full details regarding our search queries, LLM filtering pipeline, and consensus protocol are provided in Appendix A. In the subsequent sections, we cite only representative works, but a full list of papers can be found in the Appendix F, Table **??**.

4 Definitions

In this section, we formally define the lens through which we analyze the surveyed literature. We define a traditional NLP *task* as a well-defined computational problem specifying the behavior a system should exhibit. In contrast to standard tasks (e.g., classification, text generation), in this paper we categorize works based on the *Cultural Capability* they address. The distinction is clarified and further explained in the following subsection.

4.1 Defining Cultural Capabilities

We define a **Cultural Capability (CC)** as an abstraction over NLP tasks to answer: *"What specific cultural behavior is this dataset measuring, or is this system improving?"*

A CC may be measured via various NLP tasks, and a given NLP task can be used to measure various CCs. To illustrate, a system's Cultural Knowl-

edge could be measured via the NLP task of QA (e.g., "What foods are strictly prohibited in a kosher diet?"), but could also be measured via a recipe generation task, where outputting a dish with pork is inherently penalized as a factual error.

The motivation for focusing on CCs rather than on standard NLP tasks was made early in survey phase. It enabled us to identify what kinds of CCs are being targeted by the community, and which ones may be overlooked. Looking at contributions through the lens of CCs instead of NLP tasks allows for a more nuanced understanding of how culturally competent current NLP systems are.

The CCs were not pre-defined, but rather discovered through our bottom-up survey of the literature. We identified nine core CCs that the NLP community is actively working to measure and improve:

1. **Cultural Translation:** the ability to convert text in one language to another, while preserving/accounting for cultural nuance and artifacts.
2. **Survey-based Cultural Alignment:** the capability of predicting the distributions of answers to value surveys conditioned on particular cultures (e.g., World Value Survey).
3. **Value-driven Cultural Alignment:** the capability to behave in such a way that agrees with a particular culture's values, or to switch between behaviors aligning with target cultures.
4. **Cultural Knowledge:** the ability to know, recall, understand, and use textual knowledge about a particular culture or cultures.
5. **Multimodal Cultural Knowledge:** the ability to know, recall, understand, and use visual knowledge about a particular culture or cultures.
6. **Sociolinguistic Competence:** the ability to understand, produce, or use language in such a way that it aligns with a particular culture or cultures' expected linguistic behavior.
7. **Cultural Safety and Harm Reduction:** the ability to understand, censor, correct, or avoid text that contains harms or potential harms in accordance with a target culture or cultures.
8. **Cultural Education:** the ability to aid in educative processes, conditioned on a target culture.
9. **Computational Cultural Representation:** the capability to computationally represent, compare, modify, or edit representations of culture or its features.

4.2 Scope and Contribution Categorization

Survey Scope: We limit the scope of this survey to include only works whose overarching aim is to improve or measure these CCs of NLP systems. As a result, we excluded works that use existing NLP models for computational social science (CSS). For example, Garimella et al. (2016) use NLP methods to answer the question *How are the same words used differently in different cultures?* This is valuable work but falls outside the scope of this paper. We view CSS as a vital pillar of Cultural NLP, but constrain our taxonomy specifically to those works that contribute a dataset or system advancement.

Contribution Types: Within the papers that meet our criteria, we first categorized them by their contribution type: **Datasets** and **Systems**. Note that many papers have both types of contributions. The main motivation for this categorization was to make it easier to analyze the broad field of Cultural NLP and extract insights about community focus and trends. Additionally, it aids researchers in efficiently finding resources via our companion webpage CULTUREMINE. In §5, we explore the Dataset branch, analyzing the methodologies used to curate cultural data. In §6, we analyze the Systems branch, detailing the NLP methods used for Evaluation, Data Generation, and Improvement.

5 Datasets for Cultural Capabilities

We surveyed 320 dataset papers, and analyze *techniques* and *sources* for dataset creation, followed by insights into common dataset creation pipelines. The goal of this section is to leave the reader with an overview of *how culture is being injected into datasets*.

5.1 Dataset Creation Methods

We broadly categorize cultural NLP dataset creation into four approaches: human-generated data (§5.1.1), adaptation from existing language data sources (§5.1.2), datasets grounded in cultural research and frameworks (§5.1.3), and synthetic data generation (§5.1.4). Figure 2 shows our taxonomy.

5.1.1 Human-Sourced Data

Crowdsourcing workers are widely used for large-scale data collection and annotation requiring general human judgment rather than specialized cultural expertise: asking contributors to submit diverse images or questions (Cahyawijaya et al., 2025b; Arora et al., 2025), collecting opinions through voting (Falk et al., 2024), rating (Casola et al., 2024), or abusive content labeling (Muhammad et al., 2025).

Cultural experts or native speakers are employed

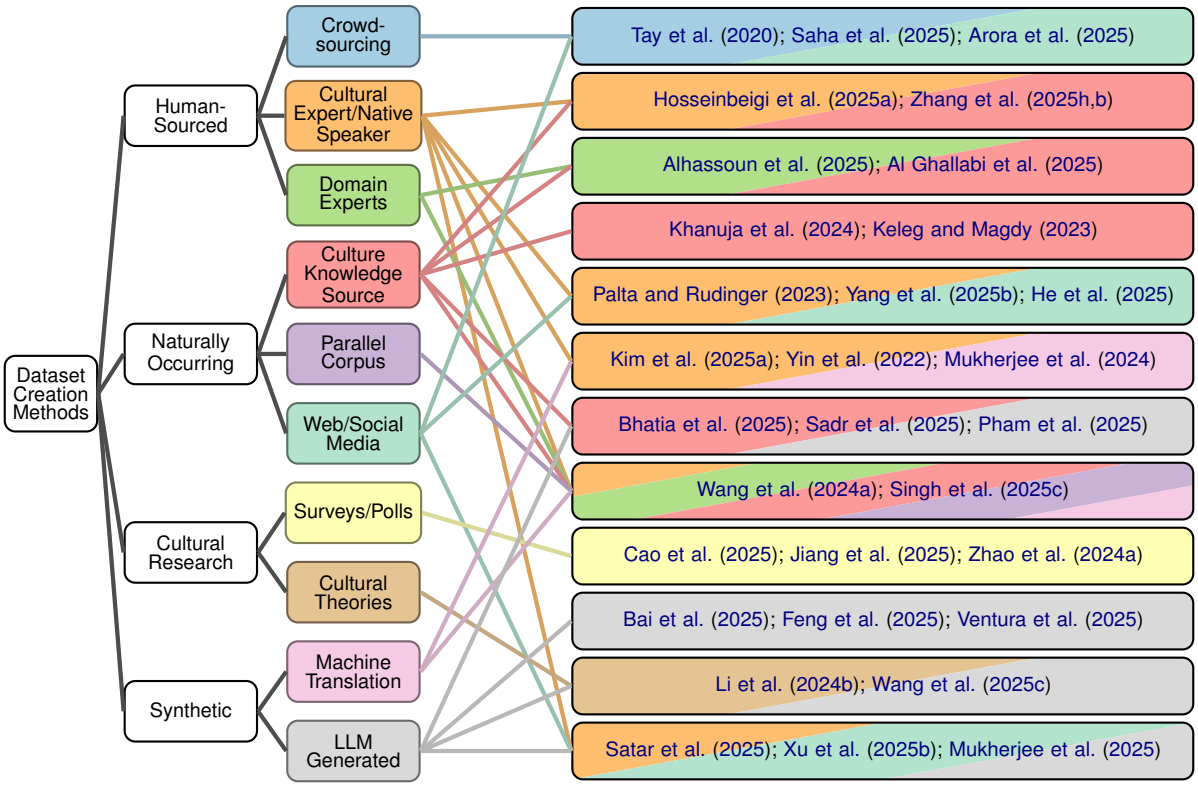

Figure 2: How are Cultural Datasets being created? Our taxonomy for Dataset Creation methods (solid color boxes), and example papers employing these methods (striped color boxes). Note that a paper can combine any number of data creation methods.

for tasks requiring cultural or linguistic expertise, such as labeling cultural aspects (Maji et al., 2025b; Alwajih et al., 2025a; Xie et al., 2025), creating culturally grounded content (Zhan et al., 2024; Guo et al., 2025), or assisting with translation (Kim et al., 2025a; Montalan et al., 2025).

Domain experts help ensure datasets align with the domains: law specialists developing annotation guidelines (Ullah et al., 2024), historians and archaeologists ensuring artifact accuracy (Ghaboura et al., 2025), and experts defining dataset taxonomies (Vasilev et al., 2025).

5.1.2 Naturally Occurring Sources

Web data provides unstructured cultural signals in online reviews (Zou et al., 2025) and social media to capture cross-cultural language use (Kumar and Jurgens, 2025; Wuraola et al., 2024; Abdelkadir et al., 2024; Kiesel et al., 2022; Liu et al., 2025e), and images for visual QA and reasoning (Liu et al., 2021, 2025e; Bayramli et al., 2025).

Parallel Corpora contain comparable content across languages, such as culturally relevant entities (Yao et al., 2024a; Conia et al., 2024), rules (Haberland et al., 2024), and topical images (Schneider and Sitaram, 2024).

Culture knowledge sources offer explicit and fine-grained cultural concepts and artifacts, reflecting the highest level of culture sensitivity: modifying existing datasets for a particular culture (Wang et al., 2024d; Grandury et al., 2025; Son et al., 2025), and filtering pre-existing cultural resources (Dai et al., 2025). Educational and domain-specific materials are also widely used: children's books (Khanuja et al., 2024), exams (Cheng et al., 2025), e-learning platforms (Pramodya et al., 2025), research journals (e.g., PubMed) (Nimo et al., 2025), literature (AbuHaija et al., 2025), and Wikipedia (Magdy et al., 2025; Bhatia et al., 2025).

5.1.3 Cultural Research

Culture Value Surveys, such as the World Values Survey (WVS) (Inglehart et al., 2000; Haerpfer et al., 2022), provide standardized value-oriented questions and empirical responses that researchers use to construct datasets: expanding question sets (Xu et al., 2025a), predicting answer distributions (Cao et al., 2025), or selecting dialogue topics based on response patterns (Ma et al., 2025b).

Cultural/Social Science Theories also guide

dataset creation. Frameworks such as Hofstede's Cultural Dimensions (Hofstede, 1984), the Theory of Basic Human Values (Schwartz, 1992), and negotiation theory (Aslani et al., 2016) are used to refine value taxonomies for question generation (Cahyawijaya et al., 2025a), categorize values (Yao et al., 2024b), and guide annotators (Hale et al., 2025).

5.1.4 Synthetic Generation

Machine Translation is used to expand datasets across languages. Work applies Google Translate to translate simple sentences (Belay et al., 2025; Yin et al., 2022) and literature (Thai et al., 2022). LLMs are used for translation (Onohara et al., 2025; Kim and Kim, 2025), augmentation (Masala et al., 2024), and verification of cultural alignment (Putri et al., 2024). Other work uses specialized translation models NLLB Team et al. (2022) for quality check (Aakanksha et al., 2024; Nguyen et al., 2024b), or trains models to support low-resource dialects (Mousi et al., 2025).

LLM Generation is increasingly used in dataset construction: generating culturally specific content such as stereotypes (Jha et al., 2023; Sahoo et al., 2024), moral scenarios (Liu et al., 2024a; Dey et al., 2025), and norms (CH-Wang et al., 2023); standardizing data formats (Chiu et al., 2025; Umbet et al., 2025), and adapting content across cultural contexts (Joshi et al., 2025; Putri et al., 2024). Beyond text, LLMs are used to synthesize images (Kim et al., 2025b), generate image captions (Bai et al., 2025), and label videos (Chen et al., 2025d).

5.2 Insights into Dataset Creation Pipelines

Our analysis reveals that modern cultural dataset creation rarely relies on a single, isolated method. Figure 2 illustrates common combinations of dataset creation approaches (additional details in Appendix C.3) with representative examples. We identify two dominant pipelines.

Pipeline 1: Source-Grounded Human Curation combines cultural knowledge sources with human expertise. Researchers typically adapt materials from books, image collections, and local websites, then ask humans to create prompts (Isbarov et al., 2025), questions (Nayak et al., 2024), and QA pairs (Limkonchotiwat et al., 2025), or annotate through quality checks (Kim et al., 2024a; Kim and Lee, 2025), label assignment (Maji et al., 2025a; Zhang et al., 2025b), or culture-specific translation (Winata et al., 2025), and enrich existing cultural datasets with expert knowledge (Cheng et al., 2025; Romanyshyn et al., 2024).

Pipeline 2: LLM Co-Creation uses LLMs as as generative **amplifiers** or **structuring tools** in dataset creation, with human input, theoretical frameworks, or existing language resources, including expanding seeded content with additional situated examples (Xu et al., 2025a; Zhan et al., 2024; Li et al., 2025); generating culturally diverse content followed by human verification (Urailertprasert et al., 2024; Qiu et al., 2025; CH-Wang et al., 2023); extending topic coverage (Wang et al., 2024f; Cahyawijaya et al., 2025a; Chiu et al., 2025); and extracting or generating additional information from existing language sources (Bhatia et al., 2025; Arnardóttir et al., 2025).

Despite the common use of these pipelines, we find several exceptions across **Cultural Capabilities** (Figure 9). Sociolinguistic Competence emphasize crowdsourced interactions to capture natural language use while avoiding explicit cultural cues that could create evaluation shortcuts. Cultural Translation rely more on language professionals (Akinade et al., 2023; Tapo et al., 2025a; Zhang et al., 2025e), and draw on web-based multilingual content, which captures contemporary expressions that often remain untranslated. Cultural Safety and Harm Reduction similarly depend on web data, particularly social media, to capture potentially harmful language needed for moderation (Maronikolakis et al., 2022; Mia et al., 2025). In contrast, Computational Cultural Representation use LLM-based synthetic generation to construct structured cultural knowledge that is difficult to obtain directly from raw data (Acquaye et al., 2024; Ziems et al., 2023; Pujari and Goldwasser, 2025).

We also identify the **conceptual rigor trade-off** in the literature. Compared to other data construction approaches, culture research based methods are used less frequently. This pattern suggests that current cultural NLP dataset creation relies more heavily on coarse-grained, **proxy-based data** sources than on structured guidance from well-established **conceptual frameworks in social science**. Such a trend highlights a potential trade-off between scalability and practical convenience on the one hand, and depth, conceptual rigor, and precision of cultural representation on the other.

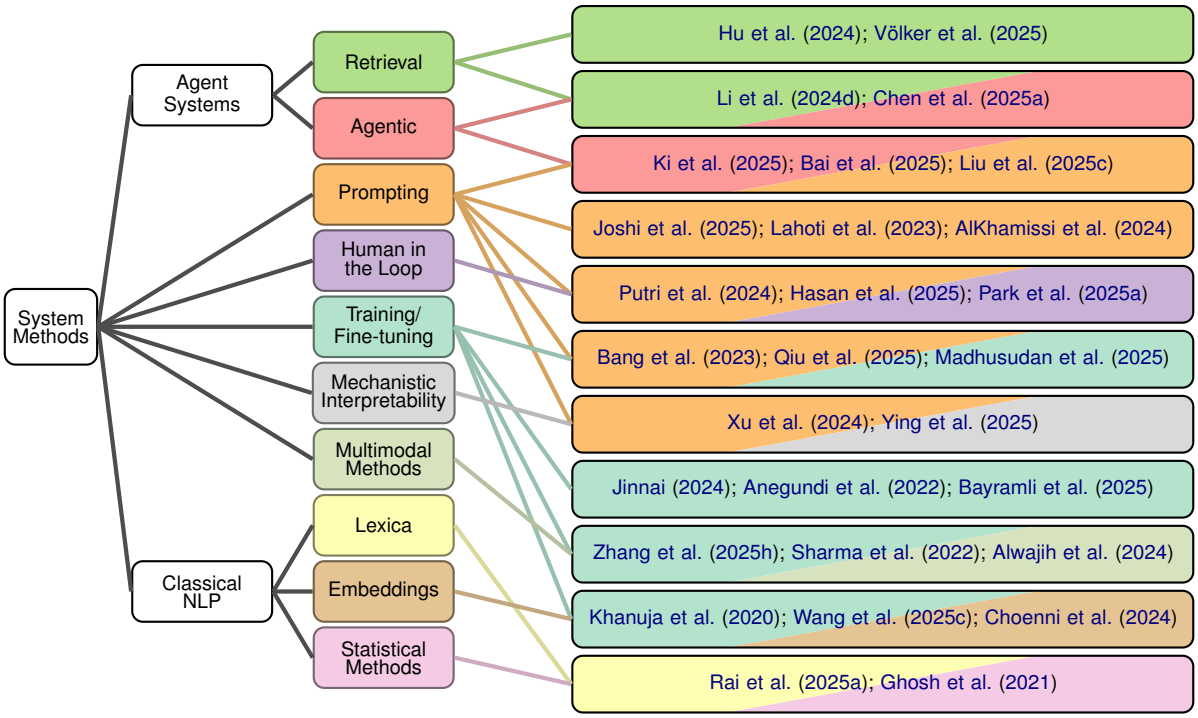

Figure 3: What methods are used to improve Cultural Capabilities? Our taxonomy of Systems (solid color boxes), and example papers employing these methods (striped color boxes). Note that a paper may combine any number of methods.

6 Systems for Cultural Capabilities

In this section we first discuss three different *kinds* of systems that we identified during our survey (§6.1), and then summarize some technical approaches adopted to improve CCs in NLP systems (§6.2).

6.1 Methodology Goals

Among all the papers that contributed a system, we identified three distinct overarching goals:

System Improvement for CC: Works that seek to directly *improve* an NLP System's CCs, e.g., (Ma et al., 2025b; Cao et al., 2025; Li et al., 2024d).

System Evaluation for CC: Works that propose a novel system to *evaluate* an NLP System's CCs, e.g., (Zhao et al., 2025; Mukherjee et al., 2023; Zhang et al., 2025a).

Cultural Data Generation: Works that propose a generalizable framework to *generate* culturally informed data, e.g., (Keleg and Magdy, 2023; Fung et al., 2023; Hasan et al., 2025).

The three system goals each have a very different impact on research, which makes this an important distinction. We allow the ability to query papers by this distinction in CULTUREMINE, and discuss insights from this classification in §7.

6.2 Technical Approaches to Methodology

Fig. 3 illustrates the taxonomy used to classify methods based on their technical approaches. Prominent themes for each approach are discussed below.

6.2.1 Prompting Based Methods

We identified over 70 papers that contributed a novel system utilizing prompting in various ways. One common approach is to vary the language of the prompt, to either probe a model (Kim and Kim, 2025; Zhao et al., 2025), or to encourage culturally aligned behavior (Feng et al., 2024a). Other methods use specialized prompts, such as injection of cultural knowledge (Shaikh et al., 2023; Ma et al., 2025b), guiding principles (AlKhamissi et al., 2024), multi-step prompting (Hobson et al., 2024), and cloze templates (Ramezani and Xu, 2023).

6.2.2 Agent Systems: Retrieval and Multi-Model Systems

We found Agent Systems, often with a "retriever agent" to be a common system proposed. What is the motivation for using an agentic framework? We find two recurring themes: (1) To allow for culturally diverse interactions and output (Ki et al., 2025; Li et al., 2024d; White et al., 2024; Bai et al., 2025;

Feng et al., 2024b), (2) To allow for specialized roles in complex systems made up of smaller modules (Anik et al., 2025; Wu et al., 2024; Yuan et al., 2024; Liang et al., 2025). For systems that utilized a retriever, what was the role of the retriever? The most common role we found was to retrieve diverse kinds of cultural information such as entities (Conia et al., 2024), recipes (Hu et al., 2024), medical text (Calvo-Bartolomé et al., 2025), poetry (Chen et al., 2025a). This would be explained by the fact that the most common CC associated with retrievers was Cultural Translation.

6.2.3 Human in the Loop Systems

We identified 13 papers that proposed a system involving a human in the loop. The main role of humans in these systems was to provide cultural expertise, which can look like judgments and corrections (Pujari and Goldwasser, 2025; Ziems et al., 2025), culturally informed seed data (Hasan et al., 2025; Rachamalla et al., 2025; Putri et al., 2024), adversarial input (Chiu et al., 2025), and even outside culture judgment (Park et al., 2025b).

6.2.4 Training Based Methods

Many papers propose novel training paradigms to improve the CC of NLP Systems. By far the most common one we identify is SFT on LLMs for some form of cultural alignment (Choenni et al., 2024; Cao et al., 2025; Xu et al., 2025a; Dai et al., 2025). Another common trend we notice is that of training smaller models for specialized tasks, such as Bert (Devlin et al., 2019) to identify values (Kiesel et al., 2022) or predict stereotypes (Kim and Johnson, 2025), e5 model (Wang et al., 2024b) to align sociocultural concepts (Wang et al., 2025c), or evaluation metrics trained to reflect human judgment (Bayramli et al., 2025). Finally, there are several papers that introduce novel architectures for specialized tasks such as subjective prediction (Parappan and Henao, 2025), predicting social relationships from videos (Zhang et al., 2025h), and transfer learning across languages (Ringel et al., 2019).

6.2.5 Classical NLP

In this subsection we discuss approaches which are often paired; embeddings, lexica, and methods over them like clustering, summarization.

Lexica: We identified 2 main ways that lexica are used in this field: (1) discovering cultural differences like ideologies (Milbauer et al., 2021), perspectives (Gutiérrez et al., 2016), expressions of mental health (Rai et al., 2025a), and personal values (Wilson et al., 2016), and (2) quantifying linguistic aspects like style (Havaldar et al., 2023a), bias, (Naous and Xu, 2025; Friedman et al., 2019) cultural awareness (Zhao et al., 2025; Caplan et al., 2025), harm (Menis Mastromichalakis et al., 2025), and concepts (Li and Zhang, 2023).

Embeddings: Lexical methods are often paired with embedding based methods for purposes like enriching the lexica (Havaldar et al., 2024, 2023a), comparing similarities or differences between cultures (Sun et al., 2021; Milbauer et al., 2021), and detecting cultural biases (Friedman et al., 2019). In contrast, (Rai et al., 2025a; Caplan et al., 2025) specifically avoid embeddings to prevent issues arising from bias in embeddings. Other works, exploit the bias in embeddings to measure abstract concepts like values, ideologies and identities (Cahyawijaya et al., 2025a; Milbauer et al., 2021; Ventura et al., 2025; Havaldar et al., 2024), and perceptions and pragmatics (Sun et al., 2021; Lin et al., 2018; White et al., 2024). Another common use of embeddings is to create a shared representation space for multiple cultures, in order to discover similarities and differences between them (Choenni et al., 2024; Zhou et al., 2023).

Statistical Methods: These methods encompass a wide variety of tools like clustering, topic modeling, evaluation metrics etc. We highlight two interesting recurring themes; the first is the use of clustering and topic modeling to discover common themes from large noisy data (Cuevas et al., 2025; Milbauer et al., 2021; Hobson et al., 2024; Pujari and Goldwasser, 2025), and the second is the use of various metrics to measure abstract concepts like cultural alignment (Wang et al., 2024c), ideologies (Milbauer et al., 2021), and values (Xu et al., 2024).

6.2.6 Mechanistic Interpretability

We identified few methods using mechanistic interpretability, which indicates it being an underexplored method for cultural NLP. The papers we did identify propose novel ways to answer why models display the cultural tendencies that they do (Xu et al., 2024; Ying et al., 2025).

6.2.7 Multimodal Methods

Several papers we surveyed proposed methods that included modalities beyond text. A very interesting line of work uses vision to account for low resource languages that do not have a surplus of textual data (Li and Zhang, 2023; Chen et al., 2024; R et al.,

2025). Another motivation for such methods is cultural competence over non-text modalities like vision (Sharma et al., 2022; Alwajih et al., 2024; Zhang et al., 2025c,h; Khanuja et al., 2025).

7 Observations and Recommendations for Future Work

7.1 Trends in Proposed Future Work

We analyzed the future work sections of the surveyed papers and found that community-proposed directions predominantly fall into two themes (details in Appendix B). First, researchers frequently call for **scaling cultural and linguistic coverage** to address data scarcity, often advocating for richer, multimodal benchmarks (Wang et al., 2024a; Maji et al., 2025a). Second, there is a strong push to **model cultural complexity** more accurately. Authors emphasize moving away from static, monolithic labels by operationalizing culture as a dynamic distribution (Havaldar et al., 2023a; Ziems et al., 2023), incorporating intersectional and fine-grained geographic variables beyond language (Falk et al., 2024; Koto et al., 2024), and developing evaluation frameworks explicitly accounting for pluralism (Zhou et al., 2023; Miehling et al., 2025).

7.2 Our Observations and Recommendations

In this section we discuss four broader patterns that offer promising avenues for the community.

Prioritizing system improvements over additional dataset creation. Dataset creation substantially outpaces the development of methods that directly improve CCs. This gap is especially pronounced for Cultural Knowledge (>100 surveyed **dataset** contributions), Cultural Safety and Harm Reduction (>60), and Value-driven Cultural Alignment (>50), each of which has fewer than 20 surveyed **system** improvement contributions. The abundance of such datasets (Singh et al., 2025c; Sahoo et al., 2025; Bui et al., 2025b; Shetty et al., 2025) indicates that the community has built a robust foundation for measuring these capabilities. We recommend that future efforts emphasize developing systems specifically designed to improve performance on these existing datasets, as done by (Ki et al., 2025; Wang et al., 2025b; Feng et al., 2024a; Parappan and Henao, 2025).

Reusing existing data generation frameworks for scalable data collection. When the creation of new data is desired, we observe an opportunity to reduce redundant effort. We identified several sophisticated, generalizable Cultural Data Generation systems (Ziems et al., 2025; Caplan et al., 2025). These frameworks are designed to produce datasets dynamically by conditioning on target variables (e.g., country, language, or source collection). However, we noticed limited follow-up reuse of these pipelines. Leveraging these existing generative frameworks can accelerate research when specific cultural data is scarce, providing a scalable alternative to curating datasets from scratch.

Developing specialized evaluation methods for implicit cultural nuances. Our survey suggests that measurement paradigms for different CCs are evolving at different rates. We use the ratio of novel *Evaluation System* papers to *Dataset* papers to approximate this focus: a higher ratio indicates active development of bespoke measurement frameworks, whereas a lower ratio implies a reliance on existing metrics. For example, capabilities involving abstract constructs, such as Value-driven Cultural Alignment (0.30) and Sociolinguistic Competence (0.21), exhibit relatively high ratios. This suggests that **standard metrics often fall short** for these CCs, prompting researchers to build specialized evaluators (Yao et al., 2024b; Shen et al., 2025; Casola et al., 2024; Ying et al., 2025). Conversely, capabilities like Multimodal Cultural Knowledge (0.02) and Cultural Translation (0.11) show significantly lower ratios, as they frequently **rely on established, reference-based metrics** like BLEU or ROUGE. While these standard metrics provide valuable baselines, we encourage the continued development of specialized evaluation methodologies for these domains as well.

This opportunity is particularly visible in multimodal work. Existing literature provides excellent coverage of visually salient cultural artifacts, such as food, clothing, and landmarks (Nayak et al., 2024; Bhatia et al., 2024; Maji et al., 2025c). Moving forward, the field is well-positioned to tackle subtler, "below-the-iceberg" phenomena (Hall, 1976) essential for real-world interaction, such as conversational grounding, gestures, paralinguistic interpretation, and cross-dialect understanding, e.g., (Sasu et al., 2025; Zhang et al., 2025h).

Evaluating and optimizing multiple CCs jointly. Our taxonomy defines CC as a collection of distinct competencies (§4.1), yet current research predominantly isolates them. Among surveyed papers con-

tributing *system improvements*, 82% target exactly one CC; for *system evaluation*, the figure is 94%. While this focused scope is necessary for foundational research, real-world deployments require systems to juggle **multiple CCs simultaneously**. Improvement on one CC does not inherently transfer to another (Zhang et al., 2025d); a culturally capable system deployed in the world may need to retrieve accurate cultural facts, communicate appropriately, and adapt to pluralistic norms concurrently. We encourage future work to report cross-CC transfer and trade-offs, and to develop comprehensive benchmarks that evaluate multiple CCs jointly.

8 CULTUREMINE: Community Resource

In this section we provide a brief overview of CULTUREMINE, which contains all the papers we surveyed, tagged with the dataset taxonomy and/or the systems taxonomy and relevant CCs. Additionally, each paper is tagged for culturally relevant metadata such as what proxy it used for culture (language, country, other). When applicable, papers are also tagged for the exact languages or regions that they account for. All the papers can be queried via drop-down filters, which makes it very easy for researchers to find a collection of work that is relevant for them! The top section of CULTUREMINE has a dynamic visualization giving an overview of papers and updates based on selected fields.

Additionally, we include a submission form that researchers can fill out to add more papers. We intend to keep the UI updated regularly, and hope that it can be a true mine for cultural NLP!

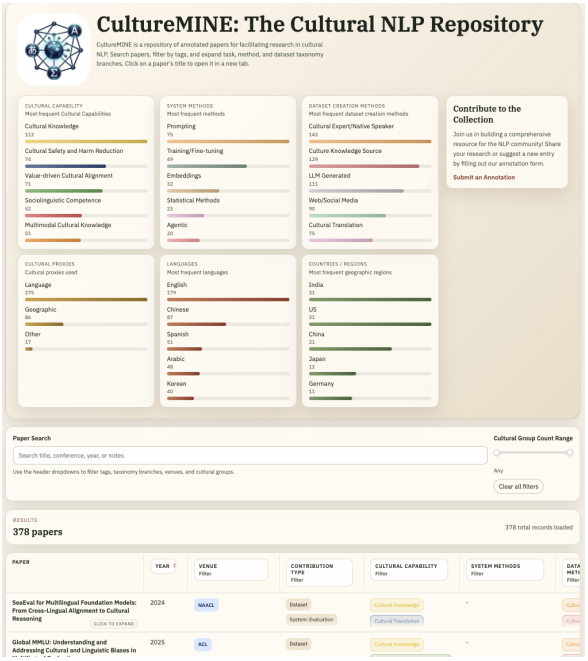

Figure 4: CULTUREMINE interface has dynamic visualizations in the top section. The fields in the bottom, which include our taxonomies for datasets and systems, can be used as filters to find relevant papers. The filters also include information such as cultural proxies used, what cultures, and what languages!

9 Conclusion

We surveyed over 375 papers, provided an overview of recent technical methods for improving the CCs of NLP systems, proposed taxonomies to organize contributions and released our surveyed papers, tagged with our taxonomies and other rich metadata via CULTUREMINE. We hope this survey and CULTUREMINE will be a valuable resource for the cultural NLP community.

Limitations

Despite our effort to provide a comprehensive overview of the cultural capabilities of NLP systems, several limitations remain.

Scope and Source Coverage. Our survey is not exhaustive. Research on culture in NLP is distributed across multiple venues and related fields, making comprehensive coverage infeasible. We therefore focus on papers that directly address our central question: how cultural capabilities are technically represented, operationalized, or incorporated in NLP systems and research designs. For the surveyed papers, we systematically collected them from the ACL Anthology, which serves as the primary repository. As a result, research published outside the ACL Anthology may be underrepresented in our survey.

Exclusion of NLP for Social Science. An important related area not included in this survey is NLP for social science. This line of research applies natural language processing methods to study social phenomena such as cultural variation, social meaning, and contextual interpretation in large-scale text data. Such work provides valuable theoretical and empirical perspectives for computational research on culture. Establishing a more systematic connection between the NLP for social science literature and culture-focused NLP research remains an important direction for future work.

Taxonomic Boundaries. The taxonomy proposed in this survey is derived through a bottom-up analysis of the papers we reviewed. It is not

intended to be a complete theory of all possible mechanisms for investigating cultural capabilities in NLP. As the literature develops, additional dimensions and categories may emerge.

Interpretative Constraints. Our analysis is constrained by how culture is defined and described in the surveyed papers. Cultural modeling choices are often only partially formalized or embedded within broader task or dataset design decisions. Consequently, some categorization decisions require interpretation.

Conceptual Proxy Limitations. Many studies represent culture through proxies such as language, geographic region, nationality, or demographic group. These proxies capture different aspects of culture and are not conceptually equivalent. Our survey organizes the current technical design space, but it does not resolve the underlying conceptual challenges of modeling culture.

Ethical Considerations

Culture is complex, dynamic, and internally heterogeneous. However, computational studies often operationalize culture using simplified proxies such as language, nationality, geographic region, or demographic group. These proxies can be practically useful for empirical analysis, but they may also reify culture as fixed or homogeneous, obscure within-group diversity, and inadvertently reinforce stereotypes or exclusions.

The goal of this survey is to organize and analyze the technical mechanisms through which prior work incorporates culture into NLP systems. We do not treat any particular operationalization of culture as definitive or exhaustive. Rather, we view these operationalizations as modeling choices made for specific empirical purposes. We therefore encourage future research to make these assumptions explicit, carefully justify the cultural proxies used, and evaluate the potential downstream risks associated with cultural generalization in NLP systems.

Acknowledgments

We thank the anonymous reviewers and Rajkumar Pujari, for their insightful comments that helped improve the paper. This project was partially funded by NSF IIS-2048001.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12027–12049, Miami, Florida, USA. Association for Computational Linguistics.

Nuredin Ali Abdelkadir, Charles Zhang, Ned Mayo, and Stevie Chancellor. 2024. Diverse perspectives, divergent models: Cross-cultural evaluation of depression detection on Twitter. 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 672–680, Mexico City, Mexico. Association for Computational Linguistics.

Izza AbuHaija, Salim Al Mandhari, Mo El-Haj, Jonas Sibony, and Paul Rayson. 2025. The nakba lexicon: Building a comprehensive dataset from palestinian literature. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 37–47, Abu Dhabi. Association for Computational Linguistics.

Anurag Acharya, Diego Estrada, Shreeja Dahal, W. Victor H. Yarlott, Diana Gomez, and Mark Finlayson. 2024. Discovering implicit meanings of cultural motifs from text. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 46–56, Mexico City, Mexico. Association for Computational Linguistics.

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 Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. *Preprint*, arXiv:2403.15412.

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.

Ibrahim Ahmad, Shiran Dudy, Resmi Ramachandranpillai, and Kenneth Church. 2024. Are generative language models multicultural? a study on Hausa

culture and emotions using ChatGPT. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 98–106, Bangkok, Thailand. Association for Computational Linguistics.

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

Idris Akinade, Jesujoba O. Alabi, David Ifeoluwa Adelani, Clement Odoje, and Dietrich Klakow. 2023. Varepsilon kú mask: Integrating Yorùbá cultural greetings into machine translation. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 1–7, Dubrovnik, Croatia. Association for Computational Linguistics.

Wafa Al Ghallabi, Ritesh Thawkar, Sara Ghaboura, Ketan Pravin More, Omkar Thawakar, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Fann or flop: A multigenre, multiera benchmark for Arabic poetry understanding in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20224–20244, Suzhou, China. Association for Computational Linguistics.

Manal Alhassoun, Imaan Mohammed Alkhanen, Nouf Alshalawi, Ibtehal Baazeem, and Waleed Alsanie. 2025. Saudi-alignment benchmark: Assessing LLMs alignment with cultural norms and domain knowledge in the saudi context. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 130–147, Suzhou, China. Association for Computational Linguistics.

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

Saied Alshahrani, Norah Alshahrani, Soumyabrata Dey, and Jeanna Matthews. 2023. Performance implications of using unrepresentative corpora in Arabic natural language processing. In *Proceedings of ArabicNLP 2023*, pages 218–231, Singapore (Hybrid). Association for Computational Linguistics.

Shatha Altammami. 2025. Leveraging AI to bridge classical Arabic and Modern Standard Arabic for text simplification. In *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, pages 76–85, Abu Dhabi, UAE. Association for Computational Linguistics.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025a. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Fakhraddin Alwajih, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed, and Muhammad Abdul-Mageed. 2025b. PalmX 2025: The first shared task on benchmarking LLMs on Arabic and islamic culture. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 774–789, Suzhou, China. Association for Computational Linguistics.

Fakhraddin Alwajih, Samar M. Magdy, Abdellah El Mekki, Omer Nacar, Youssef Nafea, Safaa Taher Abdelfadil, Abdulfattah Mohammed Yahya, Hamzah Luqman, Nada Almarwani, Samah Aloufi, Baraah Qawasmeh, Houdaifa Atou, Serry Sibaee, Hamzah A. Alsayadi, Walid Al-Dhabyani, Maged S. Al-shaibani, Aya El aatar, Nour Qandos, Rahaf Alhamouri, and 18 others. 2025c. Pearl: A multimodal culturally-aware Arabic instruction dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23048–23079, Suzhou, China. Association for Computational Linguistics.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of Arabic multimodal large language models and benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776, Bangkok, Thailand. Association for Computational Linguistics.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran A. Q. Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged S. Al-shaibani. 2024. CIDAR: Culturally relevant instruction dataset for Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12878–12901, Bangkok, Thailand. Association for Computational Linguistics.

Selenia Anastasi, Florian Schneider, Chris Biemann, and Tim Fischer. 2024. VIDA: The visual incel data archive. a theory-oriented annotated dataset to enhance hate detection through visual culture. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 59–67, Mexico City, Mexico. Association for Computational Linguistics.

Aishwarya Anegundi, Konstantin Schulz, Christian Rauh, and Georg Rehm. 2022. Modelling cultural and socio-economic dimensions of political bias in

German tweets. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 29–40, Potsdam, Germany. KONVENS 2022 Organizers.

Mahfuz Ahmed Anik, Abdur Rahman, Azmine Toushik Wasi, and Md Manjurul Ahsan. 2025. Preserving cultural identity with context-aware translation through multi-agent AI systems. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 51–60, Albuquerque, New Mexico. Association for Computational Linguistics.

Þórunn Arnardóttir, Elías Bjartur Einarsson, Garðar Ingvarsson Juto, Þorvaldur Páll Helgason, and Hafsteinn Einarsson. 2025. WikiQA-IS: Assisted benchmark generation and automated evaluation of Icelandic cultural knowledge in LLMs. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 64–73, Tallinn, Estonia. University of Tartu Library, Estonia.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. 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.

Yasser Ashraf, Yuxia Wang, Bin Gu, Preslav Nakov, and Timothy Baldwin. 2025. Arabic dataset for LLM safeguard 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 5529–5546, Albuquerque, New Mexico. Association for Computational Linguistics.

Soroush Aslani, Jimena Ramirez-Marin, Jeanne Brett, Jingjing Yao, Zhaleh Semnani-Azad, Zhi-Xue Zhang, Catherine Tinsley, Laurie Weingart, and Wendi Adair. 2016. Dignity, face, and honor cultures: A study of negotiation strategy and outcomes in three cultures. *Journal of Organizational Behavior*, 37(8):1178–1201. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/job.2095.

Muhammad Falensi Azmi, Muhammad Dehan Al Kautsar, Alfan Farizki Wicaksono, and Fajri Koto. 2025. IndoSafety: Culturally grounded safety for LLMs in Indonesian languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9135–9166, Suzhou, China. Association for Computational Linguistics.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. 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 2970–2993, Albuquerque, New Mexico. Association for Computational Linguistics.

Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, Shanu Kumar, Sagnik Basu, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. 2025. Navigating the Cultural Kaleidoscope: A Hitchhiker's Guide to Sensitivity in Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7580–7617, Albuquerque, New Mexico. Association for Computational Linguistics.

Yejin Bang, Tiezheng Yu, Andrea Madotto, Zhaojiang Lin, Mona Diab, and Pascale Fung. 2023. Enabling classifiers to make judgements explicitly aligned with human values. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 311–325, Toronto, Canada. Association for Computational Linguistics.

Lisa Bauer, Hanna Tischer, and Mohit Bansal. 2023. Social commonsense for explanation and cultural bias discovery. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3745–3760, Dubrovnik, Croatia. Association for Computational Linguistics.

Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, and Alice Oh. 2025. Diffusion models through a global lens: Are they culturally inclusive? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31137–31155, Vienna, Austria. Association for Computational Linguistics.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025. CULEMO: Cultural lenses on emotion - benchmarking LLMs for cross-cultural emotion understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18894–18909, Vienna, Austria. Association for Computational Linguistics.

Noam K. Benkler, Scott Friedman, Sonja Schmer-Galunder, Drisana Marissa Mosaphir, Robert P. Goldman, Ruta Wheelock, Vasanth Sarathy, Pavan Kantharaju, and Matthew D. McLure. 2024. Recognizing value resonance with resonance-tuned RoBERTa task definition, experimental validation, and robust modeling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13688–13698, Torino, Italia. ELRA and ICCL.

Michael Bennie, Demi Zhang, Bushi Xiao, Jing Cao, Chryseis Xinyi Liu, Jian Meng, and Alayo Tripp. 2025. PANDA - paired anti-hate narratives dataset from Asia: Using an LLM-as-a-judge to create the first Chinese counterspeech dataset. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 1–12, Abu Dhabi, UAE. Association for Computational Linguistics.

Uri Berger and Edoardo Ponti. 2025. Cross-lingual and cross-cultural variation in image descriptions. 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 9453–9465, Albuquerque, New Mexico. Association for Computational Linguistics.

Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2025. Swan and ArabicMTEB: Dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4669–4685, Albuquerque, New Mexico. Association for Computational Linguistics.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, EunJeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782, Miami, Florida, USA. Association for Computational Linguistics.

Shaily Bhatt, Tal August, and Maria Antoniak. 2025. Research borderlands: Analysing writing across research cultures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26238–26266, Vienna, Austria. Association for Computational Linguistics.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16055–16074, Miami, Florida, USA. Association for Computational Linguistics.

Xiaojun Bi, Shuo Li, Junyao Xing, Ziyue Wang, Fuwen Luo, Weizheng Qiao, Lu Han, Ziwei Sun, Peng Li, and Yang Liu. 2025. DongbaMIE: A multimodal information extraction dataset for evaluating semantic understanding of dongba pictograms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 976–990, Suzhou, China. Association for Computational Linguistics.

Houda Bouamor, Sara Al-Emadi, Zeinab Ibrahim, Hany Fazzaa, and Aisha Al-Sultan. 2025. Capturing intra-dialectal variation in qatari Arabic: A corpus of cultural and gender dimensions. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 219–230, Suzhou, China. Association for Computational Linguistics.

Minh Duc Bui, Kyung Eun Park, Goran Glavaš, Fabian David Schmidt, and Katharina Von Der Wense. 2025a. On generalization across measurement systems: LLMs entail more test-time compute for underrepresented cultures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21262–21276, Vienna, Austria. Association for Computational Linguistics.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025b. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–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 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

Antoine Cadotte, Nathalie André, and Fatiha Sadat. 2024. Machine translation through cultural texts: Can verses and prose help low-resource indigenous models? In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 121–127, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2025a. High-dimension human value representation in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5303–5330, Albuquerque, New Mexico. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, Vicky Feliren, Bahrul Ilmi Nasution, Manuel Antonio Rufino, Genta Indra Winata, Rian Adam Rajagede, Carlos Rafael Catalan, and 73 others. 2025b. Crowdsource, crawl, or generate? creating SEA-VL, a multicultural vision-language dataset for Southeast Asia. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18685–18717, Vienna, Austria. Association for Computational Linguistics.

Lorena Calvo-Bartolomé, Valérie Aldana, Karla Cantarero, Alonso Madroñal de Mesa, Jerónimo Arenas-García, and Jordan Lee Boyd-Graber. 2025. Discrepancy detection at the data level: Toward consistent multilingual question answering. In *Proceedings of the 2025 Conference on Empirical Methods in*

Natural Language Processing, pages 22013–22054, Suzhou, China. Association for Computational Linguistics.

Yong Cao, Min Chen, and Daniel Hershcovich. 2024. 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, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. Specializing large language models to simulate survey response distributions for global populations. 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 3141–3154, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Eylon Caplan, Tania Chakraborty, and Dan Goldwasser. 2025. Splits! A Flexible Dataset and Evaluation Framework for Sociocultural Linguistic Investigation. *arXiv preprint*. ArXiv:2504.04640 [cs].

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.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. 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.

Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3548–3564, Singapore. Association for Computational Linguistics.

Kyubyung Chae, Gihoon Kim, Gyuseong Lee, Taesup Kim, Jaejin Lee, and Heejin Kim. 2025. Assessing socio-cultural alignment and technical safety of sovereign LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10579–10600, Suzhou, China. Association for Computational Linguistics.

Nandu Chandran Nair, Rajendran S. Velayuthan, Yamini Chandrashekar, Gábor Bella, and Fausto Giunchiglia. 2022. IndoUKC: A concept-centered Indian multilingual lexical resource. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2833–2840, Marseille, France. European Language Resources Association.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025a. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33019–33036, Suzhou, China. Association for Computational Linguistics.

Danlu Chen, Freda Shi, Aditi Agarwal, Jacobo Myerston, and Taylor Berg-Kirkpatrick. 2024. LogogramNLP: Comparing visual and textual representations of ancient logographic writing systems for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14238–14254, Bangkok, Thailand. Association for Computational Linguistics.

Jiale Chen, Xuelian Dong, Qihao Yang, Wenxiu Xie, and Tianyong Hao. 2025b. Can large language models translate spoken-only languages through international phonetic transcription? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23420–23435, Suzhou, China. Association for Computational Linguistics.

Kai Chen, Zihao He, Taiwei Shi, and Kristina Lerman. 2025c. STEER-BENCH: A benchmark for evaluating the steerability of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18327–18355, Suzhou, China. Association for Computational Linguistics.

Xinyu Chen, Yunxin Li, Haoyuan Shi, Baotian Hu, Wenhan Luo, Yaowei Wang, and Min Zhang. 2025d. VideoVista-CulturalLingo: 360° horizons-bridging cultures, languages, and domains in video comprehension. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27102–27128, Vienna, Austria. Association for Computational Linguistics.

Tsz Chung Cheng, Chung Shing Cheng, Chaak-ming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. 2025. HKCanto-eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 1–11, Vienna, Austria. Association for Computational Linguistics.

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

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15042–15058, Bangkok, Thailand. Association for Computational Linguistics.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Simone Conia, Min Li, Roberto Navigli, and Saloni Potdar. 2025. SemEval-2025 task 2: Entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2535–2557, Vienna, Austria. Association for Computational Linguistics.

Alejandro Cuevas, Saloni Dash, Bharat Kumar Nayak, Dan Vann, and Madeleine I. G. Daepp. 2025. Anecdoctoring: Automated red-teaming across language and place. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19055–19074, Suzhou, China. Association for Computational Linguistics.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From word to world: Evaluate and mitigate culture bias in LLMs via word association test. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24510–24526, Suzhou, China. Association for Computational Linguistics.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. 2024. "They are uncultured": Unveiling Covert Harms and Social Threats in LLM Generated Conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20339–20369, Miami, Florida, USA. Association for Computational Linguistics.

Amitava Das, Yaswanth Narsupalli, Gurpreet Singh, Vinija Jain, Vasu Sharma, Suranjana Trivedy, Aman Chadha, and Amit Sheth. 2025. YinYang-align: A new benchmark for competing objectives and introducing multi-objective preference based text-to-image alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23518–23598, Vienna, Austria. Association for Computational Linguistics.

Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: The case of Bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83, Dubrovnik, Croatia. Association for Computational Linguistics.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Nicholas Deas, Elsbeth Turcan, Ivan Ernesto Perez Mejia, and Kathleen McKeown. 2024. MASIVE: Open-ended affective state identification in English and Spanish. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20467–20485, Miami, Florida, USA. Association for Computational Linguistics.

Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. StereoKG: Data-driven knowledge graph construction for cultural knowledge and stereotypes. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 67–78, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Priyanka Dey, Aayush Bothra, Yugal Khanter, Jieyu Zhao, and Emilio Ferrara. 2025. Can LLMs express personality across cultures? introducing CulturalPersonas for evaluating trait alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20241–20262, Suzhou, China. Association for Computational Linguistics.

Ashutosh Dwivedi, Siddhant Shivdutt Singh, and Ashutosh Modi. 2025. EtiCor++: Towards understanding etiquettical bias in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9355–9376, Vienna, Austria. Association for Computational Linguistics.

Abdellah El Mekki, Houdaifa Atou, Omer Nacar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. NileChat: Towards linguistically diverse and culturally aware LLMs for local communities. In *Proceedings of the 2025 Conference on Empirical*

Methods in Natural Language Processing, pages 10967–10991, Suzhou, China. Association for Computational Linguistics.

Sugyeong Eo, Jungwoo Lim, Chanjun Park, DaHyun Jung, Seonmin Koo, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim. 2024. Detecting critical errors considering cross-cultural factors in English-Korean translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4705–4716, Torino, Italia. ELRA and ICCL.

Elena V. Epure, Guillaume Salha, Manuel Moussallam, and Romain Hennequin. 2020. Modeling the music genre perception across language-bound cultures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4765–4779, Online. Association for Computational Linguistics.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2024. Elote, choclo and mazorca: on the varieties of Spanish. 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 3689–3711, Mexico City, Mexico. Association for Computational Linguistics.

Cristina España-Bonet, Ankur Bhatt, Koel Dutta Chowdhury, and Alberto Barrón-Cedeño. 2024. When elote, choclo and mazorca are not the same. isomorphism-based perspective to the Spanish varieties divergences. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 56–77, Mexico City, Mexico. Association for Computational Linguistics.

Neele Falk, Andreas Waldis, and Iryna Gurevych. 2024. Overview of PerspectiveArg2024 the first shared task on perspective argument retrieval. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 130–149, Bangkok, Thailand. Association for Computational Linguistics.

Jizhan Fang, Tianhe Lu, Yunzhi Yao, Ziyan Jiang, Xin Xu, Huajun Chen, and Ningyu Zhang. 2025. CKnowEdit: A new Chinese knowledge editing dataset for linguistics, facts, and logic error correction in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8789–8807, Vienna, Austria. Association for Computational Linguistics.

Mohammad Rifqi Farhansyah, Iwan Darmawan, Adryan Kusumawardhana, Genta Indra Winata, Alham Fikri Aji, and Derry Tanti Wijaya. 2025. Do language models understand honorific systems in Javanese? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26732–26754, Vienna, Austria. Association for Computational Linguistics.

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

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Orevaoghene Ahia, Shuyue Stella Li, Vidhisha Balachandran, Sunayana Sitaram, and Yulia Tsvetkov. 2024a. Teaching LLMs to abstain across languages via multilingual feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4125–4150, Miami, Florida, USA. Association for Computational Linguistics.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024b. 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.

Karen Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S. Schmeisser-Nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, and 4 others. 2024. Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17764–17769, Torino, Italia. ELRA and ICCL.

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019. Relating word embedding gender biases to gender gaps: A cross-cultural analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 18–24, Florence, Italy. Association for Computational Linguistics.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. CHENGYU-BENCH: Benchmarking large language models for Chinese idiom understanding and use. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2366, Suzhou, China. Association for Computational Linguistics.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.

Sara Ghaboura, Ketan Pravin More, Ritesh Thawkar, Wafa Al Ghallabi, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Time travel: A comprehensive benchmark to evaluate LMMs on historical and cultural artifacts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23627–23641, Vienna, Austria. Association for Computational Linguistics.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.

Nevan Giuliani, Cheng Charles Ma, Prakruthi Pradeep, and Daphne Ippolito. 2024. CAVA: A tool for cultural alignment visualization & analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 153–161, Miami, Florida, USA. Association for Computational Linguistics.

María Grandury, Javier Aula-Blasco, Júlia Falcão, Clémentine Fourrier, Miguel González Saiz, Gonzalo Martínez, Gonzalo Santamaria Gomez, Rodrigo Agerri, Nuria Aldama García, Luis Chiruzzo, Javier Conde, Helena Gomez Adorno, Marta Guerrero Nieto, Guido Ivetta, Natàlia López Fuertes, Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, Helena Montoro Zamorano, Carmen Muñoz Sanz, and 6 others. 2025. La leaderboard: A large language model leaderboard for Spanish varieties and languages of Spain and Latin America. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32482–32524, Vienna, Austria. Association for Computational Linguistics.

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.

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

Abhay Gupta, Jacob Cheung, Philip Meng, Shayan Sayyed, Kevin Zhu, Austen Liao, and Sean O'Brien. 2025. EnDive: A cross-dialect benchmark for fairness and performance in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16830–16855, Suzhou, China. Association for Computational Linguistics.

E.D. Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60.

Christopher R. Haberland, Jean Maillard, and Stefano Lusito. 2024. Italian-Ligurian machine translation in its cultural context. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages @ LREC-COLING 2024*, pages 168–176, Torino, Italia. ELRA and ICCL.

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. *(No Title)*.

Younggyun Hahm, Youngbin Noh, Ji Yoon Han, Tae Hwan Oh, Hyonsu Choe, Hansaem Kim, and Key-Sun Choi. 2020. Crowdsourcing in the development of a multilingual FrameNet: A case study of Korean FrameNet. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 236–244, Marseille, France. European Language Resources Association.

James Anthony Hale, Sushrita Rakshit, Kushal Chawla, Jeanne M Brett, and Jonathan Gratch. 2025. KODIS: A multicultural dispute resolution dialogue corpus. 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 12771–12785, Albuquerque, New Mexico. Association for Computational Linguistics.

Edward T. Hall. 1976. *Beyond Culture*. Anchor Press/Doubleday, New York.

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.

Abdullah Hashmat, Muhammad Arham Mirza, and Agha Ali Raza. 2025. PakBBQ: A culturally adapted bias benchmark for QA. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16160–16172, Suzhou, China. Association for Computational Linguistics.

Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. 2024. Building knowledge-guided lexica to model cultural variation. 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 211–226, Mexico City, Mexico. Association for Computational Linguistics.

Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. 2023a. Comparing styles across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6775–6791, Singapore. Association for Computational Linguistics.

Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023b. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. 2025. Towards style alignment in cross-cultural translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32213–32230, Vienna, Austria. Association for Computational Linguistics.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, and Naihao Deng. 2025. Chumor 2.0: Towards better benchmarking Chinese humor understanding from (ruo zhi ba). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21799–21818, Vienna, Austria. Association for Computational Linguistics.

Megan Herrera, Ankit Aich, and Natalie Parde. 2022. TweetTaglish: A dataset for investigating Tagalog-English code-switching. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2090–2097, Marseille, France. European Language Resources Association.

David G Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. Story morals: Surfacing value-driven narrative schemas using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13032, Miami, Florida, USA. Association for Computational Linguistics.

Geert Hofstede. 1984. *Culture's consequences: International differences in work-related values*, volume 5. sage.

Minki Hong, Jangho Choi, and Jihie Kim. 2025. NormGenesis: Multicultural dialogue generation via exemplar-guided social norm modeling and violation recovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33793–33831, Suzhou, China. Association for Computational Linguistics.

Sara Bourbour Hosseinbeigi, Behnam Rohani, Mostafa Masoudi, Mehrnoush Shamsfard, Zahra Saaberi, Mostafa Karimi Manesh, and Mohammad Amin Abbasi. 2025a. Advancing Persian LLM evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2711–2727, Albuquerque, New Mexico. Association for Computational Linguistics.

Sara Bourbour Hosseinbeigi, MohammadAli SeifKashani, Javad Seraj, Fatemeh Taherinezhad, Ali Nafisi, Fatemeh Nadi, Iman Barati, Hosein Hasani, Mostafa Amiri, and Mostafa Masoudi. 2025b. Matina: A culturally-aligned Persian language model using multiple LoRA experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20874–20889, Vienna, Austria. Association for Computational Linguistics.

Hsin-Yi Hsieh, Shih-Cheng Huang, and Richard Tzong-Han Tsai. 2024. TWBias: A benchmark for assessing social bias in traditional Chinese large language models through a Taiwan cultural lens. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8688–8704, Miami, Florida, USA. Association for Computational Linguistics.

Hsin-Yi Hsieh, Shang Wei Liu, Chang Chih Meng, Shuo-Yueh Lin, Chen Chien-Hua, Hung-Ju Lin, Hen-Hsen Huang, and I-Chen Wu. 2025. TaiwanVQA: A benchmark for visual question answering for Taiwanese daily life. In *Proceedings of the First Workshop of Evaluation of Multi-Modal Generation*, pages 57–75, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Yujia Hu, Ming Shan Hee, Preslav Nakov, and Roy Ka-Wei Lee. 2025. Toxicity Red-Teaming: Benchmarking LLM Safety in Singapore's Low-Resource Languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12183–12201, Suzhou, China. Association for Computational Linguistics.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. 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 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. 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.

Oana Ignat, Gayathri Ganesh Lakshmy, and Rada Mihalcea. 2025. InspAIred: Cross-cultural inspiration detection and analysis in real and LLM-generated social media data. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 35–49, Albuquerque, New Mexico. Association for Computational Linguistics.

Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. 2000. World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*.

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Aizirek Turdubaeva, Ilshat Saetov, Rinat Kharisov, Saule Belginova, Ariana Kenbayeva, Amina Alisheva, Abdullatif Köksal, Samir Rustamov, and Duygu Ataman. 2025. TUMLU: A unified and native language understanding benchmark for Turkic languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22816–22838, Vienna, Austria. Association for Computational Linguistics.

Ishita and Radhika Mamidi. 2025. The evolution of gen alpha slang: Linguistic patterns and AI translation challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 678–686, Vienna, Austria. Association for Computational Linguistics.

Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Gandhi, Ambujavalli R, Manickam M, C Vaijayanthi, Krishnan Karunganni, and 2 others. 2024. IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10740–10782, Bangkok, Thailand. Association for Computational Linguistics.

Suchae Jeong, Inseong Choi, Youngsik Yun, and Jihie Kim. 2025. Culture-TRIP: Culturally-aware text-to-image generation with iterative prompt refinement. 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 9543–9573, Albuquerque, New Mexico. Association for Computational Linguistics.

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 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.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794, Vienna, Austria. Association for Computational Linguistics.

Yuu Jinnai. 2024. Does cross-cultural alignment change the commonsense morality of language models? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 48–64, Bangkok, Thailand. Association for Computational Linguistics.

Raviraj Bhuminand Joshi, Rakesh Paul, Kanishk Singla, Anusha Kamath, Michael Evans, Katherine Luna, Shaona Ghosh, Utkarsh Vaidya, Eileen Margaret Peters Long, Sanjay Singh Chauhan, and Niranjan Wartikar. 2025. CultureGuard: Towards culturally-aware dataset and guard model for multilingual safety applications. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2666–2685, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

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

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.

Antonia Karamolegkou, Malvina Nikandrou, Georgios Pantazopoulos, Danae Sanchez Villegas, Phillip Rust, Ruchira Dhar, Daniel Hershcovich, and Anders Søgaard. 2025. Evaluating multimodal language models as visual assistants for visually impaired users. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25949–25982, Vienna, Austria. Association for Computational Linguistics.

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.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Simran Khanuja, Vivek Iyer, Xiaoyu He, and Graham Neubig. 2025. Towards automatic evaluation for image transcreation. 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 7034–7047, Albuquerque, New Mexico. Association for Computational Linguistics.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics.

Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. 2025. Multiple LLM agents debate for equitable cultural alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24841–24877, Vienna, Austria. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Dahyun Kim, Sukyung Lee, Yungi Kim, Attapol Rutherford, and Chanjun Park. 2025a. Representing the under-represented: Cultural and core capability benchmarks for developing Thai large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4114–4129, Abu Dhabi, UAE. Association for Computational Linguistics.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024a. 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.

Gyeongmin Kim, Jinsung Kim, Junyoung Son, and Heuiseok Lim. 2022. KoCHET: A Korean cultural heritage corpus for entity-related tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3496–3505, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jaehong Kim, Chaeyoon Jeong, Seongchan Park, Meeyoung Cha, and Wonjae Lee. 2024b. How do moral emotions shape political participation? a cross-cultural analysis of online petitions using language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16274–16289, Bangkok, Thailand. Association for Computational Linguistics.

Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. 2025b. WHEN TOM EATS KIMCHI: Evaluating cultural awareness of multimodal large language models in cultural mixture contexts. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 143–154, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Michelle YoungJin Kim and Kristen Johnson. 2025. Korean stereotype content model: Translating stereotypes across cultures. In *Proceedings of the 3rd*

Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025), pages 59–70, Albuquerque, New Mexico. Association for Computational Linguistics.

Sean Kim and Hyuhng Joon Kim. 2025. A dual-layered evaluation of geopolitical and cultural bias in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 580–595, Vienna, Austria. Association for Computational Linguistics.

Artur Kiulian, Anton Polishko, Mykola Khandoga, Oryna Chubych, Jack Connor, Raghav Ravishankar, and Adarsh Shirawalmath. 2024. From bytes to borsch: Fine-tuning gemma and mistral for the Ukrainian language representation. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 83–94, Torino, Italia. ELRA and ICCL.

Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. Untangling hate speech definitions: A semantic componential analysis across cultures and domains. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3184–3198, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Stefan Krsteski, Borjan Sazdov, Matea Tashkovska, Branislav Gerazov, and Hristijan Gjoreski. 2025. Towards open foundation language model and corpus for Macedonian: A low-resource language. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 44–57, Vienna, Austria. Association for Computational Linguistics.

Shivani Kumar and David Jurgens. 2025. Are rules meant to be broken? understanding multilingual moral reasoning as a computational pipeline with UniMoral. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5890–5912, Vienna, Austria. Association for Computational Linguistics.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10383–10405, Singapore. Association for Computational Linguistics.

Tian Lan, Xiangdong Su, Xu Liu, Ruirui Wang, Ke Chang, Jiang Li, and Guanglai Gao. 2025. McBE: A multi-task Chinese bias evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6033–6056, Vienna, Austria. Association for Computational Linguistics.

Anton Lavrouk, Tarek Naous, Alan Ritter, and Wei Xu. 2025. What are foundation models cooking in the post-soviet world? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20687–20709, Suzhou, China. 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.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024b. 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.

Thibaud Leteno, Irina Proskurina, Antoine Gourru, Julien Velcin, Charlotte Laclau, Guillaume Metzler, and Christophe Gravier. 2025. HISTOIRES-MORALES: A French dataset for assessing moral alignment. 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 2590–2612, Albuquerque, New Mexico. Association for Computational Linguistics.

Bryan Li, Fiona Luo, Samar Haider, Adwait Agashe, Siyu Li, Runqi Liu, Miranda Muqing Miao, Shriya Ramakrishnan, Yuan Yuan, and Chris Callison-Burch. 2025. Multilingual retrieval augmented generation for culturally-sensitive tasks: A benchmark for cross-lingual robustness. In *Findings of the Association*

for Computational Linguistics: ACL 2025, pages 4215–4241, Vienna, Austria. Association for Computational Linguistics.

Bryan Li, Aleksey Panasyuk, and Chris Callison-Burch. 2024a. Uncovering differences in persuasive language in Russian versus English Wikipedia. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 21–35, Miami, Florida, USA. Association for Computational Linguistics.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: boosting cross-cultural understanding in large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024c. CulturePark: Boosting Cross-cultural Understanding in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pages 65183–65216. Curran Associates, Inc.

Sha Li, Revanth Gangi Reddy, Khanh Duy Nguyen, Qingyun Wang, Yi Fung, Chi Han, Jiawei Han, Kartik Natarajan, Clare R. Voss, and Heng Ji. 2024d. Schema-guided culture-aware complex event simulation with multi-agent role-play. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 372–381, Miami, Florida, USA. Association for Computational Linguistics.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024e. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA. Association for Computational Linguistics.

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. 2024f. 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.

Zhi Li and Yin Zhang. 2023. Cultural concept adaptation on multimodal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 262–276, Singapore. Association for Computational Linguistics.

Jinggui Liang, Dung Vo, Yap Hong Xian, Hai Leong Chieu, Kian Ming A. Chai, Jing Jiang, and Lizi Liao. 2025. Colloquial singaporean English style transfer with fine-grained explainable control. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26962–26983, Vienna, Austria. Association for Computational Linguistics.

Xixian Liao, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, Javier García Gilabert, Miguel Claramunt Argote, Ella Bohman, and Maite Melero. 2025. Culture-aware machine translation: the case study of low-resource language pair Catalan-Chinese. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 150–161, Geneva, Switzerland. European Association for Machine Translation.

Peerat Limkonchotiwat, Pume Tuchinda, Lalita Lowphansirikul, Surapon Nonesung, Panuthep Tasawong, Alham Fikri Aji, Can Udomcharoenchaikit, and Sarana Nutanong. 2025. WangchanThaiInstruct: An instruction-following dataset for culture-aware, multitask, and multi-domain evaluation in Thai. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3535–3558, Suzhou, China. Association for Computational Linguistics.

Bill Yuchen Lin, Frank F. Xu, Kenny Zhu, and Seungwon Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, Melbourne, Australia. Association for Computational Linguistics.

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

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

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haijiang Liu, Qiyuan Li, Chao Gao, Yong Cao, Xiangyu Xu, Xun Wu, Daniel Hershcovich, and Jinguang Gu. 2025c. Beyond demographics: Enhancing cultural value survey simulation with multi-stage personality-driven cognitive reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural*

Language Processing, pages 18406–18428, Suzhou, China. Association for Computational Linguistics.

Jiarui Liu, Yueqi Song, Yunze Xiao, Mingqian Zheng, Lindia Tjuatja, Jana Schaich Borg, Mona T. Diab, and Maarten Sap. 2025d. Synthetic socratic debates: Examining persona effects on moral decision and persuasion dynamics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16428–16458, Suzhou, China. Association for Computational Linguistics.

Xuelin Liu, Yanfei Zhu, Shucheng Zhu, Pengyuan Liu, Ying Liu, and Dong Yu. 2024a. Evaluating moral beliefs across LLMs through a pluralistic framework. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4740–4760, Miami, Florida, USA. Association for Computational Linguistics.

Yang Liu, Jiahuan Cao, Hiuyi Cheng, Yongxin Shi, Kai Ding, and Lianwen Jin. 2025e. MCS-bench: A comprehensive benchmark for evaluating multimodal large language models in Chinese classical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10435–10492, Vienna, Austria. Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, and Nancy Chen. 2024b. Optimizing code-switching in conversational tutoring systems: A pedagogical framework and evaluation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 500–515, Kyoto, Japan. Association for Computational Linguistics.

Andrés Lou, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, and Víctor Sánchez-Cartagena. 2024. Curated datasets and neural models for machine translation of informal registers between Mayan and Spanish vernaculars. 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 2838–2850, Mexico City, Mexico. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Weicheng Ma, John J. Guerrerio, and Soroush Vosoughi. 2025a. Scalable and culturally specific stereotype dataset construction via human-LLM collaboration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23928–23956, Suzhou, China. Association for Computational Linguistics.

Weicheng Ma, Hefan Zhang, Shiyu Ji, Farnoosh Hashemi, Qichao Wang, Ivory Yang, Joice Chen, Juanwen Pan, Michael Macy, Saeed Hassanpour, and Soroush Vosoughi. 2025b. Enhancing LLM-based persuasion simulations with cultural and speaker-specific information. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14955–14976, Suzhou, China. Association for Computational Linguistics.

Sangmitra Madhusudan, Robert Morabito, Skye Reid, Nikta Gohari Sadr, and Ali Emami. 2025. Fine-Tuned LLMs are "Time Capsules" for Tracking Societal Bias Through Books. 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 2329–2358, Albuquerque, New Mexico. Association for Computational Linguistics.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. JAWAHER: A multidialectal dataset of Arabic proverbs for LLM benchmarking. 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 12320–12341, Albuquerque, New Mexico. Association for Computational Linguistics.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka, and Sriparna Saha. 2025a. SANSKRITI: A comprehensive benchmark for evaluating language models' knowledge of Indian culture. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4434–4451, Vienna, Austria. Association for Computational Linguistics.

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

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, and Sriparna Saha. 2025c. DRISHTIKON: A Multimodal Multilingual Benchmark for Testing Language Models' Understanding on Indian Culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1289–1313, Suzhou, China. Association for Computational Linguistics.

Ananya Malik, Nazanin Sabri, Melissa M. Karnaze, and Mai ElSherief. 2025. Are LLMs Empathetic to All? Investigating the Influence of Multi-Demographic Personas on a Model's Empathy. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24938–24959, Suzhou, China. Association for Computational Linguistics.

Antonis Maronikolakis, Abdullatif Köksal, and Hinrich Schuetze. 2024. Sociocultural knowledge is needed for selection of shots in hate speech detection tasks. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 1–13, St. Julian's, Malta. Association for Computational Linguistics.

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

Orfeas Menis Mastromichalakis, Jason Liartis, Kristina Rose, Antoine Isaac, and Giorgos Stamou. 2025. Don't Erase, Inform! Detecting and Contextualizing Harmful Language in Cultural Heritage Collections. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21836–21850, Vienna, Austria. Association for Computational Linguistics.

Md Ayon Mia, Akm Moshiur Rahman Mazumder, Khadiza Sultana Sayma, Md Fahim, Md Tahmid Hasan Fuad, Muhammad Ibrahim Khan, and Akmmahbubur Rahman. 2025. BANMIME : Misogyny detection with metaphor explanation on Bangla memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17813–17839, Suzhou, China. Association for Computational Linguistics.

Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M. Daly, Kush R. Varshney, Eitan Farchi, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, Miao Liu, and Prasanna Sattigeri. 2025. Evaluating the prompt steerability 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 7874–7900, Albuquerque, New Mexico. Association for Computational Linguistics.

Jeremiah Milbauer, Adarsh Mathew, and James Evans. 2021. Aligning multidimensional worldviews and discovering ideological differences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4832–4845, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, and 35 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Luca Mitran, Sophie Wu, and Andrew Piper. 2025. Probing narrative morals: A new character-focused MFT framework for use with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28514–28529, Suzhou, China. Association for Computational Linguistics.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8770–8785, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Church, and Mohamed Elhoseiny. 2024. No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20939–20962, Miami, Florida, USA. Association for Computational Linguistics.

Hadi Mohammadi, Yasmeen F. S. S. Meijer, Efthymia Papadopoulou, and Ayoub Bagheri. 2025. Do large language models understand morality across cultures? In *Proceedings of the 2nd LUHME Workshop*, pages 30–39, Bologna, Italy. UP - Universidade

do Porto (https://doi.org/10.21747/978-989-9193-73-4/lan2), LIACC - Laboratório de Inteligência Artificial e Ciência de Computadores da Universidade do Porto, CLUP - Centro de Linguística da Universidade do Porto, UEF - The University of Eastern Finland and UAH - Universidad de Alcalá.

Jann Railey Montalan, Jimson Paulo Layacan, David Demitri Africa, Richell Isaiah S. Flores, Michael T. Lopez Ii, Theresa Denise Magsajo, Anjanette Cayabyab, and William Chandra Tjhi. 2025. Batayan: A Filipino NLP benchmark for evaluating large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31239–31273, Vienna, Austria. Association for Computational Linguistics.

Erfan Moosavi Monazzah, Vahid Rahimzadeh, Yadollah Yaghoobzadeh, Azadeh Shakery, and Mohammad Taher Pilehvar. 2025. PerCul: A story-driven cultural evaluation of LLMs in Persian. 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 12670–12687, Albuquerque, New Mexico. Association for Computational Linguistics.

Guy Mor-Lan, Naama Rivlin-Angert, Yael R. Kaplan, Tamir Sheafer, and Shaul R. Shenhav. 2025. HebID: Detecting social identities in Hebrew-language political text. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9850–9870, Suzhou, China. Association for Computational Linguistics.

David R. Mortensen, Xinyu Zhang, Chenxuan Cui, and Katherine J. Zhang. 2022. A Hmong corpus with elaborate expression annotations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4992–5000, Marseille, France. European Language Resources Association.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

Hamdy Mubarak, Abubakr Mohamed, and Majd Hawasly. 2025. AraSafe: Benchmarking safety in Arabic LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9976–9992, Suzhou, China. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, Ebrahim Chekol Jibril, Elyas Abdi Ismail, Esubalew Alemneh, Hagos Tesfahun Gebremichael, Lukman Jibril Aliyu, Meriem Beloucif, Oumaima Hourrane, Rooweither Mabuya, Salomey Osei, and 8 others. 2025. AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages. 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 1854–1871, Albuquerque, New Mexico. Association for Computational Linguistics.

Anjishnu Mukherjee, Aylin Caliskan, Ziwei Zhu, and Antonios Anastasopoulos. 2024. Global gallery: The fine art of painting culture portraits through multilingual instruction tuning. 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 6398–6415, Mexico City, Mexico. Association for Computational Linguistics.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global Voices, local biases: Socio-cultural prejudices across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15828–15845, Singapore. Association for Computational Linguistics.

Sourabrata Mukherjee, Atharva Mehta, Sougata Saha, Akhil Arora, and Monojit Choudhury. 2025. Women, infamous, and exotic beings: A comparative study of honorific usages in Wikipedia and LLMs for Bengali and Hindi. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19103–19126, Suzhou, China. Association for Computational Linguistics.

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.

Shahriar Kabir Nahin, Rabindra Nath Nandi, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Kowsher, Apu Chandraw Shill, Md Ibrahim, Mehadi Hasan Menon, Tareq Al Muntasir, and Firoj Alam. 2025. TituLLMs: A family of Bangla LLMs with comprehensive benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24922–24940, Vienna, Austria. Association for Computational Linguistics.

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

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

Janki Atul Nawale, Mohammed Safi Ur Rahman Khan, Janani D, Mansi Gupta, Danish Pruthi, and Mitesh M Khapra. 2025. FairI tales: Evaluation of fairness in Indian contexts with a focus on bias and stereotypes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30331–30380, Vienna, Austria. Association for Computational Linguistics.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd Van Steenkiste, Yash Goyal, Karolina Stanczak, and Aishwarya Agrawal. 2025. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20918–20953, Suzhou, China. Association for Computational Linguistics.

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

Ri Chi Ng, Nirmalendu Prakash, Ming Shan Hee, Kenny Tsu Wei Choo, and Roy Ka-wei Lee. 2024. SGHateCheck: Functional tests for detecting hate speech in low-resource languages of Singapore. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 312–327, Mexico City, Mexico. Association for Computational Linguistics.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024a. Cultural commonsense knowledge for intercultural dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 1774–1784, New York, NY, USA. Association for Computing Machinery.

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. 2024b. 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.

Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. 2025. CROPE: Evaluating in-context adaptation of vision and language models to culture-specific concepts. 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 7917–7936, Albuquerque, New Mexico. Association for Computational Linguistics.

Charles Nimo, Shuheng Liu, Irfan Essa, and Michael L. Best. 2025. Africa health check: Probing cultural bias in medical LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32219–32232, Suzhou, China. Association for Computational Linguistics.

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

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

Kayode Olaleye, Arturo Oncevay, Mathieu Sibue, Nombuyiselo Zondi, Michelle Terblanche, Sibongile Mapikitla, Richard Lastrucci, Charese Smiley, and Vukosi Marivate. 2025. AfroCS-xs: Creating a compact, high-quality, human-validated code-switched dataset for African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33391–33410, Vienna, Austria. 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.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016. PROMETHEUS: A corpus of proverbs annotated with metaphors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3787–3793, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Saurabh Kumar Pandey, Harshit Budhiraja, Sougata Saha, and Monojit Choudhury. 2025. CULTURALLY YOURS: A reading assistant for cross-cultural content. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 208–216, Abu Dhabi, UAE. Association for Computational Linguistics.

Yurii Paniv, Artur Kiulian, Dmytro Chaplynskyi, Mykola Khandoga, Anton Polishko, Tetiana Bas, and Guillermo Gabrielli. 2025. Benchmarking multimodal models for Ukrainian language understanding across academic and cultural domains. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 14–26, Vienna, Austria (online). Association for Computational Linguistics.

Mohammed Fayiz Parappan and Ricardo Henao. 2025. Learning subjective label distributions via sociocultural descriptors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20322–20338, Suzhou, China. Association for Computational Linguistics.

ChaeHun Park, Yujin Baek, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2025a. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-VLM collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21960–21974, Vienna, Austria. Association for Computational Linguistics.

Junyeong Park, Seogyeong Jeong, Seyoung Song, Yohan Lee, and Alice Oh. 2025b. LLM-C3MOD: A human-LLM collaborative system for cross-cultural hate speech moderation. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 71–88, Albuquerque, New Mexico. Association for Computational Linguistics.

Seoyoon Park, Jaehee Kim, and Hansaem Kim. 2025c. Too polite to be human: Evaluating LLM empathy in Korean conversations via a DCT-based framework. In *Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025)*, pages 76–89, Vienna, Austria. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

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

Viet Thanh Pham, Shilin Qu, Farhad Moghimifar, Suraj Sharma, Yuan-Fang Li, Weiqing Wang, and Reza Haf. 2024. Multi-cultural norm base: Frame-based norm discovery in multi-cultural settings. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 24–35, Miami, FL, USA. Association for Computational Linguistics.

Prisca Piccirilli, Alexander Fraser, and Sabine Schulte im Walde. 2024. VOLIMET: A Parallel Corpus of Literal and Metaphorical Verb-Object Pairs for English–German and English–French. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 222–237, Mexico City, Mexico. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Susanna Paoli, Alba Cercas Curry, and Dirk Hovy. 2024. Divine LLaMAs: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4346–4366, Miami, Florida, USA. Association for Computational Linguistics.

Rhitabrat Pokharel and Ameeta Agrawal. 2025. neDIOM: Dataset and analysis of Nepali idioms. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 160–171, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Diaz, Ding Wang, and Gregory Serapio-García. 2024. GRASP: A disagreement analysis framework to assess group associations in perspectives. 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 3473–3492, Mexico City, Mexico. Association for Computational Linguistics.

Ashmari Pramodya, Nirasha Nelki, Heshan Shalinda, Chamila Liyanage, Yusuke Sakai, Randil Pushpananda, Ruvan Weerasinghe, Hidetaka Kamigaito, and Taro Watanabe. 2025. SinhalaMMLU: A comprehensive benchmark for evaluating multitask language understanding in Sinhala. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32943–32961, Suzhou, China. Association for Computational Linguistics.

Juan Prieto, Cristian Martinez, Melissa Robles, Alberto Moreno, Sara Palacios, and Rubén Manrique. 2024. Translation systems for low-resource colombian indigenous languages, a first step towards cultural

preservation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 7–14, Mexico City, Mexico. Association for Computational Linguistics.

Rajkumar Pujari and Dan Goldwasser. 2025. LLM-human pipeline for cultural grounding of conversations. 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 1029–1048, Albuquerque, New Mexico. 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.

Karthick Narayanan R, Siddharth Singh, Saurabh Singh, Aryan Mathur, Ritesh Kumar, Shyam Ratan, Bornini Lahiri, Benu Pareek, Neerav Mathur, Amalesh Gope, Meiraba Takhellambam, and Yogesh Dawer. 2025. Field to model: Pairing community data collection with scalable NLP through the LiFE suite. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 76–84, Vienna, Austria. Association for Computational Linguistics.

Neel Prabhanjan Rachamalla, Aravind Konakalla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. 2025. Pragyaan: Designing and curating high-quality cultural post-training datasets for Indian languages. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 285–321, Suzhuo, China. Association for Computational Linguistics.

Sunny Rai, Khushi Shelat, Devansh Jain, Ashwin Kishen, Young Min Cho, Maitreyi Redkar, Samindara Hardikar-Sawant, Lyle Ungar, and Sharath Chandra Guntuku. 2025a. Cross-cultural differences in mental health expressions on social media. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 132–142, Albuquerque, New Mexico. Association for Computational Linguistics.

Sunny Rai, Khushang Zaveri, Shreya Havaldar, Soumna Nema, Lyle Ungar, and Sharath Chandra Guntuku. 2025b. Social norms in cinema: A cross-cultural analysis of shame, pride and prejudice. 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 11396–11415, Albuquerque, New Mexico. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2025. The discordance between embedded ethics and cultural inference in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14715–14736, Suzhou, China. Association for Computational Linguistics.

Artem Reshetnikov and Maria-Cristina Marinescu. 2025. Caption generation in cultural heritage: Crowdsourced data and tuning multimodal large language models. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 42–50, Albuquerque, New Mexico. Association for Computational Linguistics.

Dor Ringel, Gal Lavee, Ido Guy, and Kira Radinsky. 2019. Cross-cultural transfer learning for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3873–3883, Hong Kong, China. Association for Computational Linguistics.

Keonwoo Roh, Yeong-Joon Ju, and Seong-Whan Lee. 2025. XLQA: A benchmark for locale-aware multilingual open-domain question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28809–28821, Suzhou, China. Association for Computational Linguistics.

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 67–74, Torino, Italia. ELRA and ICCL.

Donya Rooein, Vilém Zouhar, Debora Nozza, and Dirk Hovy. 2025. Biased Tales: Cultural and Topic Bias in Generating Children's Stories. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 52–72, Suzhou, China. Association for Computational Linguistics.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in Arab culture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 7695–7710, Vienna, Austria. Association for Computational Linguistics.

Nikta Gohari Sadr, Sahar Heidariasl, Karine Megerdoomian, Laleh Seyyed-Kalantari, and Ali Emami. 2025. We politely insist: Your LLM must learn the Persian art of taarof. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1819–1838, Suzhou, China. Association for Computational Linguistics.

Hamidreza Saffari, Mohammadamin Shafiei, Donya Rooein, Francesco Pierri, and Debora Nozza. 2025. Can I introduce my boyfriend to my grandmother? evaluating large language models capabilities on Iranian social norm classification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6075–6089, Albuquerque, New Mexico. Association for Computational Linguistics.

Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta, and Monojit Choudhury. 2025. Reading between the lines: Can LLMs identify cross-cultural communication gaps? 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 8043–8067, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Pramit Sahoo, Maharaj Brahma, and Maunendra Sankar Desarkar. 2025. DIWALI - diversity and inclusivity aWare cuLture specific items for India: Dataset and assessment of LLMs for cultural text adaptation in Indian context. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33599–33626, Suzhou, China. Association for Computational Linguistics.

Edward Sapir. 1929. The status of linguistics as a science. *Language*, 5(4):207–214.

David Sasu, Zehui Wu, Ziwei Gong, Run Chen, Pengyuan Shi, Lin Ai, Julia Hirschberg, and Natalie Schluter. 2025. Akan cinematic emotions (ACE): A multimodal multi-party dataset for emotion recognition in movie dialogues. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9820–9831, Vienna, Austria. Association for Computational Linguistics.

Burak Satar, Zhixin Ma, Patrick Amadeus Irawan, Wilfried Ariel Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. 2025. Seeing culture: A benchmark for visual reasoning and grounding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22227–22243, Suzhou, China. Association for Computational Linguistics.

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

Florian Schneider and Sunayana Sitaram. 2024. M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4309–4345, Miami, Florida, USA. Association for Computational Linguistics.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. volume 25 of *Advances in Experimental Social Psychology*, pages 1–65. Academic Press.

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.

Andrea Seveso, Daniele Potertì, Edoardo Federici, Mario Mezzanzanica, and Fabio Mercorio. 2025. ITALIC: An Italian culture-aware natural language benchmark. 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 1469–1478, Albuquerque, New Mexico. Association for Computational Linguistics.

Karolina Seweryn, Anna Kołos, Agnieszka Karlińska, Katarzyna Lorenc, Katarzyna Dziewulska, Maciej Chrabaszcz, Aleksandra Krasnodebska, Paula Betscher, Zofia Cieślińska, Katarzyna Kowol, Julia Moska, Dawid Motyka, Paweł Walkowiak, Bartosz Żuk, and Arkadiusz Janz. 2025. PLLuM-align: Polish preference dataset for large language model alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23879–23908, Suzhou, China. Association for Computational Linguistics.

Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, Mohammed Irfan Kurpath, Yahya Hmaiti, Go Inoue, Jean Lahoud, Md. Safirur Rashid, Shadid Intisar Quasem, Maheen Fatima, Franco Vidal, Mykola Maslych, Ketan Pravin More, Sanoojan Baliah, Hasindri Watawana, Yuhao Li, Fabian

Farestam, and 10 others. 2025. A culturally-diverse multilingual multimodal video benchmark & model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19998–20022, Suzhou, China. Association for Computational Linguistics.

Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569, Toronto, Canada. Association for Computational Linguistics.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022. DISARM: Detecting the victims targeted by harmful memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States. Association for Computational Linguistics.

Hua Shen, Nicholas Clark, and Tanu Mitra. 2025. Mind the value-action gap: Do LLMs act in alignment with their values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118, Suzhou, China. Association for Computational Linguistics.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. 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 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Shurong Sheng, Luc Van Gool, and Marie-Francine Moens. 2016. A dataset for multimodal question answering in the cultural heritage domain. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 10–17, Osaka, Japan. The COLING 2016 Organizing Committee.

Anudeex Shetty, Amin Beheshti, Mark Dras, and Usman Naseem. 2025. VITAL: A new dataset for benchmarking pluralistic alignment in healthcare. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22954–22974, Vienna, Austria. Association for Computational Linguistics.

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.

Daiki Shiono, Ana Brassard, Yukiko Ishizuki, and Jun Suzuki. 2025. Evaluating model alignment with human perception: A study on shitsukan in LLMs and LVLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11428–11444, Abu Dhabi, UAE. Association for Computational Linguistics.

Pratik Rakesh Singh, Kritarth Prasad, Mohammadi Zaki, and Pankaj Wasnik. 2025a. Graph-assisted culturally adaptable idiomatic translation for Indic languages. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7029–7044, Vienna, Austria. Association for Computational Linguistics.

Punit Kumar Singh, Nishant Kumar, Akash Ghosh, Kunal Pasad, Khushi Soni, Manisha Jaishwal, Sriparna Saha, Syukron Abu Ishaq Alfarozi, Asres Temam Abagissa, Kitsuchart Pasupa, Haiqin Yang, and Jose G Moreno. 2025b. Let's Play Across Cultures: A Large Multilingual, Multicultural Benchmark for Assessing Language Models' Understanding of Sports. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15194–15241, Suzhou, China. Association for Computational Linguistics.

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

Sunayana Sitaram, Adrian de Wynter, Isobel McCrum, Qilong Gu, and Si-Qing Chen. 2025. A multilingual, culture-first approach to addressing misgendering in LLM applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31159–31183, Suzhou, China. Association for Computational Linguistics.

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.

Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya Joshi. 2025. Evaluating dialect robustness of language models via conversation understanding. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 24–38, Abu Dhabi. Association for Computational Linguistics.

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. 2021. Cross-cultural similarity features for cross-lingual transfer learning of pragmatically motivated tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2403–2414, Online. Association for Computational Linguistics.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward informal language processing: Knowledge of slang in large language models. 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 1683–1701, Mexico City, Mexico. Association for Computational Linguistics.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. SEA-HELM: Southeast Asian holistic evaluation of language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12308–12336, Vienna, Austria. Association for Computational Linguistics.

Xin Tan, Bowei Zou, and AiTi Aw. 2025. A benchmark for translations across styles and language variants. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2389–2402, Suzhou, China. Association for Computational Linguistics.

Eshaan Tanwar, Anwoy Chatterjee, Michael Saxon, Alon Albalak, William Yang Wang, and Tanmoy Chakraborty. 2025. Do You Know About My Nation? Investigating Multilingual Language Models' Cultural Literacy Through Factual Knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14956–14979, Suzhou, China. Association for Computational Linguistics.

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. _eprint: https://academic.oup.com/pnasnexus/article-pdf/3/9/pgae346/59151559/pgae346.pdf.

Allahsera Auguste Tapo, Kevin Assogba, Christopher M Homan, M. Mustafa Rafique, and Marcos Zampieri. 2025a. Bayelemabaga: Creating resources for Bambara NLP. 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 12060–12070, Albuquerque, New Mexico. Association for Computational Linguistics.

Allahsera Auguste Tapo, Nouhoum Coulibaly, Seydou Diallo, Sebastien Diarra, Christopher M Homan, Mamadou K. Keita, and Michael Leventhal. 2025b. GAIfE: Using GenAI to improve literacy in low-resourced settings. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7929–7944, Albuquerque, New Mexico. Association for Computational Linguistics.

Yi Tay, Donovan Ong, Jie Fu, Alvin Chan, Nancy Chen, Anh Tuan Luu, and Chris Pal. 2020. Would you rather? a new benchmark for learning machine alignment with cultural values and social preferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5369–5373, Online. Association for Computational Linguistics.

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.

Mukhammed Togmanov, Nurdaulet Mukhituly, Diana Turmakhan, Jonibek Mansurov, Maiya Goloburda, Akhmed Sakip, Zhuohan Xie, Yuxia Wang, Bekassyl Syzdykov, Nurkhan Laiyk, Alham Fikri Aji, Ekaterina Kochmar, Preslav Nakov, and Fajri Koto. 2025. KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge of Kazakhstan. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14403–14416, Vienna, Austria. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Ah Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, Dietrich Klakow, and Seid Muhie Yimam. 2024. EthioLLM: Multilingual large language models for Ethiopian languages with task evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6341–6352, Torino, Italia. ELRA and ICCL.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott A. Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.

Jackson Trager, Francielle Vargas, Diego Alves, Matteo Guida, Mikel K. Ngueajio, Ameeta Agrawal, Yalda Daryani, Farzan Karimi Malekabadi, and Flor Miriam Plaza-del Arco. 2025. MFTCXplain: A multilingual benchmark dataset for evaluating the moral reasoning of LLMs through multi-hop hate speech explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15709–15740, Suzhou, China. Association for Computational Linguistics.

Ayuto Tsutsumi and Yuu Jinnai. 2025. Do large language models know folktales? a case study of yokai in Japanese folktales. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16124–16146, Vienna, Austria. Association for Computational Linguistics.

Faizad Ullah, Ali Faheem, Ubaid Azam, Muhammad Sohaib Ayub, Faisal Kamiran, and Asim Karim. 2024. Detecting cybercrimes in accordance with Pakistani law: Dataset and evaluation using PLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4717–4728, Torino, Italia. ELRA and ICCL.

Sanzhar Umbet, Sanzhar Murzakhmetov, Beksultan Sagyndyk, Kirill Yakunin, Timur Akishev, and Pavel Zubitski. 2025. KazBench-KK: A cultural-knowledge benchmark for Kazakh. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 38–57, Vienna, Austria. Association for Computational Linguistics.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast Asian cultural context dataset for visual question answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Bangkok, Thailand. Association for Computational Linguistics.

Viacheslav Vasilev, Julia Agafonova, Nikolai Gerasimenko, Alexander Kapitanov, Polina Mikhailova, Evelina Mironova, and Denis Dimitrov. 2025. RusCode: Russian cultural code benchmark for text-to-image generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7656–7672, Albuquerque, New Mexico. Association for Computational Linguistics.

Justin Vasselli, Eunike Andriani Kardinata, Yusuke Sakai, and Taro Watanabe. 2025. Multilingual dialogue generation and localization with dialogue act scripting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32896–32911, Suzhou, China. Association for Computational Linguistics.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2025. Navigating cultural chasms: Exploring and unlocking the cultural POV of text-to-image models. *Transactions of the Association for Computational Linguistics*, 13:142–166.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. MILU: A multi-task Indic language understanding benchmark. 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 10076–10132, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Tom Völker, Jan Pfister, and Andreas Hotho. 2025. SALT at SemEval-2025 task 2: A SQL-based approach for LLM-free entity-aware-translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 852–864, Vienna, Austria. 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.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Minghan Wang, Viet Thanh Pham, Farhad Moghimifar, and Thuy-Trang Vu. 2025a. Proverbs run in pairs: Evaluating proverb translation capability of large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1646–1662, Vienna, Austria. Association for Computational Linguistics.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025b. Multilingual prompting for improving LLM

generation diversity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6367–6389, Suzhou, China. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024c. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. 2024d. KULTURE bench: A benchmark for assessing language model in Korean cultural context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 914–927, Tokyo, Japan. Tokyo University of Foreign Studies.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024e. CMB: A comprehensive medical benchmark in Chinese. 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 6184–6205, Mexico City, Mexico. Association for Computational Linguistics.

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

Zeqiang Wang, Jon Johnson, and Suparna De. 2025c. DLIR: Spherical adaptation for cross-lingual knowledge transfer of sociological concepts alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2061–2075, Suzhou, China. Association for Computational Linguistics.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7900–7932, Miami, Florida, USA. Association for Computational Linguistics.

Yuting Wei, Yuanxing Xu, Xinru Wei, Simin Yang, 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.

Isadora White, Sashrika Pandey, and Michelle Pan. 2024. Communicate to play: Pragmatic reasoning for efficient cross-cultural communication. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12201–12216, Miami, Florida, USA. Association for Computational Linguistics.

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.

Steven Wilson, Rada Mihalcea, Ryan Boyd, and James Pennebaker. 2016. Disentangling topic models: A cross-cultural analysis of personal values through words. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 143–152, Austin, Texas. Association for Computational Linguistics.

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

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

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.

Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2024. Understanding slang with LLMs: Modelling cross-cultural nuances through paraphrasing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15525–15531, Miami, Florida, USA. Association for Computational Linguistics.

Yubo Xie, Chenkai Wang, Zongyang Ma, and Fahui Miao. 2025. Are large language models chronically online surfers? a dataset for Chinese Internet meme explanation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17062–17083, Suzhou, China. Association for Computational Linguistics.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. 2024. Exploring multilingual concepts of human values in large language models: Is value alignment consistent, transferable and controllable across languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1771–1793, Miami, Florida, USA. Association for Computational Linguistics.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025a. 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.

Zhijun Xu, Siyu Yuan, Yiqiao Zhang, Jingyu Sun, Tong Zheng, and Deqing Yang. 2025b. PunMemeCN: A benchmark to explore vision-language models' understanding of Chinese pun memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18694–18710, Suzhou, China. Association for Computational Linguistics.

Zhijun Xu, Siyu Yuan, Yiqiao Zhang, Jingyu Sun, Tong Zheng, and Deqing Yang. 2025c. PunMemeCN: A Benchmark to Explore Vision-Language Models' Understanding of Chinese Pun Memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18694–18710, Suzhou, China. Association for Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Srishti Yadav, Zhi Zhang, Daniel Hershcovich, and Ekaterina Shutova. 2025. Beyond words: Exploring cultural value sensitivity in multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7607–7623, Albuquerque, New Mexico. Association for Computational Linguistics.

Silvana Yakhni and Ali Chehab. 2025. Can LLMs translate cultural nuance in dialects? a case study on Lebanese Arabic. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 114–135, Abu Dhabi, UAE. Association for Computational Linguistics.

Taisei Yamamoto, Ryoma Kumon, Danushka Bollegala, and Hitomi Yanaka. 2025. Bias mitigation or cultural commonsense? evaluating LLMs with a Japanese dataset. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17295–17313, Suzhou, China. Association for Computational Linguistics.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. 2025a. NüshuRescue: Reviving the Endangered Nüshu Language with AI. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7020–7034, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Yahan Yang, Soham Dan, Shuo Li, Dan Roth, and Insup Lee. 2025c. MrGuard: A multilingual reasoning guardrail for universal LLM safety. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27377–27396, Suzhou, China. 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.

Amir Hossein Yari and Fajri Koto. 2025. Unveiling cultural blind spots: Analyzing the limitations of mLLMs in procedural text comprehension. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20151–20170, Vienna, Austria. 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.

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

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

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Da Yin, Haoyi Qiu, Kung-Hsiang Huang, Kai-Wei Chang, and Nanyun Peng. 2024. SafeWorld: Geo-Diverse Safety Alignment. In *Advances in Neural Information Processing Systems*, volume 37, pages 128734–128768. Curran Associates, Inc.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22230–22251, Vienna, Austria. Association for Computational Linguistics.

Hao Yu, Jesujoba Oluwadara Alabi, Andiswa Bukula, Jian Yun Zhuang, En-Shiun Annie Lee, Tadesse Kebede Guge, Israel Abebe Azime, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, and 2 others. 2025. INJONGO: A multicultural intent detection and slot-filling dataset for 16 African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9429–9452, Vienna, Austria. Association for Computational Linguistics.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Tao Liu, 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.

Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng, Minghao Wu, Xuanfan Ni, Tianqi Shi, Yu Zhao, Yefeng Liu, Chenyu Zhu, Ruizhe Li, Jiahui Geng, Qing Li, Yu Tong, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. Marco-bench-MIF: On multilingual instruction-following capability of large language. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24058–24072, Vienna, Austria. 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, Zhiyuan Liao, and Yansong Feng. 2025a. Cross-lingual transfer of cultural knowledge: An asymmetric phenomenon. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 147–157, Vienna, Austria. Association for Computational Linguistics.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024a. 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.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang,

Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, and 2 others. 2025b. Can MLLMs understand the deep implication behind Chinese images? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14402, Vienna, Austria. Association for Computational Linguistics.

Jian Zhang, Junyi Guo, Junyi Yuan, Huanda Lu, Yanlin Zhou, Fangyu Wu, Qiufeng Wang, and Dongming Lu. 2025c. LLM-driven completeness and consistency evaluation for cultural heritage data augmentation in cross-modal retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19407–19417, Suzhou, China. Association for Computational Linguistics.

Jinghao Zhang, Sihang Jiang, Shiwei Guo, Shisong Chen, Yanghua Xiao, Hongwei Feng, Jiaqing Liang, Minggui HE, Shimin Tao, and Hongxia Ma. 2025d. CultureScope: A Dimensional Lens for Probing Cultural Understanding in LLMs. *arXiv preprint*. ArXiv:2509.16188 [cs].

Ran Zhang, Wei Zhao, Lieve Macken, and Steffen Eger. 2025e. LiTransProQA: An LLM-based literary translation evaluation metric with professional question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29099–29121, Suzhou, China. Association for Computational Linguistics.

Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Kiana Avestimehr, Katharine Butler, Yanjun Weng, Mi Zhang, Shrikanth Narayanan, and Salman Avestimehr. 2025f. Creating a lens of Chinese culture: A multimodal dataset for Chinese pun rebus art understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22473–22487, Vienna, Austria. Association for Computational Linguistics.

Xinyu Zhang, Pei Zhang, Shuang Luo, Jialong Tang, Yu Wan, Baosong Yang, and Fei Huang. 2025g. CultureSynth: A hierarchical taxonomy-guided and retrieval-augmented framework for cultural question-answer synthesis. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10448–10467, Suzhou, China. Association for Computational Linguistics.

Yuxuan Zhang, Yangfu Zhu, Haorui Wang, and Bin Wu. 2025h. Interesting culture: Social relation recognition from videos via culture de-confounding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5174–5184, Suzhou, China. Association for Computational Linguistics.

Zhonghe Zhang, Xiaoyu He, Vivek Iyer, and Alexandra Birch. 2024b. Cultural adaptation of menus: A fine-grained approach. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1258–1271, Miami, Florida, USA. Association for Computational Linguistics.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A. Hedderich. 2025. MAKIEval: A multilingual automatic WiKidata-based framework for cultural awareness evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23104–23136, Suzhou, China. Association for Computational Linguistics.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024a. 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.

Yuan Zhao, Ruiquan Zhang, Dengfeng Yao, and Yidong Chen. 2024b. Translation quality evaluation of sign language avatar. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 405–415, Taiyuan, China. Chinese Information Processing Society of China.

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, and Alan Ritter. 2022. Stanceosaurus: Classifying stance towards multicultural misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2132–2151, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiayou Zhong, Anudeex Shetty, Chao Jia, Xuanrui Lin, and Usman Naseem. 2025. Pluralistic alignment for healthcare: A role-driven framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31320–31343, Suzhou, China. Association for Computational Linguistics.

Haiqi Zhou, David G Hobson, Derek Ruths, and Andrew Piper. 2024. Large scale narrative messaging around climate change: A cross-cultural comparison. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 143–155, Bangkok, Thailand. Association for Computational Linguistics.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12684–12702, Singapore. Association for Computational Linguistics.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2025a. 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.

Li Zhou, Lutong Yu, Dongchu Xie, Shaohuan Cheng, Wenyan Li, and Haizhou Li. 2025b. Hanfu-bench: A multimodal benchmark on cross-temporal cultural understanding and transcreation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24616–24638, Suzhou, China. Association for Computational Linguistics.

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025c. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*.

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, William Barr Held, Jane Yu, Amir Goldberg, David Grusky, and Diyi Yang. 2025. Culture cartography: Mapping the landscape of cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1739–1757, Suzhou, China. Association for Computational Linguistics.

Chenye Zou, Xingyue Wen, Tianyi Hu, Qian Janice Wang, and Daniel Hershcovich. 2025. Do LLMs understand wine descriptors across cultures? a benchmark for cultural adaptations of wine reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1875–1894, Suzhou, China. Association for Computational Linguistics.

A Detailed Literature Collection and Annotation Methodology

A.1 Phase 1: Initial Seed Collection

We took a two-phased approach to collect relevant literature, focusing on the past five years (including all available papers from 2025) to capture the modern landscape of Cultural NLP. The first phase consisted of a manual search combined with an automated scrape of the ACL Anthology. We queried for papers containing the words "culture" or "cultural" in the title, alongside at least one other keyword related to investigating or operationalizing culture in NLP. This initial highly targeted search resulted in a seed set of 180 papers.

A.2 Phase 2: Broad Scrape and LLM-Assisted Filtering

To ensure comprehensive coverage, the second phase involved a broader scrape of the ACL Anthology for any paper containing "culture" or "cultural" in either the title or abstract, returning 1,216 additional candidates. Given the volume, we employed an LLM to assist in filtering the list down to a highly relevant subset.

First, we prompted the LLM with a description of our inclusion criteria along with the title and abstract of each paper, retaining only those assigned a high relevance score. Next, we prompted the LLM with stricter inclusion guidelines, providing it with potentially relevant full-text sections of each candidate paper (e.g., Introduction, Dataset, Methods, Evaluation). To ensure we analyzed works with thorough descriptions and evaluations of proposed datasets or methods, we filtered out short papers, considering only those longer than 6 pages. This rigorous second phase yielded 277 additional highly relevant papers.

A.3 Manual Annotation and Consensus Strategy

The combined pool of candidate papers was manually processed by annotators who are Computer Science PhD students actively involved in NLP research. Papers were annotated across three main dimensions: (1) Cultural Capabilities addressed, (2) dataset creation methodology, and (3) technical system methods utilized. Throughout the annotation process, we performed a manual filtering step to discard papers that did not pose a novel contribution in any of these three aspects. Specifically, we removed works that solely evaluated existing models on existing datasets, or papers utilizing NLP tools for pure sociolinguistic or social science analyses without proposing a technical system or dataset improvement. This exclusion phase removed 79 papers, leaving a final corpus of 378 papers to be read thoroughly and annotated.

To ensure high-quality and consistent annotations, regular meetings were held between the surveyors. During these meetings, precise definitions of the labels were discussed, disagreements were resolved, and edge cases were evaluated. We also conducted cross-validations on small subsamples of papers to purposefully identify and resolve points of contention. Once this iterative process concluded and the taxonomy definitions were firmly established, the team went back and corrected all previously labeled papers to guarantee strict consistency and adherence to the final taxonomy across the entire 378-paper corpus.

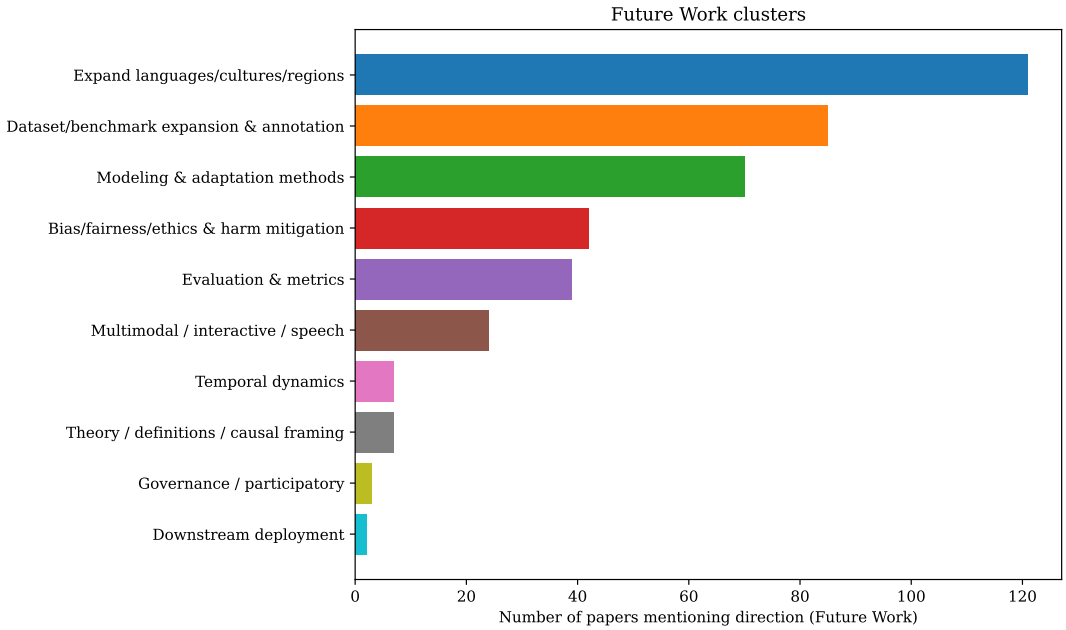

Figure 5: Frequency of future-work clusters in the papers

B Future Work Cluster Analysis

We analyze the Future Work sections recorded in our annotated spreadsheet. The aggregated results are summarized in Figure 5. We identify recurring directions through a transparent and reproducible clustering procedure. Because individual papers often propose multiple extensions, clustering is treated as a multi-label assignment problem. A single future work entry can therefore belong to multiple clusters.

B.1 Cluster schema

We define a small set of coarse themes that repeatedly appear across the surveyed corpus. Each cluster represents a practical research direction rather than a topical category. The clusters and their descriptions are as follows:

- **Expand languages and cultures**: extending coverage to additional languages, dialects, regions, or cultural contexts.
- **Dataset and benchmark expansion**: collecting new datasets, enlarging benchmarks, adding annotations, or improving dataset curation and documentation.
- **Modeling and adaptation methods**: developing improved modeling, training, fine-tuning, alignment, transfer, retrieval, or adaptation approaches that explicitly address cultural phenomena.
- **Bias, fairness, ethics, and harm mitigation**: measuring or mitigating culturally shaped bias, stereotyping, toxicity, or other harmful outputs.
- **Evaluation and metrics**: proposing stronger evaluation metrics, human evaluation protocols, robustness checks, or generalization tests for cultural capabilities.
- **Multimodal or interactive settings**: extending cultural modeling to speech, audio, vision, dialogue systems, or other interactive environments.
- **Temporal dynamics**: modeling culture as a time-varying process, including updates, drift, or longitudinal change.
- **Theory, definitions, and causal framing**: clarifying conceptual definitions of culture or proposing causal or mechanistic interpretations of cultural variables.
- **Governance and participatory approaches**: incorporating community participation, consent mechanisms, stewardship practices, or governance frameworks.
- **Downstream deployment**: evaluating integration into real-world systems and conducting deployment-oriented validation.

B.2 Assignment procedure

Cluster assignments are produced using a dictionary-based matching procedure. For each cluster, we construct a small set of indicative surface forms, such as "multilingual", "dialect", "benchmark", "annotation", "robustness", "toxicity", "speech", or "drift". A cluster is marked as present when any of its indicative patterns appear in the corresponding future work text using case-insensitive matching. This design prioritizes interpretability and reproducibility: cluster definitions are explicit, and assignments can be verified by inspecting the matched expressions. The clustering procedure is intentionally lightweight and surface form-driven. It may undercount papers that describe a direction using uncommon phrasing, and it does not disambiguate cases where a keyword appears with a different meaning. We therefore treat the clusters as a descriptive summary of recurring future work themes rather than as an exhaustive taxonomy.

C Analysis of Methods used for CCs

C.1 Methods across cultural capabilities

Figure 6 presents a 3×3 grid of stacked bar charts comparing method families across cultural capabilities. Each subplot corresponds to one cultural capability. The x-axis lists methodological families, while the y-axis reports the number of contributions. Each bar is partitioned into segments representing four contribution types: System Contribution, Model Evaluation, Data Generation, and System Improvement. Overall, Cultural Knowledge and Value-driven Cultural Alignment contain the largest number of contributions, whereas Cultural Education, Survey-based Cultural Alignment, and Multimodal Cultural Knowledge appear relatively sparse. Across cultural capabilities, prompting-based methods are the most frequently used approach, with training-based methods and lexica, embeddings, or other classical NLP techniques also appearing regularly. Some cultural capabilities display clearer methodological specialization. Cultural Computational Representation relies more heavily on classical NLP-style approaches, while Cultural Translation more often adopts agent-based and prompting-based strategies.

A key observation is that cultural NLP research is methodologically heterogeneous. Different tasks tend to attract different technical approaches rather than converging on a single dominant paradigm. Current work is largely centered on prompting and training-based techniques, particularly for tasks involving cultural knowledge, value alignment, and safety, reflecting the broader influence of large language models in recent NLP research. In contrast, tasks related to cultural representation and sociolinguistic competence maintain stronger connections to earlier NLP traditions. The figure also highlights several relatively underexplored areas, including cultural education, multimodal cultural reasoning, and mechanistic interpretability, indicating opportunities for future work to expand both task coverage and methodological diversity.

C.2 Cultural capabilities across methods

Figure 7 presents the distribution of cultural capabilities across method families using a 2×3 grid of bar charts. Each subplot corresponds to a method family, and the horizontal axis lists cultural capabilities, while the vertical axis reports the number of contributions. The six method families are retrieval and multimodel approaches, human-in-the-loop methods, lexica, embeddings, classical NLP techniques, multimodal approaches, prompting-based methods, and training-based methods. Across methods, prompting-based approaches exhibit the largest overall volume and cover the widest range of cultural capabilities, followed by training-based methods and lexica, embeddings, and classical NLP techniques. In contrast, human-in-the-loop and multimodal approaches account for substantially fewer contributions and are concentrated in a limited set of cultural capabilities. At the capability level, cultural knowledge, value-driven cultural alignment, sociolinguistic competence, and cultural computational representation appear repeatedly across multiple method families. Cultural education and survey-based cultural alignment remain comparatively limited. The stacked bar segments indicate that many methods support multiple contribution types, including system contribution, model evaluation, data generation, and system improvement, rather than focusing on a single contribution category.

The figure further suggests that the field is structured around a small set of widely reusable technical paradigms, particularly prompting-based and training-based approaches, which support a broad range of cultural capabilities. This pattern indicates that much of cultural NLP research adapts general-purpose large language model techniques to culture-related problems rather than develop-

ing highly task-specific architectures. At the same time, method families differ in flexibility. Lexica, embeddings, and classical NLP techniques remain particularly relevant for representational and sociolinguistic analysis, whereas multimodal and human-in-the-loop approaches appear more specialized and comparatively underexplored. Overall, the distribution indicates that research diversity varies not only across cultural capabilities but also across technical methods, with several approaches functioning as central methodological hubs while others remain peripheral.

C.3 Cultural Capabilities Across Dataset Creation Methods

Figure 8 presents an UpSet-style visualization of dataset creation method combinations, stacked by cultural capability. The bottom matrix indicates which dataset creation ingredients appear in each combination pattern. These ingredients include Cultural Experts or Native Speakers, Culture Knowledge Source, Synthetic Generation, Web Data, Expert Annotation, Translation, Crowdsourced Annotation, Common Language Data, Cultural and Social Science Theories, and Culture Value Surveys. The bars above the matrix report the frequency of each ingredient combination, with segments colored by cultural capability category. The bars on the left indicate the marginal frequency of individual ingredients. Cultural Experts or Native Speakers, Culture Knowledge Source, Synthetic Generation, and Web Data appear most frequently, while Cultural and Social Science Theories and Culture Value Surveys occur relatively rarely. At the combination level, the most frequent patterns reach approximately the high teens, while many other combinations appear only four to six times. Across these combinations, Cultural Knowledge contributes the largest share in many of the most common patterns, while Value-driven Cultural Alignment, Sociolinguistic Competence, and Cultural Translation also appear across multiple combinations.

Figure 9 shows a heap map of dataset creation method vs CCs. A key observation is that dataset creation for cultural natural language processing is typically compositional rather than based on a single source. Many studies construct datasets by combining several ingredients, most commonly human expertise from native speakers, external cultural knowledge resources, synthetic data generation, and web-collected data. This indicates that cultural information is rarely operationalized through a single data pipeline. Instead, researchers assemble datasets by integrating human knowledge, structured cultural resources, and scalable generation strategies. The figure also reveals an imbalance in the maturity of different cultural data construction approaches. High-frequency combinations are dominated by Cultural Knowledge-oriented pipelines, suggesting that dataset development has progressed most strongly for knowledge-centered tasks. In contrast, theoretically grounded sources such as cultural value surveys and cultural or social science theories remain uncommon. Overall, the distribution of combinations suggests that current practice prioritizes pragmatic and scalable data construction strategies, while theoretically grounded but less reusable approaches remain underexplored.

D CULTUREMINE

We accompany our survey with a web-interface, called CULTUREMINE, in the hope that it will facilitate research in cultural NLP. CULTUREMINE allows users to filter papers according to our taxonomy, as well as by the cultural groups they study. For example, if a user is interested in Cultural Safety and Harm Reduction in Korean, they can easily apply filters (shown in Figures 10 and 11) to find relevant papers (shown in Figure 12) and click on a paper's title to read it. In addition to providing direct, easy access to papers, CULTUREMINE provides dynamic bar charts at the top of the page, showing the most common Cultural Capabilities, System Methods, and Dataset Creation Methods used in the set of papers that satisfy the active set of filters. It also includes similar charts for the cultural proxies used, languages studied, and regions or countries studied (shown in Figure 13). To keep up with the fast-pace of research in cultural NLP, CULTUREMINE includes a link to a Google Form that allows anyone to submit an annotation for a paper that is not currently covered. Submitted annotations will be reviewed and/or modified manually before being added to ensure that all annotations displayed are of high-quality.

E Cultural Proxies and Groups

We perform a brief analysis of the cultures studied by the papers included in our survey. Figure 14 breaks down the types of cultural proxies used. 72.8% use language as the main proxy for culture, 22.2% use a geographic proxy (Country or Geo-

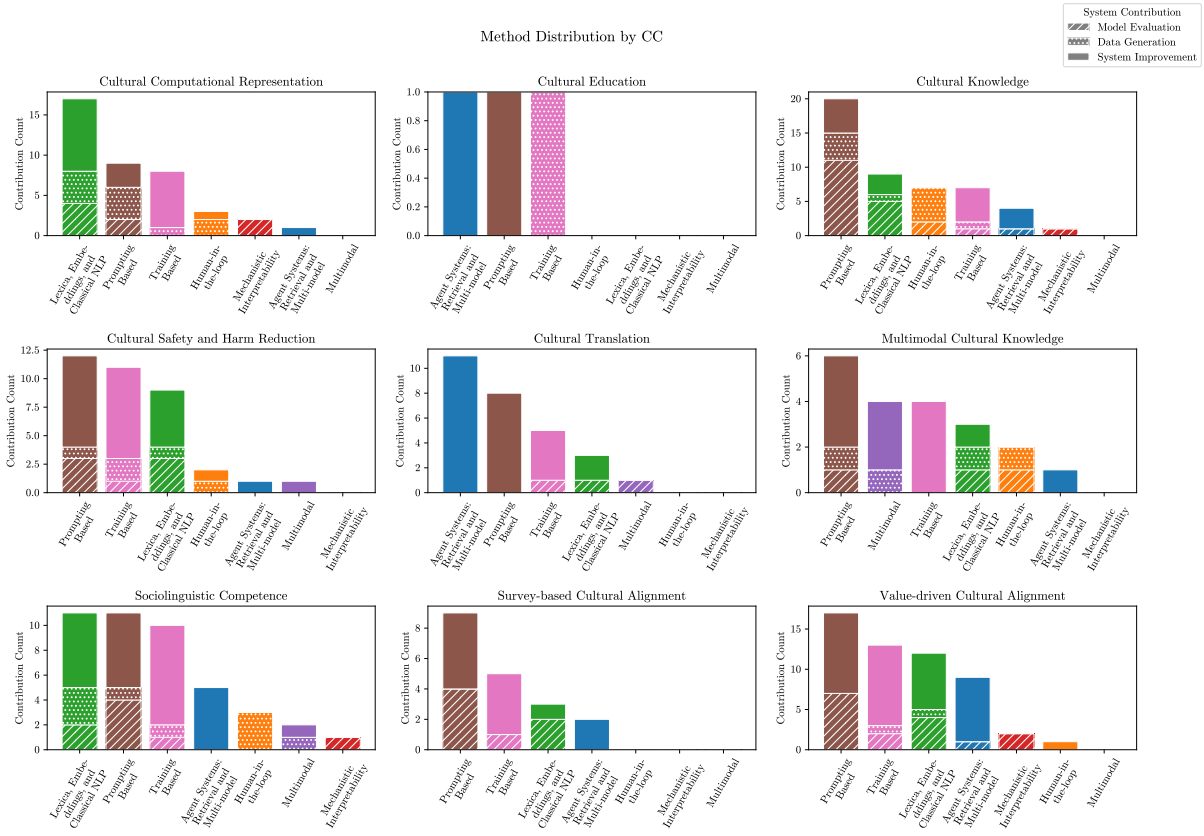

Figure 6: Method distribution across cultural capabilities

graphic Region), and only 5% use some other proxy. As discussed in (Zhou et al., 2025c), these commonly used proxies for culture can be problematic as the groups within them are often composed of multiple cultures. We further analyze the specific cultural groups included in papers that consider at most 10 different groups (235 using language and 50 using geography). Of the papers that use language, there is a very long tail distribution; over half include English and around a third include Chinese, with all but 9 languages—out of 113—being covered less than 20 times (Figure 15). We see a similar trend across 79 total geographic proxies, with around half including US and India, around a third including China, and all but 9 groups covered 5 or fewer times (Figure 16).

F List of Surveyed Papers

For compact presentation, we abbreviate cultural capability categories using acronyms in the following table **??** of surveyed papers. Specifically, CT = Cultural Translation; SCA = Survey-based Cultural Alignment; VCA = Value-driven Cultural Alignment; CK = Cultural Knowledge; MCK = Multimodal Cultural Knowledge; SC = Sociolinguistic Competence; CSHR = Cultural Safety and Harm Reduction; CE = Cultural Education; and CCR = Cultural Computational Representation. These abbreviations are used only for table presentation and correspond directly to the capability categories described in the main text.

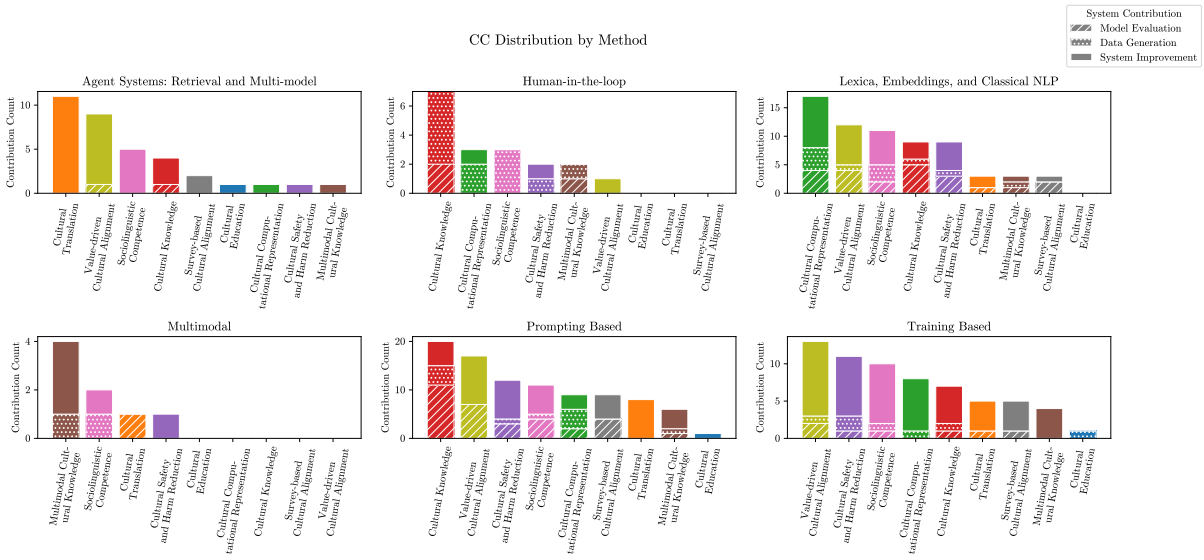

Figure 7: Cultural capability distribution across methods

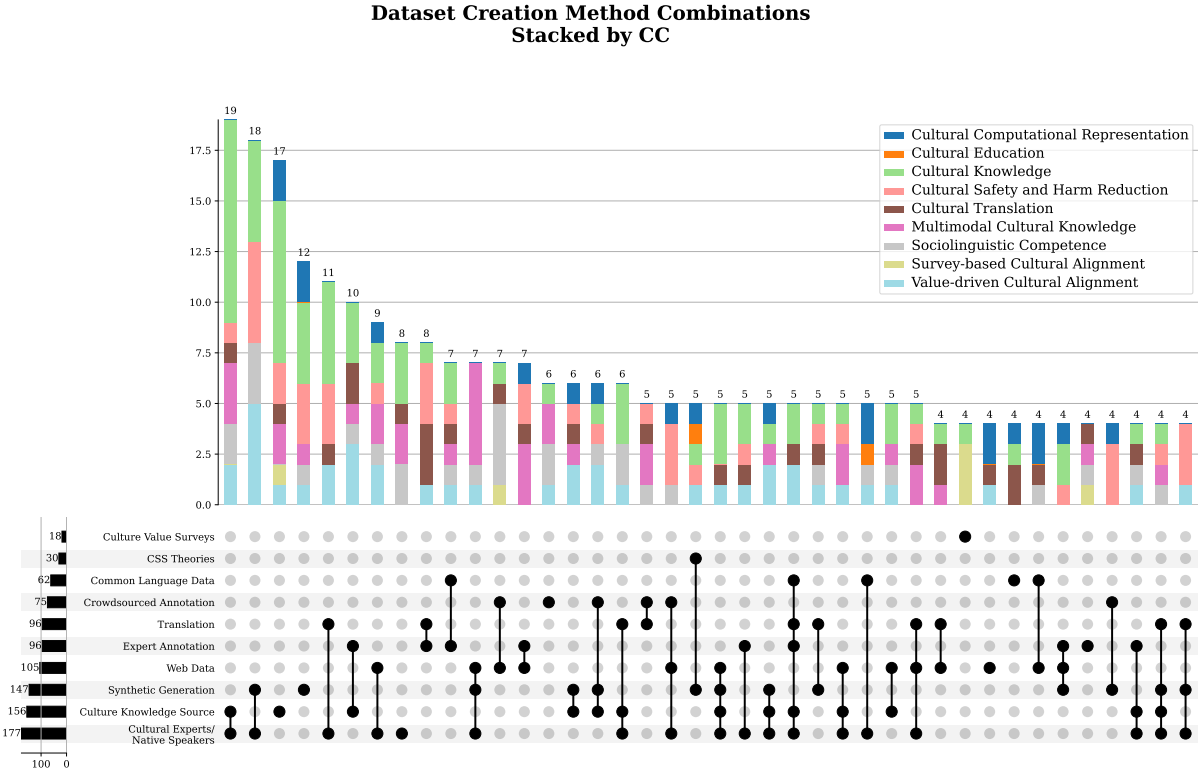

Figure 8: Dataset creation method combinations stacked by cultural capability

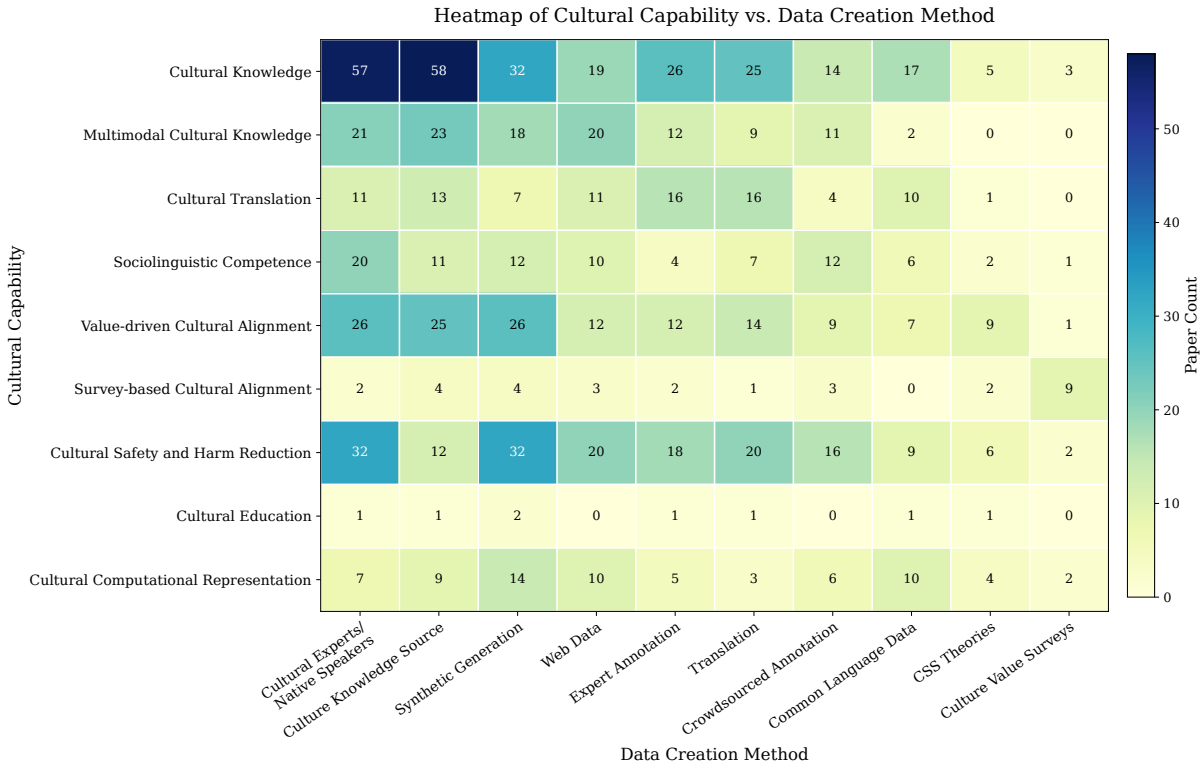

Figure 9: Heatmap of cultural capability by dataset creation method

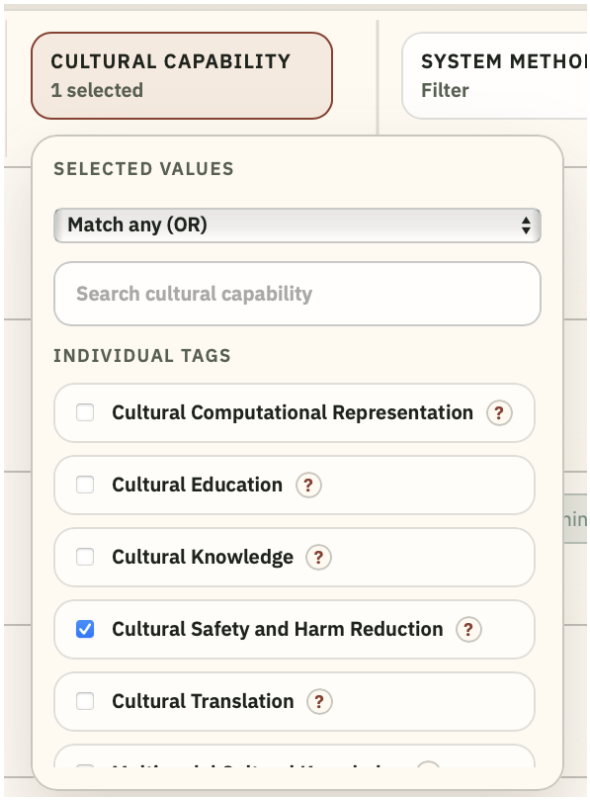

Figure 10: Cultural Capabilities Filter in CULTUREMINE

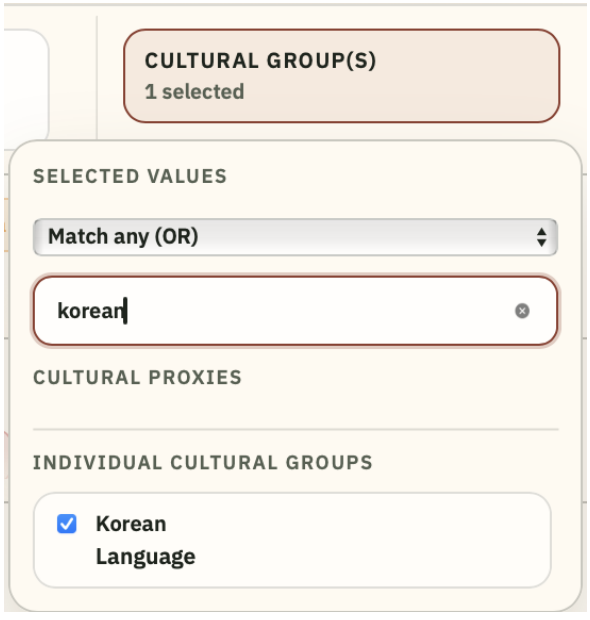

Figure 11: Cultural Groups Filter in CULTUREMINE

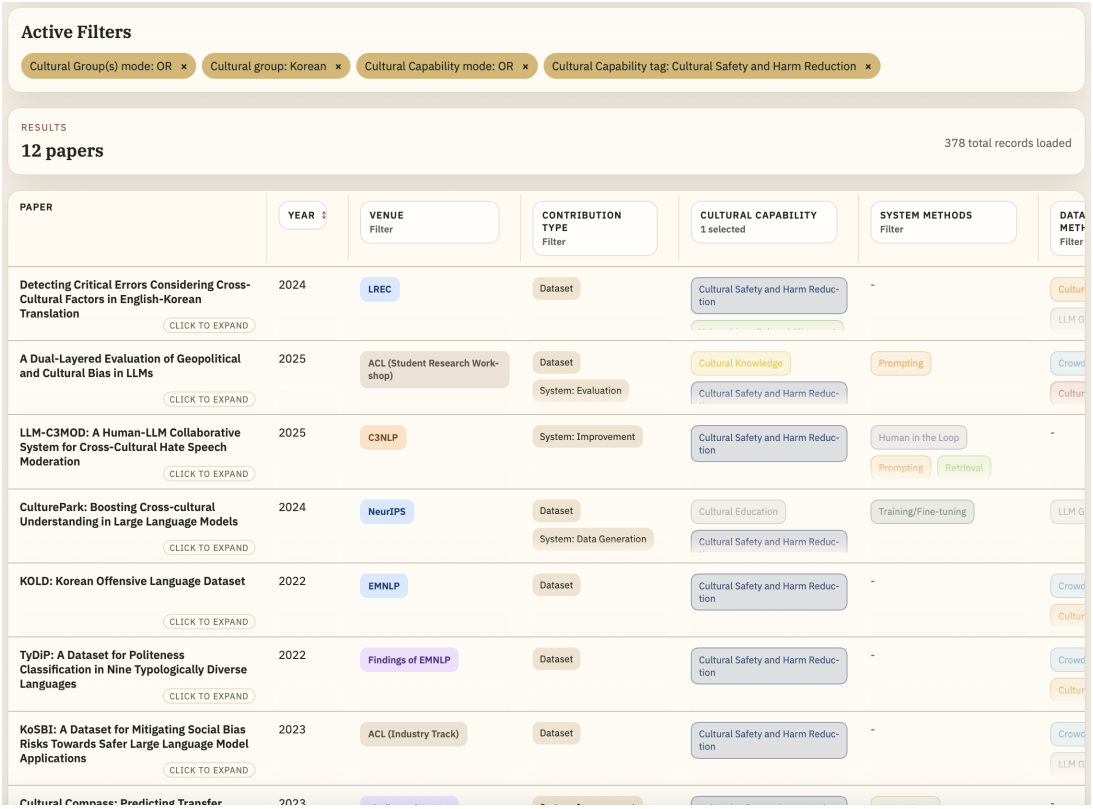

Figure 12: Filtered Paper List in CULTUREMINE

Figure 13: Screenshot of the Plots in CULTUREMINE

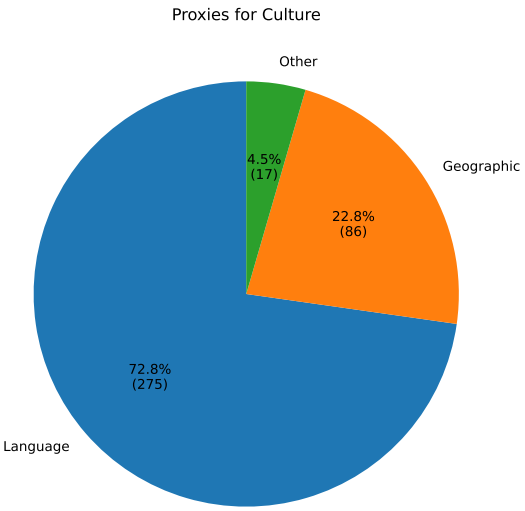

Figure 14: Pie Chart of Cultural Proxies Used

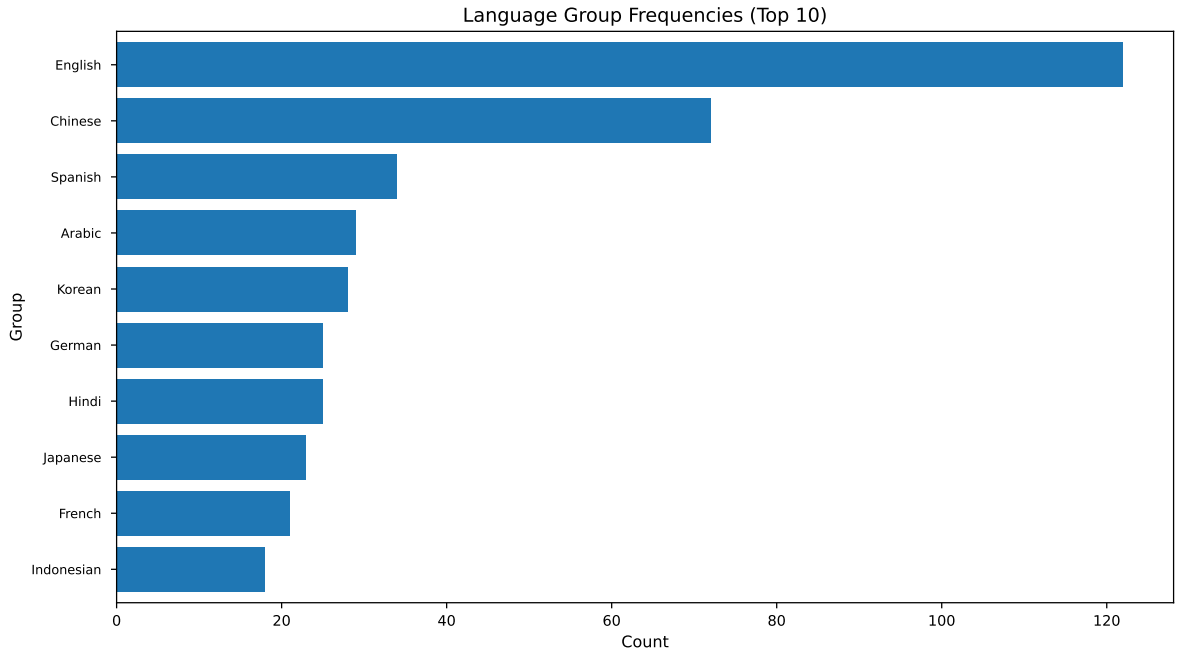

Figure 15: Bar Chart of the Frequencies of the 10 Most Commonly Studied Languages

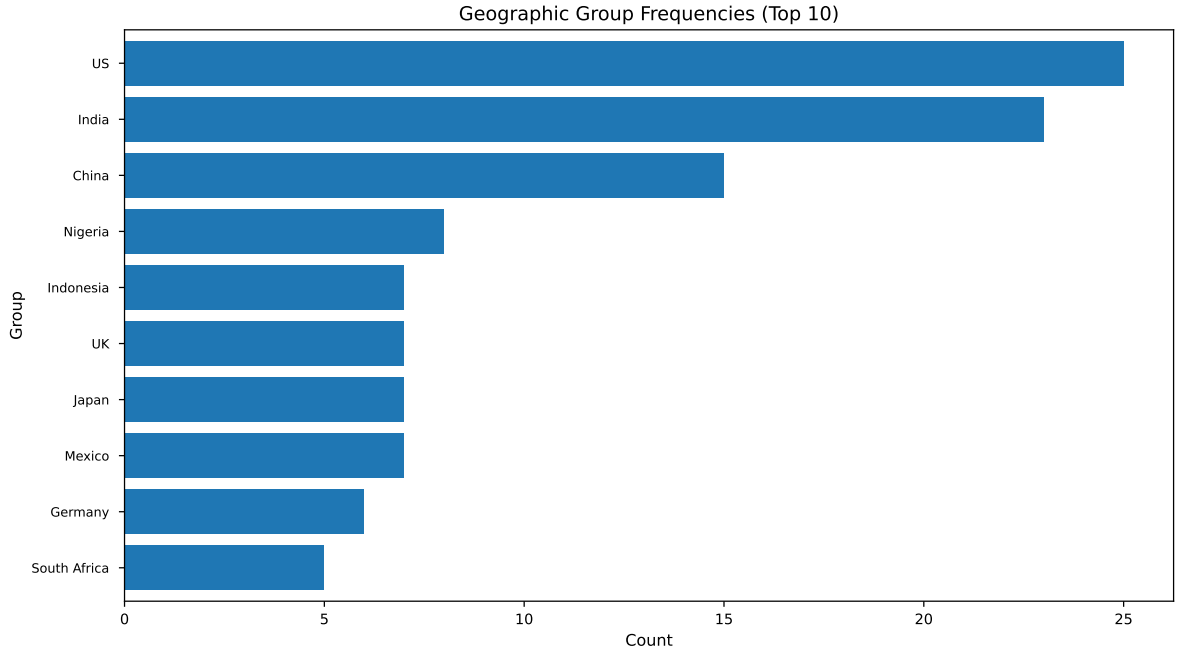

Figure 16: Bar Chart of the Frequencies of the 10 Most Commonly Studied Geographic

Paper	Culture capability
Wang et al. (2024a)	CK, CT, VCA
Singh et al. (2025c)	CK, VCA
Khanuja et al. (2020)	SC
Joshi et al. (2025)	CSHR
Nguyen et al. (2024a)	CCR, SC
Tao et al. (2024)	SCA
Nayak et al. (2024)	MCK
Xu et al. (2024)	VCA, CCR
Arora et al. (2023)	VCA, CCR
Cao et al. (2023)	SCA
Conia et al. (2024)	CT
Sun et al. (2021)	SC
AlKhamissi et al. (2024)	SCA
Masoud et al. (2025)	SCA
Jinnai (2024)	SC
Xu et al. (2025a)	SCA
Ki et al. (2025)	VCA
Sun et al. (2024)	CK
Havaldar et al. (2023b)	VCA
Ahmad et al. (2024)	VCA
Bui et al. (2025b)	CSHR
Bhatia et al. (2024)	MCK
Hale et al. (2025)	VCA, SC
Cecilia Liu et al. (2024)	CK, SC
Schneider and Sitaram (2024)	MCK
Cahyawijaya et al. (2025b)	MCK
Tay et al. (2020)	SCA
Zheng et al. (2022)	VCA
Mohamed et al. (2022)	MCK
Anegundi et al. (2022)	CSHR, CCR
Jha et al. (2023)	CSHR, CCR
Akinade et al. (2023)	CT
Das et al. (2023)	CSHR
Bauer et al. (2023)	VCA
Mukherjee et al. (2023)	CSHR
Shaikh et al. (2023)	SC
Palta and Rudinger (2023)	CK
Hu et al. (2023)	VCA
Choenni et al. (2024)	CCR
Naous et al. (2024)	VCA
Urailertprasert et al. (2024)	MCK
Prieto et al. (2024)	CT
Wang et al. (2024f)	VCA
Zhou et al. (2024)	VCA
Hu et al. (2024)	CT
Khanuja et al. (2024)	MCK
Wuraola et al. (2024)	SC
Li et al. (2024e)	MCK
Putri et al. (2024)	CK
Mohamed et al. (2024)	MCK
Giuliani et al. (2024)	CK
Li et al. (2024d)	VCA
Cao et al. (2024)	VCA
Zhan et al. (2024)	VCA, SC
Hsieh et al. (2024)	CSHR
White et al. (2024)	SC
Yao et al. (2024a)	CT
Bhatt and Diaz (2024)	VCA

Table continues

Paper	Culture capability
Cadotte et al. (2024)	CT
Kim et al. (2024a)	CK
Eo et al. (2024)	CSHR, VCA
Zhao et al. (2024a)	SCA, CK
Fort et al. (2024)	CSHR
Havaldar et al. (2024)	VCA
Lee et al. (2024b)	CSHR
Abdelkadir et al. (2024)	SC
Wang et al. (2024d)	CK
Haberland et al. (2024)	CT
Tonneau et al. (2024)	CSHR
Yakhni and Chehab (2025)	CT
Arora et al. (2025)	CK
Belay et al. (2025)	CK
Park et al. (2025a)	MCK
Chiu et al. (2025)	CK
Yang et al. (2025b)	SC
Havaldar et al. (2025)	CT
Wu et al. (2025)	CK
Zhang et al. (2025a)	CK
Kim and Kim (2025)	CSHR, CK
Ignat et al. (2025)	SC
Park et al. (2025b)	CSHR
Rai et al. (2025a)	SC
Kim et al. (2025b)	CSHR
Kim et al. (2025a)	VCA
Mousi et al. (2025)	CK, SC
Pandey et al. (2025)	CSHR
Cheng et al. (2025)	CK
Umbet et al. (2025)	CK
Qiu et al. (2025)	VCA
Bhatia et al. (2025)	CCR
Yadav et al. (2025)	SCA
Vasilev et al. (2025)	MCK
Li et al. (2025)	VCA, CK
Maji et al. (2025a)	CK
Singh et al. (2025a)	CT
Schneider et al. (2025)	CK, MCK
Hasan et al. (2025)	CK
Kim and Lee (2025)	CK
Zhang et al. (2025f)	MCK
Ghaboura et al. (2025)	MCK
Reshetnikov and Marinescu (2025)	MCK
Anik et al. (2025)	CT
Onohara et al. (2025)	CK, MCK
Seveso et al. (2025)	CK
Bai et al. (2025)	CK
Winata et al. (2025)	MCK
Naous and Xu (2025)	CK
Saha et al. (2025)	SC
Berger and Ponti (2025)	MCK
Jeong et al. (2025)	MCK
Arnardóttir et al. (2025)	CK
Ventura et al. (2025)	CK, MCK
Paniv et al. (2025)	MCK
Ringel et al. (2019)	SC
Lin et al. (2018)	SC
Gutiérrez et al. (2016)	VCA, SC

Table continues

Paper	Culture capability
Sheng et al. (2016)	MCK
Wilson et al. (2016)	SC, VCA
Friedman et al. (2019)	CSHR, CCR
Maji et al. (2025b)	MCK
Cuevas et al. (2025)	CSHR
Sahoo et al. (2025)	CK
Satar et al. (2025)	MCK
Zhang et al. (2025d)	CK, VCA
Limkonchotiwat et al. (2025)	CK
Cao et al. (2025)	SCA
Jiang et al. (2025)	SCA
Ramezani and Xu (2023)	VCA
Cahyawijaya et al. (2025a)	CCR
Yin et al. (2024)	CSHR
Li et al. (2024c)	CE, CSHR, VCA
Ziems et al. (2025)	VCA, CK
Epure et al. (2020)	CCR
Hahm et al. (2020)	SC
Yin et al. (2021)	MCK
Milbauer et al. (2021)	CCR
Liu et al. (2021)	MCK
Ghosh et al. (2021)	CSHR
Kiesel et al. (2022)	SCA, VCA
Kim et al. (2022)	CK
Yin et al. (2022)	CK
Thai et al. (2022)	CT
Jeong et al. (2022)	CSHR
Maronikolakis et al. (2022)	CSHR
Nadejde et al. (2022)	CT, SC
Sharma et al. (2022)	CSHR
Srinivasan and Choi (2022)	CSHR
Herrera et al. (2022)	SC
Chandran Nair et al. (2022)	CCR
Mortensen et al. (2022)	CCR
Deshpande et al. (2022)	CSHR
Ziems et al. (2023)	CCR
Lee et al. (2023)	CSHR
Alshahrani et al. (2023)	CCR
Li and Zhang (2023)	MCK
CH-Wang et al. (2023)	CCR
Havaldar et al. (2023a)	CCR
Lahoti et al. (2023)	CK, VCA
Koto et al. (2023)	CK
Fung et al. (2023)	CCR
Keleg and Magdy (2023)	CCR
Kabra et al. (2023)	CK
Huang and Yang (2023)	CCR
Zhou et al. (2023)	CSHR
Bang et al. (2023)	VCA, CSHR
Wang et al. (2024c)	SCA
Zhang et al. (2024a)	CCR
Alwajih et al. (2024)	MCK
Chen et al. (2024)	CT
Casola et al. (2024)	CSHR, SC
Nguyen et al. (2024b)	CT, CSHR
Falk et al. (2024)	CCR
Zhao et al. (2024b)	CT
Pham et al. (2024)	CCR

Table continues

Paper	Culture capability
Feng et al. (2024a)	CK
Feng et al. (2024b)	SCA, VCA, CCR
Lovenia et al. (2024)	CT, CK, MCK, SC, CSHR, VCA
Watts et al. (2024)	CK
Acquaye et al. (2024)	CK
Aakanksha et al. (2024)	CSHR
Hobson et al. (2024)	VCA
Davani et al. (2024)	CSHR
Dammu et al. (2024)	CSHR, SC
Deas et al. (2024)	CCR
Wu et al. (2024)	CT
Yuan et al. (2024)	CK
Javed et al. (2024)	CCR
Lee et al. (2024a)	CK, SCA
Yu et al. (2024)	CK, VCA
Li et al. (2024f)	CK
Alyafeai et al. (2024)	CK
Kim et al. (2024b)	VCA
Wei et al. (2024)	CK
Plaza-del Arco et al. (2024)	CCR
Liu et al. (2024a)	VCA
Shi et al. (2024)	CK, CCR
Yüksel et al. (2024)	CK
Masala et al. (2024)	CK
Huang and Xiong (2024)	CSHR
Ullah et al. (2024)	CSHR
Seth et al. (2024)	CK
Agarwal et al. (2024)	VCA
Tonja et al. (2024)	CSHR, CT
Yarlott et al. (2024)	CK
Son et al. (2024)	CK
Benkler et al. (2024)	VCA
Grigoreva et al. (2024)	CSHR
Maronikolakis et al. (2024)	CSHR
Lou et al. (2024)	CT
Prabhakaran et al. (2024)	CSHR
España-Bonet and Barrón-Cedeño (2024)	CCR
Shen et al. (2024)	CK
Wang et al. (2024e)	CK
Mukherjee et al. (2024)	CK
Huang et al. (2024)	CK, VCA
Yao et al. (2024b)	VCA
Acharya et al. (2024)	CK
Liu et al. (2024b)	SC, CE
Piccirilli et al. (2024)	SC
Koto et al. (2024)	CK
Romanyshyn et al. (2024)	CK
Kiulian et al. (2024)	CK
España-Bonet et al. (2024)	CCR
Li et al. (2024a)	CCR, SC
Zhang et al. (2024b)	CT
Anastasi et al. (2024)	CSHR
Ng et al. (2024)	CSHR
Liu et al. (2025b)	VCA
Kumar and Jurgens (2025)	VCA
Sadallah et al. (2025)	CK

Table continues

Paper	Culture capability
Fang et al. (2025)	CK
Yu et al. (2025)	SC
Liu et al. (2025e)	MCK, CK
Zhang et al. (2025b)	MCK
Togmanov et al. (2025)	CK
Yari and Koto (2025)	CK
Bui et al. (2025a)	CK
Menis Mastromichalakis et al. (2025)	CSHR
Ying et al. (2025)	CK, SC
Feng et al. (2025)	CK
Isbarov et al. (2025)	VCA, CK
Shetty et al. (2025)	VCA
Zeng et al. (2025)	VCA, CK
Yerukola et al. (2025)	CSHR
Karamolegkou et al. (2025)	MCK
Bhatt et al. (2025)	SC
Farhansyah et al. (2025)	SC, CT
Liang et al. (2025)	CT
Chen et al. (2025d)	MCK
Nawale et al. (2025)	CSHR
Bayramli et al. (2025)	MCK
Montalan et al. (2025)	CK, CSHR
Grandury et al. (2025)	VCA, CK
Alwajih et al. (2025a)	CK, SC
Olaleye et al. (2025)	SC
Ishita and Mamidi (2025)	CT
Alhassoun et al. (2025)	CK, VCA
Bouamor et al. (2025)	SC
Alwajih et al. (2025b)	CK
Krsteski et al. (2025)	VCA
Kim and Johnson (2025)	CSHR
Pokharel and Agrawal (2025)	SC
Altammami (2025)	CT
Yang et al. (2025a)	CT
Shiono et al. (2025)	MCK
Kabir et al. (2025)	SCA
Rooein et al. (2025)	CSHR, CCR
Maji et al. (2025c)	MCK
Xuan et al. (2025)	SC
Sadr et al. (2025)	VCA
Fu et al. (2025)	CK
Shen et al. (2025)	VCA
Wang et al. (2025b)	CK, VCA
Azmi et al. (2025)	CSHR
El Mekki et al. (2025)	CT, VCA, CK
Hu et al. (2025)	CSHR
Ramezani and Xu (2025)	CK
Tanwar et al. (2025)	CK
Singh et al. (2025b)	CK, MCK
Hashmat et al. (2025)	CSHR
Liu et al. (2025d)	SC
Xie et al. (2025)	CK
Yamamoto et al. (2025)	CSHR
Aji and Cohn (2025)	CK, CT
Mia et al. (2025)	MCK, CSHR
Chen et al. (2025c)	CK
Liu et al. (2025c)	SCA
Xu et al. (2025c)	MCK

Table continues

Paper	Culture capability
Mukherjee et al. (2025)	VCA
Zhang et al. (2025c)	MCK
Shafique et al. (2025)	MCK
Al Ghallabi et al. (2025)	CK
Parappan and Henao (2025)	CSHR
Lavrouk et al. (2025)	MCK, CK
Calvo-Bartolomé et al. (2025)	CK
Chen et al. (2025b)	CT
Seweryn et al. (2025)	CK, CSHR, VCA
Ma et al. (2025a)	CSHR
Nyandwi et al. (2025)	MCK
Dai et al. (2025)	SCA
Zhou et al. (2025b)	MCK
Yang et al. (2025c)	CSHR
Mitran et al. (2025)	SCA
Roh et al. (2025)	CK
Zhang et al. (2025e)	CT
Sitaram et al. (2025)	CSHR
Zhong et al. (2025)	VCA
Nimo et al. (2025)	CK, CSHR
Guo et al. (2025)	CK
Vasselli et al. (2025)	SC, CT
Pramodya et al. (2025)	CK
Chen et al. (2025a)	CT
Hong et al. (2025)	SC, VCA
Hsieh et al. (2025)	MCK
R et al. (2025)	SC
Hosseinbeigi et al. (2025a)	CK
Korre et al. (2025)	CSHR
Saffari et al. (2025)	VCA, CSHR
Tapo et al. (2025b)	CE
Wang et al. (2025a)	CT
Lan et al. (2025)	CSHR
Dwivedi et al. (2025)	VCA
Sasu et al. (2025)	SC
Susanto et al. (2025)	CK, VCA, SC
Tsutsumi and Jinnai (2025)	CK
Hosseinbeigi et al. (2025b)	VCA, CK, CT
He et al. (2025)	CK
Das et al. (2025)	MCK, CSHR
Nahin et al. (2025)	CK
Bi et al. (2025)	MCK
Zou et al. (2025)	CT
Wang et al. (2025c)	CCR, CK
Tan et al. (2025)	CT
Zhang et al. (2025h)	SC
Mor-Lan et al. (2025)	CCR
Mubarak et al. (2025)	CSHR
Zhang et al. (2025g)	CK
Chae et al. (2025)	CK, VCA, CSHR
Ma et al. (2025b)	SC
Trager et al. (2025)	CSHR
Gupta et al. (2025)	SC
Dey et al. (2025)	SCA
Nayak et al. (2025)	MCK
Villa-Cueva et al. (2025)	CT
Wibowo et al. (2024)	CK
Alwajih et al. (2025c)	MCK
Zhao et al. (2025)	CK

Table continues

Paper	Culture capability
Malik et al. (2025)	SC, VCA
Caplan et al. (2025)	SC
Mohammadi et al. (2025)	SCA
Bennie et al. (2025)	CSHR
Rachamalla et al. (2025)	CK, CSHR, SC
Liao et al. (2025)	CT
Pujari and Goldwasser (2025)	VCA, CCR
Muhammad et al. (2025)	CSHR
Madhusudan et al. (2025)	CCR
Leteno et al. (2025)	VCA
Son et al. (2025)	CK
Ashraf et al. (2025)	CSHR
Khanuja et al. (2025)	MCK
Banerjee et al. (2025)	CSHR
Miehling et al. (2025)	VCA
Nikandrou et al. (2025)	MCK
Pham et al. (2025)	VCA, CSHR
Zhou et al. (2025a)	CK
Verma et al. (2025)	CK
Rai et al. (2025b)	VCA
Mitchell et al. (2025)	CSHR
Tapo et al. (2025a)	CT
Magdy et al. (2025)	CK
Moosavi Monazzah et al. (2025)	CK
AbuHaija et al. (2025)	CCR
Völker et al. (2025)	CT
Conia et al. (2025)	CT
Park et al. (2025c)	VCA
Srirag et al. (2025)	SC
Sahoo et al. (2024)	CSHR
Özbal et al. (2016)	VCA, SC