Survey

Survey of Cultural Awareness in Language Models: Text and Beyond

Siddhesh Pawar[1*], Junyeong Park[2*], Jiho Jin[2], Arnav Arora[1], Junho Myung[2], Srishti Yadav[1], Faiz Ghifari Haznitrama[2], Inhwa Song[2], Alice Oh[2], and Isabelle Augenstein[1]

[1]University of Copenhagen, Denmark
sipa@di.ku.dk

[2]KAIST, Republic of Korea
jjjunyeong9986@kaist.ac.kr

Large-scale deployment of large language models (LLMs) in various applications, such as chatbots and virtual assistants, requires LLMs to be culturally sensitive to the user to ensure inclusivity. Culture has been widely studied in psychology and anthropology, and there has been a recent surge in research on making LLMs more culturally inclusive, going beyond multilinguality and building on findings from psychology and anthropology. In this article, we survey efforts towards incorporating cultural awareness into text-based and multimodal LLMs. We start by defining cultural awareness in LLMs, taking definitions of culture from the anthropology and psychology literature as a point of departure. We then examine methodologies adopted for creating cross-cultural datasets, strategies for cultural inclusion in downstream tasks, and methodologies that have been used for benchmarking cultural awareness in LLMs. Further, we discuss the ethical implications of cultural alignment, the role of human–computer interaction in driving cultural inclusion in LLMs, and the role of cultural alignment in driving social science research. We finally provide pointers to future research based on our findings about gaps in the literature.[1]

1. Introduction

Language models are deployed in various user-facing applications, such as recommender systems (Bao et al. 2023), customer service (Pandya and Holia 2023), and search applications (Xiong et al. 2024), which are increasingly used by people in all aspects of their lives, including education (Kasneci et al. 2023), public health (De Angelis et al.

* Equal contributions.

[1] We additionally organize the papers covered by this survey at https://github.com/siddheshih/culture-awareness-llms.git.

Action Editor: Eduardo Blanco. Submission received: 27 October 2024; revised version received: 10 March 2025; accepted for publication: 18 April 2025.

https://doi.org/10.1162/coli.a.14

2023), and professional writing (Jakesch et al. 2023). These models reflect the Western perspective, as they are predominantly trained on Western-centric data (Durmus et al. 2023). This skewed perspective can lead to stereotyping and alienation of users, propagation of stereotypes due to a lack of cultural understanding (e.g., flattening of cultural identities), or responding in a culturally insensitive way (Cao et al. 2022, 2023). Therefore, cultural awareness is one of the critical factors that should be considered when creating natural language processing (NLP) models.

In this article, we provide a comprehensive survey of the steps that the NLP community has taken to make language models more culturally inclusive. Furthermore, with advancements in multimodal foundation models and their adaption to NLP tasks (Fei et al. 2022), we also examine efforts towards cultural inclusion in multimodal NLP systems (i.e., multimodal systems with language understanding as one of their components). As the notion of culture used by the NLP community (to define and ensure cultural inclusion in NLP systems) is adopted from social science research, we start by defining "cultural awareness in LLMs" based on definitions of culture in the psychology and anthropology literature. We then consolidate the research that examines cultural inclusion in LLMs and multimodal models, including benchmark creation, training data creation, alignment methodologies, and evaluation methodologies. We also discuss the role of cultural alignment in accelerating social science research. Human–computer interaction (HCI) also plays a role in ensuring cultural alignment in LLM, as studying how different cultures react to certain levels of cultural (mis)alignment and the study of varied expectations of people with different cultural backgrounds falls in the realm of HCI research (Weidinger et al. 2023). Finally, we discuss the ethical and safety implications of current research directions and provide potential research avenues that the community could take to foster cultural inclusion in language models. While recent surveys (Liu, Gurevych, and Korhonen 2024; Adilazuarda et al. 2024) focus on the cultural alignment of LLMs in NLP and provide taxonomies for grouping current cultural alignment works, we consolidate the literature from a broader scope. We survey and compare efforts towards incorporation and conceptualization of culture in NLP systems, and our survey spans several modalities, including images, videos, and audio, along with text. We position our survey at the intersection of NLP, multimodality, and social science.

The key contributions and research goals of this survey are as follows:

1. We review 300+ papers to provide an overview of the current state of benchmarks and methods used for cultural inclusion in multimodal language models (we organize the papers in §4, §5, §6).
2. We provide an overview of common data sources used for creating cultural alignment datasets and how current benchmark creation and culturally relevant fine-tuning dataset creation methodologies leverage these common sources (§2); we also discuss ethical implications and limitations of the dataset creation methodologies (§8).
3. We provide an overview of the coverage of current datasets for geographical regions and cultures (§7) and discuss measures that the community could take to foster equity in cultural inclusion (§9).
4. We also examine the societal impact and implications of deploying LLMs with or without cultural awareness and discuss the role of HCI research in cultural alignment (§8).

Literature Collection Strategy. We start by searching papers on arXiv, Google Scholar, and ACL Anthology and then filter the papers using the PRISMA methodology (Page et al. 2021). We provide the keywords used and filtering criteria in Appendix 1. The majority of our papers are articles published in the ACL anthology and regional ACL chapters, EMNLP, ICLR, and ICML; computer vision conferences (ICCV and CVPR); and HCI venues (CHI, CSCW, Proceedings of the ACM on Human–Computer Interaction, International Journal of Human-Computer Interaction). The inclusion of cultural aspects in the NLP and CV communities is recent, with most (benchmark) papers published post-2016, so we consider cultural inclusion benchmarks after 2016. For methodology-based papers, as the scope of our survey is limited to LLMs, we consider post-2018 papers that examine cultural inclusion in pretrained LLMs (Han et al. 2021). We also consider recent submissions to arXiv to include recent NLP and social science papers, as the publication cycles for social science journals are typically 1–3 years. The scope of this paper is limited to social science papers that study cultural inclusion and cultural analysis in the context of LLMs, so we do not consider any specific social science journal. For HCI-related papers, while there has been substantive work and case studies on the deployment of LLMs across cultural groups, we focus on works on improving and evaluating cultural alignment in LLMs using HCI.

We define culture in §2 and organize our paper into three major parts. The first part discusses data sources and methodologies the community has used to create datasets and benchmarks for the cultural inclusion of LLMs (§3). The second part discusses the methodologies and state of benchmarks that have been used or created for improving cultural awareness in LLMs across modalities (§4, §5, §6). Finally, we discuss our observations: the state of cultural inclusion (§7), ethical issues related to cultural alignment, and the role of cultural alignment in accelerating social science research (§8), and future research directions (§9) in the last part. In each of the subsections in §4, §5, and §6, we identify specific research gaps and, based on the research gaps, provide concrete suggestions for future research in §9.

2. Definitions of Culture and Methodology

Culture is a complex construct and has been studied in psychology and anthropology with different considerations and assumptions. We adopt White's (1959) view of culture, as they consolidate its definitions from the psychological and anthropological perspectives, distinguishing between human behavior and the study of culture. The psychological perspective considers the study of human behavior as the central part of the analysis and sees culture as an extension of human behavior. One of the main goals of cultural psychology is to study changes in human behavior with respect to culture. The theory and methods in cultural psychology begin with the assumption that psychological processes are socioculturally grounded. On the other hand, the anthropological perspective sees culture as an abstraction of human behavior. The abstraction is necessary to discard unimportant details and focus on actual human interactions, depending on the context. While White's (1959) definition represents an early, influential Western view that tends to treat culture as a set of shared practices, Hall (1976) further expands the definition by introducing the concept of high- and low-context communication, which underscores how cultural practices are expressed differently in non-Western contexts. In many eastern and indigenous societies, communication is characterized by a high-context orientation where meaning is implicitly conveyed through shared histories, rituals, nonverbal cues, and relational dynamics rather than

solely through explicit verbal language. In contrast, where low-context communication prevails, meaning is typically conveyed through explicit direct language. In Japan, for example, haragei ("belly talk") relies on unspoken understandings shaped by decades of social reciprocity, contrasting sharply with Western contractual individualism, implying a high-context nature of Japanese culture. Hall's definition does not privilege non-Western cultures over Western; rather, it provides a continuum that accommodates both extremes. These perspectives reveal that in many non-Western contexts, culture is not merely an abstract set of ideas, but a lived epistemology, where understanding is deeply intertwined with daily communal practices, environmental stewardship, and spiritual traditions (Ito, Walker, and Liang 2014; Tomaselli and Xanthaki 2021).

Both perspectives (psychological and anthropological) are important when considering cultural awareness in LLMs. The anthropological perspective looks at understanding the context and interpreting different elements of tasks based on the context, while the psychological perspective deals with how to process the current information (task and the context) to produce a response. The Sapir–Whorf hypothesis (Whorf 2012) from linguistic anthropology provides a unifying framework by suggesting that language reflects and shapes cultural cognition. Its strong form—linguistic determinism—asserts that language fundamentally constrains perception and thought; however, this view has largely been discredited in favor of the weaker form, linguistic relativity, which posits that language subtly influences how individuals internalize and express cultural norms and values. The design of current LLMs is to mimic human behavior as closely as possible without consideration of the context (such as writing a correct summary and seeing how factually accurate it is rather than seeing how the summary would be based on the context, considering context would involve considering the knowledge level of the user, and the purpose of the summary). The context consists of social factors, which also form an essential part of the language, and culture is one of the main components of social factors; for a detailed discussion on modeling social factors of context into NLP systems, we refer the reader to Hovy and Yang (2021). The two perspectives on culture guide the factors to consider while designing culturally aware LLMs.

From an anthropological perspective, culture is actions, things, and concepts viewed in the context of other actions and things. For instance, going for a vote is just an act in itself; it gains significance when considered in the context of democracy, autocracy, and so on. Thus, the definition of culture also considers other humans' behaviors. The locus of culture (or understanding of the cultural context) consists of three dimensions: (1) "Within humans" (such as concepts, traditions, beliefs, social practices, etc.); (2) between humans "social interaction among human beings"; and (3) outside of humans but "within the patterns of social interaction" (in materialized objects such as tools, arts, etc.). Dimension 1 deals with actual actions, things, and concepts (cultural knowledge and morals), while Dimensions 2 and 3 consider the context of the actions and concepts. The human dimension (Dimension 1) forms the basis for understanding the cultural elements of the task; the other two dimensions are important for generating relevant answers and when LLMs are used as agents. White (1959) groups together concepts and actions that form an identity of a culture into a broad category called "elements of culture." These elements of culture have been studied in NLP under textual information tasks that are concerned with cultural commonsense knowledge, norms, values, morals, linguistic forms, and so forth, as well as visio-linguistic parts, such as concepts (and perceptions) associated with various (physical) objects and art forms. Research on cultural psychology has shown that various aspects of visual perception, such as the perception of length, geometrical intuition, and depth, vary across people from different cultural backgrounds (Segall, Campbell, and Herskovits 1967; Jahoda and McGurk 1974).

The variance of perception across cultures, in turn, affects how differences in cultural backgrounds affect the way that individuals attend to, understand, and talk about visual content (Nisbett and Masuda 2003). The variance of perspective necessitates cultural adaptation of images and captions generated by the models. The variance of perspective across cultures has been studied in psychology across five major categories: architecture, clothing, dance and music, food and drink, and religion (Halpern 1955). As the elements of culture and their perceptions in context vary vastly among cultures, it becomes necessary to study and model the variance in these elements across cultures to create culturally inclusive language technologies.

Concerning LLMs and NLP systems in general, cultural awareness can be thought of as the ability to understand the context in which they are asked to perform a particular task and how the context (and elements of culture in the context) varies with culture (cultural competence). Cultural awareness also includes the ability to understand the variance of cultural elements across different cultures. Understating the cultural context consists of two things: (a) Recognizing the social context in which a task is performed and (b) based on the context, interpreting different elements of the task. Some tasks are context-sensitive, such as hate-speech detection (the definition of hate-speech varies from culture to culture as norms vary), while others are not (e.g., mathematical reasoning). Some recent work (e.g., AlKhamissi et al. 2024) has defined cultural alignment with respect to the model's views aligning with a group of people representing a culture; our definition of cultural awareness encompasses the definition of cultural alignment. So, when designing models for a task (understanding the input and producing an output), the LLM first needs to understand the context (where it is deployed, what is the end goal, etc.) and then decide if culture needs to be considered for that particular task. To understand the concepts, the LLMs should broadly consider the relationship (for example, how to converse when writing an application letter as a student), social context, and the "containers" of communications (such as the setup in which the LLM is deployed, the goal of the LLM) and demographics (Liu, Gurevych, and Korhonen 2024). The context can also be a design choice while creating or fine-tuning LLMs based on deployment goals (e.g., what data to collect and how the LLM will be used). For generalized LLMs, there is a need to have the capability to understand the context in LLMs. Most efforts up until now have focused on creating context-specific datasets and benchmarks, and less effort has been focused on building and testing LLMs that automatically detect the context.

Once the LLMs have recognized the context, the understanding of elements of the task and the response depends on cultural knowledge (e.g., the task: "generate a story with causal conversations happening in Korea" depends on elements such as LLMs' knowledge about the nature of causal conversations in Korea). We broadly use the term "elements" to include values, meaning of artifacts, and pragmatically motivated features relevant to the task. This raises the question: How do we enable LLMs to generate culturally appropriate responses and understand the cultural elements of a task? Explicitly modeling cultural knowledge in LLMs has been explored as one potential approach. Cultural knowledge consists of various aspects such as norms, morals, values, common sense knowledge, linguistic forms, artifacts, concepts, and meanings associated with artifacts, and so on.

In this survey, we discuss the multiple methods the community has explored to add and evaluate the cultural knowledge in LLMs. The major body of literature focuses on adding cultural knowledge to training/fine-tuning/alignment data and evaluation benchmarks. The data sources for cultural knowledge include direct sources such as sociological surveys, or indirect sources, which include task-specific datasets in which

cultural knowledge is implicitly but deliberately added. We discuss the creation of these sources in §2 and summarize task and usage-specific details in §4 for text-only datasets and in §5 and §6 for datasets for other modalities. Cross-cultural modeling remains underexplored in NLP research, as most studies in NLP examine cultural elements in isolation rather than analyzing their similarities and differences (Hershcovich et al. 2022). The study of cross-lingual similarities and differences has been central to cultural research in anthropology (Ember 2009). Modeling cross-cultural differences becomes a critical aspect to consider while building multicultural datasets, as there is a risk of flattening identities and erasing cultural boundaries if detailed culture-specific data is unavailable (this generally happens for under-represented cultures). Given the progress in creating generalized models, careful consideration should be given to monocultures and subcultures within a culture.

3. Data Creation Methodology

In this section, we examine the data source and creation methodology for culture-specific datasets and benchmarks. The dataset creation methodologies are organized into automatic pipelines (§3.1), semi-automatic pipelines (§3.2), and manual creation (§3.3). Example benchmarks and datasets organized by data resource and dataset creation methodology are listed in Figure 1.

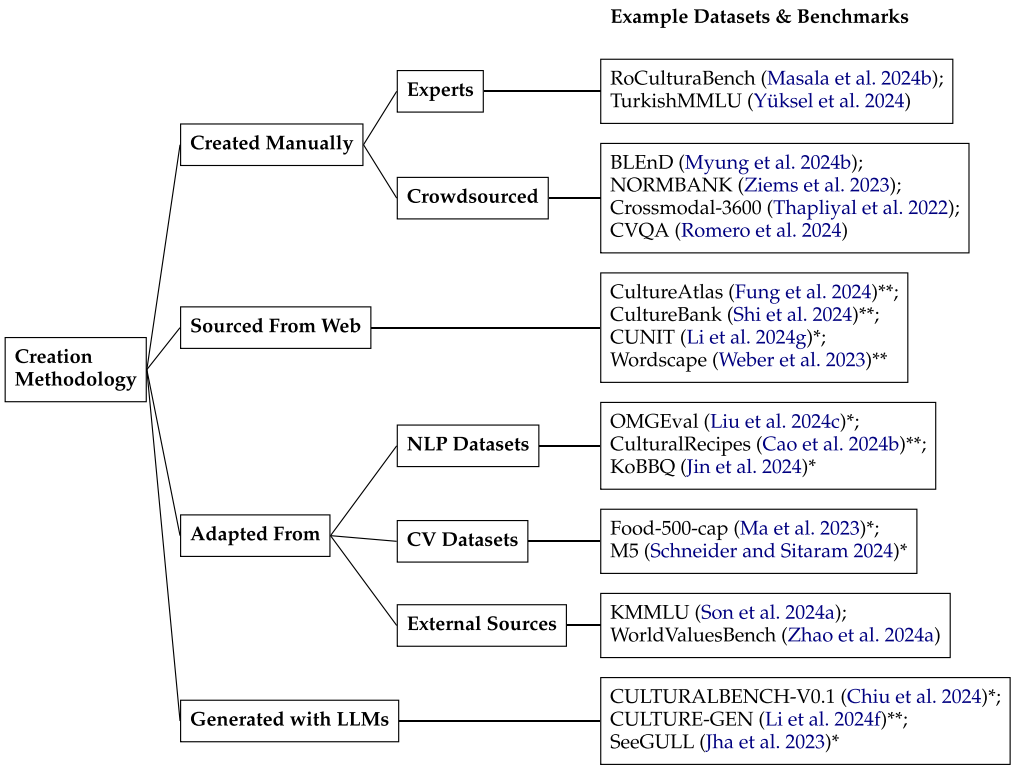


Figure 1
Overview of the data creation methodologies and example datasets and benchmarks. Datasets and benchmarks created using semi-automatic and fully automatic pipelines are marked with * and **, respectively.

3.1 Automatic Pipelines & Model-in-the-Loop

Most research focuses on automatic curation to gather cultural knowledge and create training data at scale, especially for pretraining. It relies primarily on publicly available multilingual large-scale corpora such as Wikipedia, CC100 (Conneau et al. 2020), mC4 (Xue et al. 2021), and CulturaX (Nguyen et al. 2024), which are processed raw Web text corpora gathered from public Web archives. These sources are then cleansed and filtered for specific or even multiple cultures as listed below.

- Korean (Yoo et al. 2024)
- Irish (Tran, O'Sullivan, and Nguyen 2024)
- Portuguese (Pires et al. 2023; Almeida et al. 2024)
- Arabic (Sengupta et al. 2023; Huang et al. 2024; Aloui et al. 2024)
- Chinese (Du et al. 2024)
- Taiwanese (Lin and Chen 2023)
- Persian (Abbasi et al. 2023)
- Thai (Pipatanakul et al. 2023)
- Romanian (Masala et al. 2024)
- Basque (Etxaniz et al. 2024)
- Ukrainian (Kiulian et al. 2024)
- Ethiopian (Tonja et al. 2024)
- Indonesian (Owen et al. 2024; Cahyawijaya et al. 2024b)
- Multiple cultures (ImaniGooghari et al. 2023; Nguyen et al. 2023b; Üstün et al. 2024)

The refined data are then used to train culture-specific LLMs that are tailored to the knowledge of these cultures.

Research on cultural knowledge acquisition has progressed through various model-in-the-loop techniques to enhance data quality and quantity. Early approaches focus on social media mining, as demonstrated by StereoKG (Deshpande et al. 2022), which extracts cultural knowledge from Reddit and Twitter using OpenIE (Mausam 2016). This was followed by more sophisticated pipelines like CANDLE (Nguyen et al. 2023a), which uses fine-tuned models to filter and classify cultural text corpora. Recent studies have leveraged LLMs, with MANGO (Nguyen, Razniewski, and Weikum 2024) generating culture-specific knowledge using seed data from CANDLE or ConceptNet (Speer, Chin, and Havasi 2017), and CultureAtlas (Fung et al. 2024) synthesizing cultural knowledge frames from Wikipedia-based sources.

Automation of instruction data creation has also seen significant advancement through LLM-based approaches. LLM_ADAPT (Putri et al. 2024) leverages LLMs to culturally adapt CommonsenseQA (Talmor et al. 2019) for Indonesian and Sundanese contexts. CultureBank (Shi et al. 2024) and CRAFT (Wang et al. 2024a) utilize LLMs differently: The former extracts cultural descriptions from social media, while the latter generates questions based on cultural keywords and filtered corpus chunks. Survey-based methods have also emerged, with CultureLLM (Li et al. 2024b) using LLMs

to augment the World Values Survey (WVS) data (Survey 2022) through semantic equivalence and synonym replacement. CulturePark (Li et al. 2024c) advances this approach with a multi-agent framework for cross-cultural conversations, using both the Pew Global Attitudes Survey (Center 2022) and WVS as initialization data. For cross-lingual applications, X-Instruction (Li et al. 2024d) creates instruction data through a three-step pipeline that leverages high-resource languages, using the OpenAssistant Conversations corpus (Köpf et al. 2023) and CulturaX (Nguyen et al. 2024). These developments demonstrate the effective combination of LLM-generated synthetic data with established cultural resources.

Concerning computer vision models, automatic pipelines for data creation involve using geo-localized image datasets such as datasets from photo-sharing platforms such as Flikr, Google Photos, Pinterest (Kuznetsova et al. 2020), Wikimedia, Youtube, etc. We categorize these sources as extracted using automatic pipelines because the geolocalized tags are already present when the dataset is being considered for cultural adaptation. The captions for the images are obtained from meta-data or using the associated text on Wikipedia (Srinivasan et al. 2021; Weber et al. 2023). The community has mostly used automatic pipelines to get the images, as captions obtained from metadata or from Wikipedia may not be reliable for creating benchmarks or datasets for cultural adaptation.

3.2 Semi-automatic: Human-in-the-Loop

Semi-automatic approaches combine human expertise with machine processing to achieve higher-quality datasets. Recent studies demonstrate this approach through various methodologies. Putri et al.'s (2024) LLM_GEN dataset relies on human annotators to create initial categories and concepts before generating commonsense question-answering data with LLMs. COIG-CQIA (Bai et al. 2024) uses human experts to curate Chinese instruction data sources before applying machine-based filtering. For Arabic instruction tuning, CIDAR (Alyafeai et al. 2024) translates the AlpaGasus dataset (Chen et al. 2024b) using ChatGPT (OpenAI), followed by cultural localization and linguistic review by native speakers. The STREAM framework (Wang et al. 2024f) demonstrates extensive human involvement: Annotators provide moral values to guide LLM scenario generation, manually screen the outputs, and collaborate with LLMs to evaluate selected scenarios for measuring human-LLM alignment. These approaches effectively balance quality and scalability by combining human expertise with machine processing capabilities.

For vision-language models, semi-automatic and human-in-the-loop approaches have been the most common way of creating benchmarks and datasets for cultural adaptation. One of the methods includes using geolocalized images from photo-sharing platforms such as Flickr, Google Photos, Pinterest (Kuznetsova et al. 2020), Wikimedia, and Youtube, and then using local annotators to create a variety of datasets (Thapliyal et al. 2022; Yin et al. 2021). Some studies also use pre-existing computer vision benchmarks, such as ISIA Food-500 (Min et al. 2020), Dollar Street (Dubois et al. 2023), and so on, and refine the captions (or related information such as questions in Visual Question Answering (VQA)) to include cultural information (Ma et al. 2023; Schneider and Sitaram 2024).

3.3 Manual: Handcrafted from Scratch

Manually crafted datasets, created by human experts and annotators, remain the gold standard for quality and human value alignment in NLP. While these datasets ensure

high-quality outputs aligned with human expectations and linguistic nuances, they face scalability challenges due to the significant time and financial investments required. Recent works demonstrate the diverse applications of manual dataset creation. Putri et al.'s (2024) HUMAN_GEN dataset was developed entirely from scratch, with annotators handling both concept creation and question-answering data construction. For capturing undocumented cultural knowledge, Myung et al. (2024) recruited native annotators through crowdsourcing platforms to document everyday mundane knowledge, while directly recruiting North Korean participants due to limited annotator availability. In the domain of dialectal studies, Le and Luu (2023) worked with native speakers to create a parallel corpus for Central and Northern Vietnamese dialects. Manual annotation has proven particularly valuable for subjective tasks, such as cultural-specific hate speech detection (Jeong et al. 2022) and inspiring content detection (Ignat, Lakshmy, and Mihalcea 2024).

Recent language models demonstrate the effectiveness of handcrafted datasets in enhancing cultural relevance and performance. For Korean, HyperClovaX (Yoo et al. 2024) incorporates human-annotated datasets for instruction tuning post-pretraining. For Chinese cultural knowledge, TaiwanLLM (Lin and Chen 2023) leverages human instructions and feedback from real user interactions. Arabic language models like AceGPT (Huang et al. 2024) and Jais (Sengupta et al. 2023) enhance their supervised fine-tuning with native Arabic instructions. The impact of handcrafted datasets is particularly significant for low-resource languages. Training Komodo (Owen et al. 2024) involved a collaboration with local language experts to collect data for various local languages, while Aya (Üstün et al. 2024) expands this approach to cover 101 languages. Despite these advances, a systematic approach to developing high-quality datasets for underrepresented languages remains a critical research gap, highlighting the need for broader dataset creation efforts across diverse linguistic and cultural contexts.

For vision-language data, manual data collection methods include starting with an initial list of questions and concepts and asking the annotators to search relevant images on the Internet (Baek et al. 2024; Wang et al. 2024g) followed by processing images to create various types of benchmarks (captioning, VQA, image retrieval, etc.). Native annotators can also drive the selection of cultural topics and objects to prioritize objects and concepts with significant cross-cultural differences (Wang et al. 2024g). Most video and speech datasets are also created manually, where annotators code emotions of curated videos or Web series (Amiriparian et al. 2024b; Zhao et al. 2022). There have been a few examples in the literature where annotators are asked to click on the relevant photos based on initial concepts (Romero et al. 2024). Searching for culture-related photos can be limited since there is a bias towards primarily uploading aesthetically pleasing images to the Internet, leading to a lack of photos with everyday objects and common sense knowledge. Also, the difference in online presence between cultures can lead to the discrimination of certain cultures (Liu et al. 2022d). In the upcoming sections, we discuss how data-creation methodologies introduced in this section have been used for creating task-specific data and cultural alignment of language and vision models.

4. Language Models and Culture

There has been a growing recognition of the cultural biases, stereotypes, and lack of diverse cultural knowledge present in LLMs (Hershcovich et al. 2022; Navigli, Conia, and Ross 2023). Those issues directly lead to problems, particularly in applications like dialogue systems, where LLMs may overlook users' cultural backgrounds, potentially leading to inaccurate information or the reinforcement of cultural stereotypes.

To address these limitations and make LLMs more culturally inclusive, two key approaches have emerged: (a) pretraining and fine-tuning models with culturally relevant data; and (b) prompt-based methods that do not require retraining. Section 4.1 provides a detailed explanation of these methods, focusing on how they aim to enhance LLMs' cultural adaptability. Furthermore, there is increasing attention towards developing benchmarks and evaluation frameworks that measure how well LLMs align with diverse cultural contexts. Section 4.2 elaborates on these benchmarks and evaluation frameworks. However, both alignment methodologies and evaluation techniques remain fragmented, with no universally established standards.

4.1 Cultural Alignment: Methodologies and Goals

Cultural alignment refers to the process of aligning an AI system with the set of shared beliefs, values, and norms of the group of users that interact with the system, as defined by Masoud et al. (2023) based on the foundational works of Hofstede, Hofstede, and Minkov (2010) and Bennett III, Fadil, and Greenwood (1994). The importance of cultural alignment is demonstrated by Masoud et al. (2023) through their Cultural Alignment Test (Hofstede's CAT), which reveals that current LLMs struggle to fully comprehend cultural values. Their research suggests that this limitation could be addressed through fine-tuning models with culture-specific language. Complementary studies by Li et al. (2024f) and Tao et al. (2024) reach similar conclusions, though their findings emphasize the effectiveness of prompting techniques for achieving cultural alignment. Given the importance of these findings, we examine the current state of cultural alignment in AI systems through two distinct perspectives. First, we analyze the methodologies for achieving cultural alignment, focusing on two primary approaches: model training and prompting techniques. Second, we explore how different cultural objectives influence and shape alignment efforts.

In this section, we discuss the methodologies used for cultural alignment. In general, cultural alignment can be done through two approaches: training-based (§4.1.1) and training-free (§4.1.2). In addition, there are goal-based alignment methods for specific goals, such as content moderation (§4.1.3). The studies are organized in Figure 2.

4.1.1 Training-based Methods. Training a language model is one way to achieve cultural alignment. The key differentiating factor in this approach lies in the training data, which must contain culturally relevant knowledge, norms, and values specific to the target culture, as previously discussed in §2. Training methodologies for cultural alignment can be broadly categorized into two main approaches: pretraining and fine-tuning. Pretraining is the step where the model is trained on a large corpus to learn the general features of the data, which, for the purpose of cultural alignment, includes culture-specific knowledge, norms, and values obtained inside. Pretraining can be further categorized into two strategies: initiating pretraining from scratch using culturally relevant data, or continuing from an existing pretrained LLM. Pretraining the model from scratch is expensive, as a result of training the whole model parameters and the large size of the data. Therefore, not a lot of cultural alignment is done with this method, as only HyperClovaX (Yoo et al. 2024), PersianLLaMA (Abbasi et al. 2023), and JASMINE (Billah Nagoudi et al. 2023) are pretrained from scratch.

Continued pretraining is another way of pretraining, which involves taking an existing pretrained model and training it further on culturally relevant data. This method avoids the computational expense of pretraining from scratch, and it requires only raw text data rather than the labeled datasets needed for supervised fine-tuning.

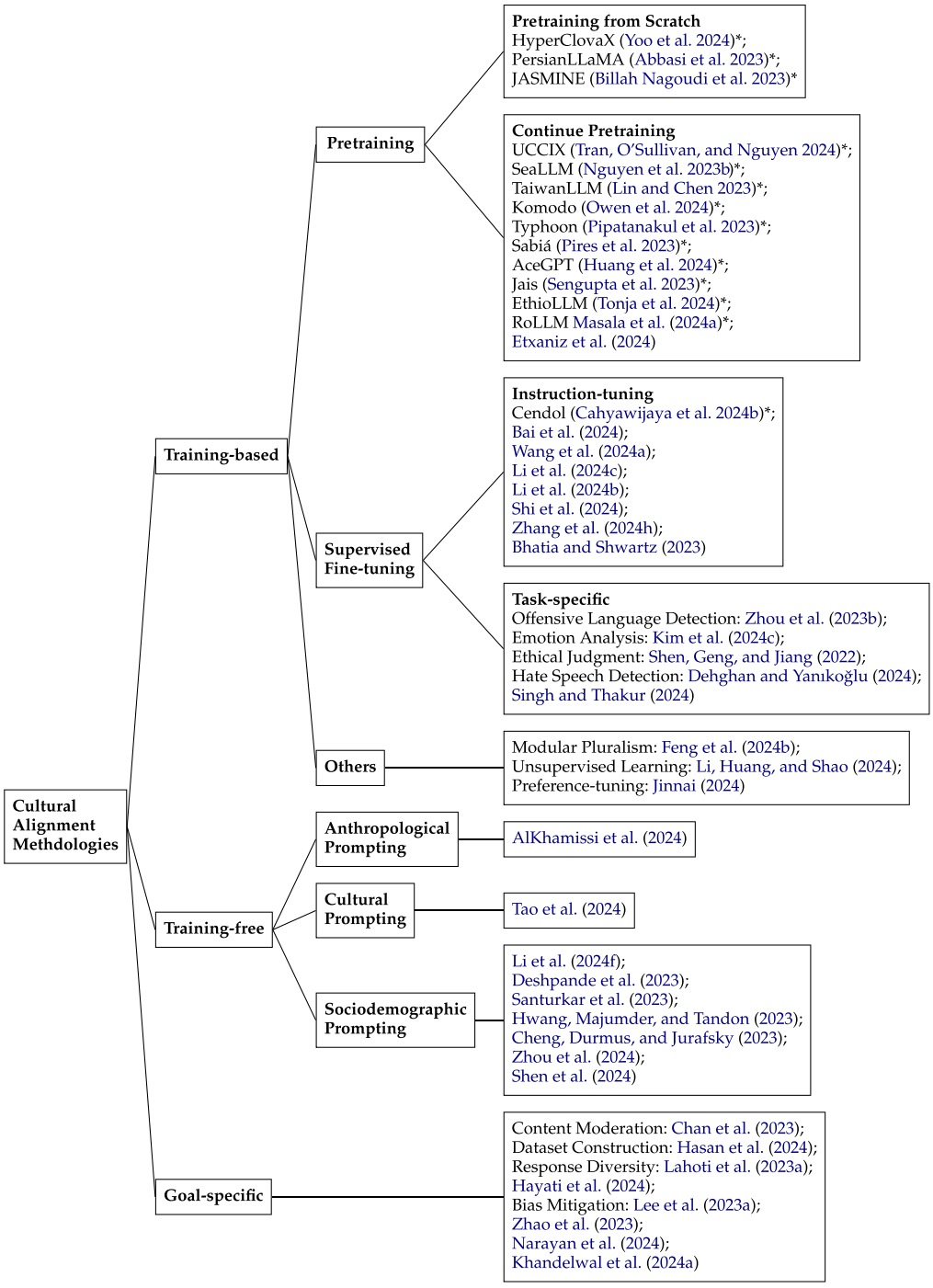


Figure 2
Cultural alignment methodologies for language models categorized by methodologies and goals. Model names are marked with *.

There are many papers which incorporate continued pretraining as part of their cultural alignment efforts as shown by the large number of culture-specific LLMs (Tran, O'Sullivan, and Nguyen 2024; Nguyen et al. 2023b; Lin and Chen 2023; Owen et al. 2024; Pipatanakul et al. 2023; Pires et al. 2023; Huang et al. 2024; Sengupta et al. 2023; Tonja et al. 2024; Masala et al. 2024; Etxaniz et al. 2024) using continued pretraining.

Fine-tuning for cultural alignment involves further training a pretrained model on culturally relevant labeled datasets. Unlike continued pretraining, which uses raw text data, fine-tuning utilizes data specifically labeled for the intended task. This approach can be applied to task-specific objectives such as hate-speech detection and emotion classification, or to general-purpose applications such as instruction-following and conversational abilities. Instruction-tuning, a specific form of fine-tuning that uses instruction-response pairs, has been widely adopted by researchers developing culture-specific LLMs (Yoo et al. 2024; Lin and Chen 2023; Owen et al. 2024; Huang et al. 2024; Cahyawijaya et al. 2024b; Sengupta et al. 2023; Masala et al. 2024; Nguyen et al. 2023b; Bai et al. 2024). Another study that leverages instruction-tuning is that of Zhang et al. (2024h), who propose a rapid adaptation method for large models in specific cultural contexts based on specific cultural knowledge and safety values data. Recent research (Li et al. 2024c,b; Wang et al. 2024a; Shi et al. 2024) demonstrates that instruction-tuning enables models to effectively reason across multiple cultures in conversations. Additionally, Bhatia and Shwartz (2023) show that fine-tuning on CANDLE data (Nguyen et al. 2023a) allows models to both capture and generate culturally nuanced commonsense knowledge.

Many studies also show a positive impact on various cultural alignment applications by using task-specific fine-tuning. For offensive language detection, Zhou et al. (2023b) achieve effective results by fine-tuning models on cultural value survey data. In hate speech detection, which is deeply intertwined with cultural contexts, researchers have utilized diverse approaches. Dehghan and Yanıkoğlu (2024) apply dual-contrastive learning when fine-tuning and incorporating paralinguistic features such as emoji, while Singh and Thakur (2024) use a federated approach that utilizes continuous adaptation and fine-tuning to detect hate speech which is highly affected by cultural nuances. For emotion analysis, Kim et al. (2024c) achieve promising results for moral emotions classification, where the model utilizes information on moral emotions embedded in the data and can perceive different emotions for different cultures. For ethical judgment, Shen, Geng, and Jiang (2022) study grounding complex narrative situations with social norms using a pretrained encoder-decoder and integrating these norms with a classification model.

Beyond pretraining and fine-tuning, some innovative approaches offer unique perspectives on training-based alignment. One such approach is Modular Pluralism (Feng et al. 2024b), which uses smaller language models alongside larger ones to guide them in incorporating cultural knowledge and values into their responses according to the given cultural context. From an unsupervised perspective, A and Augenstein (2020) use subword segmentation, language models, as well as a method for mapping between subword sequences to convert between traditional and simplified Chinese characters, which is an important aspect of understanding Chinese culture. Li, Huang, and Shao (2024) build on this and utilize an adaptive context-aware unsupervised learning framework. There is also a study by Jinnai (2024) that uses preference-tuning through Direct Preference Optimization (Rafailov et al. 2023) instead of fine-tuning to investigate how cross-cultural alignment affects an LLM's commonsense morality.

Several studies have investigated the actual impact of training for cultural alignment. Mukherjee et al. (2024a) found that while there are improvements in terms of

cultural competence, they still fall short, particularly in non-Western contexts. They highlight the need to incorporate more than the target language during the training process. Ladhak et al. (2023) find that while pretraining can make the model aligned with the specific culture, the resulting model can possess bias contained in the pretraining data. They also find that fine-tuning with smaller parameters, such as adapter-fine-tuning techniques like LoRA, provides better generalization and debiasing rather than training the entire model. Choenni, Lauscher, and Shutova (2024) investigate how cultural value shifts during fine-tuning, and find that language has a minor role in cultural shifts and positively affects alignment with human values, but it varies considerably across languages.

4.1.2 Training-free Methods. Cultural alignment in language models can be achieved without additional training, primarily through prompting techniques. Research by AlKhamissi et al. (2024), Zhou et al. (2024), and Arora, Kaffee, and Augenstein (2023) demonstrates that cultural alignment is influenced by the training data and the prompts used during inference. AlKhamissi et al. (2024) observed that models exhibit stronger cultural alignment when prompted in a culture-specific language. Building on this insight, they introduced anthropological prompting, incorporating anthropological reasoning aspects into the prompt to enhance cultural alignment. Another promising approach is the Collective, Critique, and Self-Voting method, which is proposed by Lahoti et al. (2023). Their findings suggest that language models can comprehend the concept of diversity and are capable of reasoning about and critiquing their responses to improve cultural diversity in their outputs. Tao et al. (2024) also propose a prompt methodology called cultural prompting, which instructs the language model to answer like a person from another society. They investigated the method by comparing the model responses to nationally representative survey data and found cultural prompting works quite well to increase the alignment of the model with the nationally representative survey data. Moreover, Pawar et al. (2025) show that LLMs pick up on implicit cultural information via user names, which can lead to an inadvertent reinforcement of cultural biases.

Sociodemographic prompting has gained widespread attention among researchers for cultural alignment (Deshpande et al. 2023; Santurkar et al. 2023; Hwang, Majumder, and Tandon 2023; Cheng, Durmus, and Jurafsky 2023; Zhou et al. 2024; Li et al. 2024f; Shen et al. 2024; Kwok, Bravansky, and Griffin 2024). This method involves enriching prompts with sociodemographic information or cultural context, expecting that the model's output will align with the information given in the prompt. Sociodemographic prompting has shown potential in applications such as data augmentation (Hartvigsen et al. 2022), social computing simulation (Park et al. 2022), and mitigating hallucinations (Feng et al. 2024a). An interesting approach in relation to sociodemographic prompting is multilingual feedback (Feng et al. 2024a). In that study, the authors provide additional context by probing LLM to self-reflect and provide multilingual feedback on the initial answer as reasoning for a new answer that is better aligned across diverse languages, cultures, and communities.

However, concerns have been raised regarding the robustness of sociodemographic prompting. Mukherjee et al. (2024b) found that most models exhibit similar variations in response to culturally conditioned cues as they do to non-cultural ones, particularly in terms of eliciting cultural bias. Similarly, Beck et al. (2024) observed that model outcomes vary significantly across different model types, sizes, and datasets. These findings suggest that sociodemographic prompting should be employed cautiously, especially in sensitive applications.

4.1.3 Goal-specific Alignment Strategies. In this section, we discuss research on testing and improving cultural alignment in language models for particular goals.

Chan et al. (2023) train large language models on extensive datasets of media news and articles to create culturally attuned models for content moderation; the goal is to capture the nuances of communication and offensive content across cultures. Lahoti et al. (2023) propose metrics to measure diversity in LLM-generated responses along the people and culture axes and propose a new prompting technique to self-improve the diversity of responses in LLMs to represent diverse opinions. Along similar lines, Hayati et al. (2024) propose a step-by-step recall prompting-based method to increase the diversity of responses (with cultural diversity increase being one of the outcomes). Lee et al. (2023) provide a (fine-tuning) dataset specific to Korean culture for mitigating social bias in generated content. Zhou et al. (2023b) study the importance of cultural features in determining the success of transfer learning in the case of offensive language detection. Khandelwal et al. (2024) provide a dataset of Indian stereotypes and anti-stereotypes and propose interventions to reduce both stereotypical and anti-stereotypical biases in language models, thereby aligning them with Indian culture. Narayan et al. (2024) propose a framework to quantify and mitigate biases within LLMs by creating a new metric which detects, measures, and mitigates racial and cultural biases in LLMs without reliance on demographic annotations. Hasan et al. (2024) propose a language-independent framework to construct culturally and regionally aligned question answering (QA) datasets in native languages for LLM evaluation and demonstrate the efficacy of the framework by designing a multilingual natural QA dataset, MultiNativQA, consisting of around 64k manually annotated QA pairs in 7 languages, ranging from high to extremely low resources, based on queries from native speakers from 9 regions covering 18 topics. Zhao et al. (2023) introduce CHBias, a dataset for bias evaluation and mitigation of Chinese conversational language models with culture-specific biases.

Takeaways from §4.1. Various alignment methods, such as continued pretraining, fine-tuning, instruction tuning, and prompt tuning, are used to create more culturally aware LLMs. However, most efforts are concentrated on aligning LLMs with individual local cultures, with limited research dedicated to developing cross-cultural LLMs that encompass comprehensive knowledge of multiple cultures. More work is needed in this area to enhance the cross-cultural understanding of LLMs.

4.2 Benchmarks and Evaluation

In this section, we provide an overview of various benchmarks designed to assess cultural elements through text-based tasks. The cultural elements are categorized into eight domains as specified in Figure 3. The papers were categorized using a bottom–up

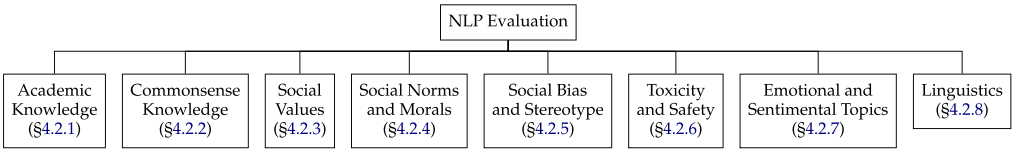


Figure 3
An overview of domains of text-based culturally aware benchmarks.

clustering approach, in which the authors manually annotated each paper's category based on its task and then clustered the papers according to similar tasks.

The Academic Knowledge section (§4.2.1) focuses on evaluating knowledge sourced from human educational materials. The Commonsense Knowledge section (§4.2.2) covers diverse datasets and benchmarks that assess general cultural knowledge, such as food, family, holidays, sports, and entertainment. In the Social Values section (§4.2.3), social science studies are used to evaluate LLMs' alignment with human social values. The Social Norms and Morals section (§4.2.4) examines specific cultural norms and morals, exploring how these values shift depending on the social context. In the Social Bias and Stereotypes section (§4.2.5), the focus is on adapting bias benchmarks to local languages and cultures, expanding to cross-cultural perspectives. The Toxicity and Safety section (§4.2.6) addresses offensive and hate speech detection in local languages and cultures. The Emotional and Sentimental Topics section (§4.2.7) explores psychological cultural difference including emotion prediction and sentiment analysis. Lastly, the Linguistics section (§4.2.8) delves into how culture is reflected in language, the ways language varieties and literary forms embody cultural elements, and how translation and dialogue systems can become more culturally aware.

Each cultural element is evaluated through element-specific approaches. For instance, commonsense knowledge is typically assessed using multiple-choice questions (MCQ) or short-answer questions that require cultural knowledge. Meanwhile, social values are often examined using sociological surveys like the WVS to test for cross-cultural differences in LLMs' understanding of social values.

4.2.1 Academic Knowledge. Human educational resources such as exam questions or textbooks are being utilized to assess language understanding and general knowledge capability of LLMs. For instance, the Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al. 2021) is sourced from practice exam questions, such as Graduate Record Examination (GRE). This dataset is commonly used to evaluate LLMs' language understanding and problem-solving abilities across various domains, including science, technology, engineering, and mathematics (STEM), humanities, and social science. Among these academic domains, fields such as history, law, and literature in particular, often require knowledge specific to a certain region. Thus, benchmarks have been developed from local educational materials to evaluate regional knowledge. The overall hierarchy of the papers in this section is specified in Figure 4.

One of the shortcomings of the MMLU dataset is that it primarily focuses on knowledge related to the United States. Addressing this, the dataset has been adapted to create several linguistically and culturally specific benchmarks, including ArabicMMLU (Koto et al. 2024a), CMMLU (Li et al. 2024e), IndoMMLU (Koto et al. 2023), JapaneseMMLU (JMMLU) (Yin et al. 2024), KMMLU (Son et al. 2024b), TMMLU (Hsu et al. 2023),

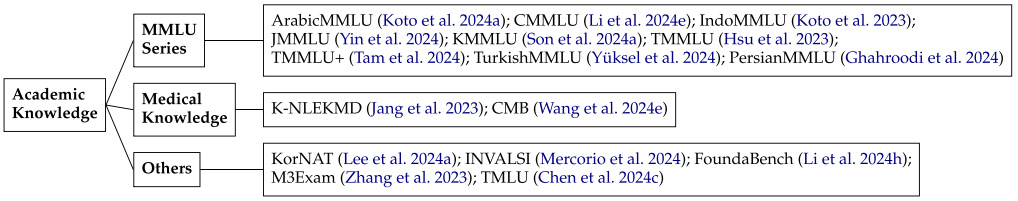


Figure 4
Academic knowledge evaluation benchmarks.

Table 1
Details of the MMLU series benchmarks.

	Languages	Evaluation method	Creation method	Educational stages	Domains	Size(k)	Cultural question ratio(%)
MMLU (Hendrycks et al. 2021)	English	MCQ	Manually created	Elementary High school College Professional	Humanities Social Science STEM	15.9	–
ArabicMMLU (Koto et al. 2024a)	Modern Standard Arabic	MCQ	Manually created	Primary school Middle school High school University Professional	Humanities Social Science STEM Language	14.5	57.7
CMMLU (Li et al. 2024f)	Mandarin Chinese	MCQ	Manually created	Primary school Middle/high school College Professional	Humanities Social Science STEM	11.5	∼25.3
IndoMMLU (Koto et al. 2023)	Indonesian and local languages	MCQ	Manually created by experts	Primary school Junior high school Senior high school University	Humanities Social Science STEM Indonesian Language Local Languages Local Cultures	14.9	46
JMMLU (Yin et al. 2024)	Japanese	MCQ	Adapted from MMLU and manually created by experts	Elementary High school College Professional	Humanities Social Science STEM	7.5	–
KMMLU (Son et al. 2024a)	Korean	MCQ	Automatically extracted	Expert	Humanities Social Science STEM Applied Science	35	20.4
TMMLU (Hsu et al. 2023)	Traditional Chinese	MCQ	Manually created	Elementary to Professional	Vocational to Academic Fields	3.3	–
TMMLU+ (Tam et al. 2024)	Traditional Chinese	MCQ	Adapted from TMMLU, TTQA, and manually created	Primary, Secondary, Undergraduate, Professional	Humanities Social Science STEM	22.7	–
TurkishMMLU (Yüksel et al. 2024)	Turkish	MCQ	Manually created by experts	High school	Humanities Social Science Math Natural Sciences Language	10	–
PersianMMLU (Ghahroodi et al. 2024)	Persian	MCQ	Automatically extracted	Lower primary school, Upper primary school, Lower secondary school, Upper secondary school	Humanities Social Science Natural Science Mathematics	20	–

TMMLU+ (Tam et al. 2024), TurkishMMLU (Yüksel et al. 2024), and PersianMMLU (Ghahroodi et al. 2024). In particular, IndoMMLU also includes nine local cultures and eight local languages in Indonesia, and ArabicMMLU is sourced from eight different countries in North Africa, the Levant, and the Gulf. The details about MMLU-series benchmarks are shown in Table 1.

All benchmarks use the MCQ format, although the number of candidate answers vary. Most benchmarks are built based on local exam questions and educational materials, with the exception of JMMLU (Yin et al. 2024) and TMMLU+ (Tam et al. 2024). In more detail, JMMLU is partially composed of translated questions from the MMLU dataset (Hendrycks et al. 2021) and TMMLU+ includes subjects from TMMLU (Hsu et al. 2023) and TTQA (Ennen et al. 2023). Although the benchmarks encompass a diverse range of knowledge from K–12 education to professional and even industrial knowledge, each benchmark is split into different educational stages because each country has different educational curricula. KMMLU (Son et al. 2024b) and TurkishMMLU are specialized for expert-level and high school-level questions, respectively.

Beyond the MMLU-series benchmarks, KorNAT (Lee et al. 2024a), INVALSI (Mercorio et al. 2024), FoundaBench (Li et al. 2024h), and TMLU (Chen et al. 2024c) are created to test educational knowledge in South Korea, Italy, China, and Taiwan, respectively. KorNAT (Lee et al. 2024a) includes social value and common knowledge datasets. Specifically, the common knowledge dataset is developed based on the

national compulsory education curriculum, covering seven subjects from the Korean GED syllabus. All questions are manually created by rephrasing the reference materials to MCQ format questions. Similarly, the INVALSI benchmark (Mercorio et al. 2024) is structured based on the INVALSI test, a popular educational assessment criteria across Italy. The INVALSI test includes various domains, including mathematics, but it especially focuses on assessing a student's linguistic proficiency through various tasks. It consists of both the MCQ and multiple complex choice questions format. Half the questions in FoundaBench (Li et al. 2024i) evaluate Chinese K–12 subject knowledge. K–12 education in China refers to compulsory primary and secondary education in China. Other than collecting questions from Chinese academic exams, they also automatically generate questions with GPT-4 (OpenAI 2023). For automatic generation, they first extract key contents from collected documents, then manually formulate and refine optimal prompts through several iterations. Traditional Chinese, also known as Traditional Mandarin, is predominantly used in Taiwan, Hong Kong, and Macao but remains underrepresented compared with Simplified Chinese, which is the standard in Mainland China. Thus, TMLU (Chen et al. 2024c) aims to evaluate knowledge and reasoning capability of LLMs in the context of Taiwanese Mandarin across 37 subjects including social science, STEM, humanities, and Taiwan-specific content. M3Exam (Zhang et al. 2023) specifically focuses on evaluating LLMs in a multilingual context. It includes nine languages from high-resource languages like English and Chinese, to extremely low-resource languages such as Javanese, each reflecting a distinct cultural background. They recruit native speakers from each region to manually collect official graduation exams in primary, middle, and high school.

Within the medical domain, CMB (Wang et al. 2024e) and K-NLEKMD (Jang et al. 2023) evaluate region-specific medical knowledge in Korea and China. Medical knowledge is often shaped by regional factors such as climate, diet, and ethnicity, leading to unique medical systems in each country. CMB (Wang et al. 2024e) is a comprehensive medical benchmark in Chinese that covers six categories of medical knowledge, including the Physician, Nurse, and Pharmacist domains. The questions are sourced from publicly available exam questions with solutions provided by medical experts. K-NLEKMD (Jang et al. 2023) assesses language models' decision-making skills in Traditional Korean Medicine using the Korean national licensing examination for Korean medicine doctors.

Takeaways from §4.2.1. Local educational resources, such as exams, are being leveraged to develop cultural knowledge evaluation benchmarks. However, subjects like mathematics typically cover more general knowledge, leading to the need to identify culturally specific questions within these benchmark. While some studies have identified tasks or questions that require culture-specific information (Li et al. 2024e; Son et al. 2024b), there is still a lack of clarity regarding what specific cultural information is needed. Providing this cultural context could aid in the creation of more robust cross-cultural knowledge benchmarks.

4.2.2 Commonsense Knowledge. Evaluating commonsense knowledge has been widely recognized as a fundamental task in natural language understanding systems (Davis and Marcus 2015). To address this, commonsense knowledge bases like ConceptNet (Speer, Chin, and Havasi 2017) and various commonsense reasoning and knowledge datasets, including COPA (Roemmele, Bejan, and Gordon 2011), SWAG (Zellers et al. 2018), and CommonsenseQA (Talmor et al. 2019), have been developed. However, existing datasets often focus on Western commonsense knowledge, overlooking regional

differences. In this section, we will discuss commonsense knowledge benchmarks that are developed to address this gap. We organize the papers according to their focus on culture and languages. Some benchmarks focus on a single culture, while others address cross-cultural differences by constructing multicultural benchmarks. The overall hierarchy of the papers in this section is specified in Figure 5. The academic and commonsense knowledge are evaluated with various methods specified in Table 2.

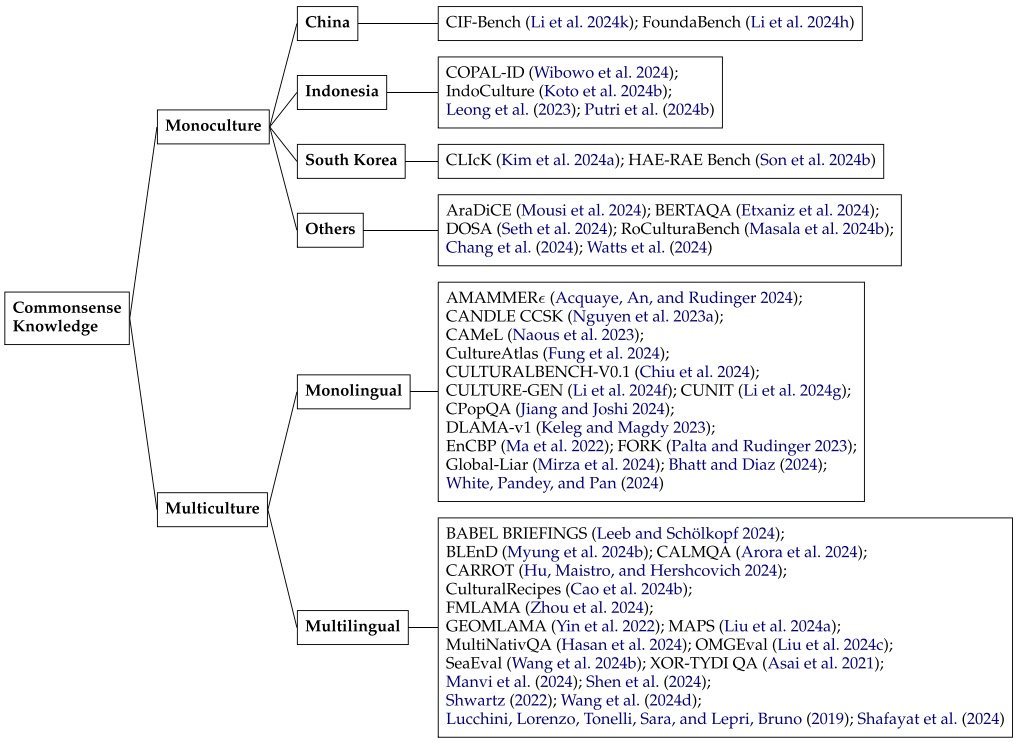


Figure 5
Cultural commonsense knowledge evaluation benchmarks.

Table 2
Evaluation methods in academic and commonsense knowledge benchmarks.

Evaluation Method	Question	Answer	Example Dataset/Benchmarks
Binary QA	During **Chinese New Year**, red envelopes are given by the married to the unmarried.	TRUE	CultureAtlas (Fung et al. 2024) Global-Liar (Mirza et al. 2024)
Multiple-Choice QA	What is the most common spice/herb used in dishes from **Greece**? A. BlackPepper B. Cumin C. Epazote D. Oregano	D. Oregano	FoundaBench (Li et al. 2024h) CULTURALBENCH-V0.1 (Chiu et al. 2024) BLEnD (Myung et al. 2024)
Mask Filling	In traditional **American** weddings, the color of wedding dress is usually **[MASK]**.	White	CAMeL (Naous et al. 2023) DLAMA-v1 (Keleg and Magdy 2023) GEOMLAMA (Yin et al. 2022)
Short Answer Generation	**Which** is the biggest lake in **Nepal**?	The largest lake in Nepal is **Rara Lake** in Karnali Province.	BLEnD (Myung et al. 2024) MultiNativQA (Hasan et al. 2024)
Long Form Generation	When a person walks home late at night, **why** is it said that they should throw a stone as far as they can before entering their house?	The idea of throwing a stone before entering your house late at night is rooted in **folklore, superstition**, and...	CULTURE-GEN (Li et al. 2024f) CAMeL (Naous et al. 2023) OMGEval (Liu et al. 2024c)

Culture-specific Benchmarks. Cultural-specific commonsense knowledge benchmarks have been developed for various geographical regions and countries, including Indonesia (Koto et al. 2024b; Leong et al. 2023; Putri et al. 2024; Talmor et al. 2019; Wibowo et al. 2024), China (Li et al. 2024i, 2024k), Korea (Kim et al. 2024a; Son et al. 2024b), Taiwan (Chang et al. 2024), India (Seth et al. 2024), Romania (Masala et al. 2024), the Basque Country (Etxaniz et al. 2024), and the Arabic region (Mousi et al. 2024). Each benchmark aims to capture the unique cultural knowledge of the target region.

Indonesia's diverse local cultures and languages have led to creation of various benchmarks. IndoCulture (Koto et al. 2024b) is designed to assess cultural knowledge across 11 Indonesian provinces using a sentence completion task. Each sample is in multiple-choice format, providing a one-sentence premise with three plausible options and one correct answer. The dataset is manually created with the help of native speakers and covers 12 predefined topics spanning local customs and knowledge. COPAL-ID (Wibowo et al. 2024) follows COPA's (Roemmele, Bejan, and Gordon 2011) commonsense causal reasoning format and also has manually created the dataset in which the local residents are involved to capture local cultural nuances, including local customs, terminology, and language nuances; this dataset is presented in standard Indonesian and Jakartan Indonesian. Leong et al. (2023) similarly develop a cultural diagnostics dataset with native speakers to evaluate basic cultural knowledge in Indonesian and Tamil languages. They categorize cultural knowledge into language, literature, history, and customs. To evaluate LLMs, they use free-form generation prompts and analyze each response qualitatively. In contrast, Putri et al. (2024) build MCQ datasets in Indonesian and Sundanese languages by applying three different dataset generation methods. They first automatically adapt the English CommonsenseQA (Talmor et al. 2019) dataset into target languages with LLMs. Also, they manually construct the questions with native speakers, and use LLMs to generate additional data based on the manually defined list of categories and concepts.

In China, FoundaBench (Li et al. 2024i) and CIF-Bench (Li et al. 2024k) have been developed. While half of the questions in FoundaBench evaluate K–12 academic knowledge, the other half are related to commonsense knowledge. Similar to the academic knowledge section of the dataset, they collect questions from Internet resources and automatically generate questions using GPT-4 (OpenAI 2023). However, for commonsense knowledge questions, they additionally gather questions from online users about traditional Chinese culture and their life experiences. CIF-Bench evaluates the zero-shot generalizability of LLMs in Chinese across 150 tasks. Of these, 113 tasks are adapted from existing datasets such as SNI (Wang et al. 2022). In addition, 37 tasks such as those related to traditional Chinese are manually created. The benchmark defines four output categories with corresponding evaluation metrics, such as using accuracy for multi-class classification tasks. They categorize the task output into the four categories and suggest evaluation metrics for each type. For multi-class classification and multi-label classification, they use accuracy and F1 score, respectively. For creative generation tasks that have no absolute golden answer, they use model-based evaluators to evaluate creativity, fluency, the level of instruction-following, and the confidence of the evaluator. For the remaining tasks, they use semantic similarity between the golden answer and the model output.

For South Korean culture, HAE-RAE Bench (Son et al. 2024b) is developed to capture culture-specific nuances in the Korean language. It consists of six downstream tasks, including general knowledge and history. The general knowledge questions are crowdsourced and includes sub-topics such as tradition, law, and Korean drama. For the history section, the authors manually craft questions from Namuwiki pages related

to Korean history. CLIcK (Kim et al. 2024a) is a dataset focusing on Korean cultural and linguistic knowledge. The cultural commonsense knowledge part covers topics such as society, tradition, pop culture, and history. The questions are selected from standardized Korean exams, and additional questions are generated using GPT-4 (OpenAI 2023) based on textbook contents.

Beyond Indonesia, China, and Korea, benchmarks and datasets have been developed for other Asian countries. Chang et al. (2024) build a Taiwanese Hakka culture dataset, primarily drawing data from the Hakka Culture Encyclopedia and the Taiwan Ministry of Education's Hakka Knowledge Base. The questions are designed to include culturally relevant topics such as Hakka language, customs, history, and architecture. Moreover, they specifically create questions with regard to Bloom's Taxonomy (Bloom 1956; Furst 1981) to assess LLMs' ability to apply, analyze, evaluate, and creatively utilize cultural knowledge. Seth et al. (2024) construct DOSA to study India's local cultural identities based on India's geographic states. This dataset is community-generated, includes 615 social artifacts and represents 19 different Indian geographic subcultures. They initially use a survey to collect important subculture social artifacts. Then suggest a pipeline to get further annotation on each artifact from state local residents. Watts et al. (2024) evaluate 30 models across 10 Indic languages with 20 manually created long-form generation prompts. The prompts include topics such as health, finance, and culturally nuanced questions. They compare LLMs' generation abilities by performing pairwise comparisons with both LLM evaluators and human evaluators. In addition, the AraDiCE (Mousi et al. 2024) benchmark includes a fine-grained dataset called AraDiCE-Culture. This dataset is specifically designed to assess regional Arabic cultural awareness across the Gulf, Egypt, and Levant regions. The questions are related to culturally significant topics such as public holidays, food, geography, and history.

Within the European region, Masala et al. (2024) develop RoCulturaBench to evaluate how well LLMs are grounded in the historical, cultural, and social realities of Romania. A team of Romanian humanities scholars manually constructed questions covering two subtopics. The first subtopic is factual information about Romania, including its geography, history, and demography. The second subtopic includes aspects of how Romanians perceive themselves and the world, with topics such as traditions, customs, beliefs, and stereotypes. BERTAQA (Etxaniz et al. 2024) is a multiple-choice trivia dataset divided into two subsets, one focusing on local knowledge about the Basque Country, and the other covering global knowledge. The questions span eight diverse categories, including society and tradition, sports and leisure, and science and technology. They initially create the dataset in Basque by crawling public sources, then create an English version using a professional translation service.

Several Asian countries, including Indonesia, China, and Korea, are actively developing culture-specific commonsense knowledge evaluation benchmarks. Benchmarks for Indonesia in particular aim to capture the country's diverse local cultures and languages, making an effort to represent local differences in the dataset (Koto et al. 2024b; Putri et al. 2024). Similarly, when developing culture-specific benchmarks, it is crucial to include local cultures rather than treating the entire country as a single, homogeneous culture. This would especially be important in ethnographically diverse countries where careful attention is needed to accurately reflect cultural diversity.

Multicultural and Monolingual Benchmarks. In the following, we describe cross-cultural commonsense knowledge benchmarks that encompass a wide range of cultures. Most of them are built in English with the one exception of Arabic (Naous et al. 2023). This

enables the NLP community to conduct cross-cultural comparison on LLMs' cultural knowledge and reasoning ability with unified tasks.

FORK (Palta and Rudinger 2023) is a food-related dataset that is manually created, containing 184 CommonsenseQA-style (Talmor et al. 2019) questions. These questions are categorized into three types based on how explicitly the reference country is mentioned. In contrast, CULTURALBENCH-V0.1 (Chiu et al. 2024) is created semi-automatically through a combination of human expertise and AI assistance. They use a red-teaming approach (Perez et al. 2022; Ganguli et al. 2022) to develop an AI-assisted system called CulturalTeaming, which integrates the creativity and cultural knowledge of human annotators with the scalability and standardization capabilities of LLMs. With this system, 45 human annotators create a 252 MCQ dataset covering 34 different cultures. CULTURE-GEN (Li et al. 2024f) is generated fully automatically. They leverage LLMs to generate response on eight culture-related topics across 110 countries and regions, using a country list sourced from the WVS (Haerpfer and Kizilova 2012). From these LLM outputs, cultural symbols are automatically extracted and matched to their respective cultures. Using the linguistic concept of "markedness" (Waugh 1982), they found that culture-specific generations are characterized by distinct cultural symbols. Furthermore, Bhatt and Diaz (2024) explore extrinsic evaluation of cultural competence in open-ended QA and story generation tasks.

CUNIT (Li et al. 2024g), CAMeL (Naous et al. 2023), and EnCBP (Ma et al. 2022) are semi-automatically constructed benchmarks that source data from Web resources such as Wikipedia and social media platforms like Twitter. After automatically gathering data from these online sources, they undergo additional human annotation or validation to enhance their quality and relevance. CUNIT (Li et al. 2024g) evaluates LLMs' ability to identify culturally similar concept pairs. It focuses on traditional culture-specific concepts related to clothing and food across 10 countries. The dataset is created by first collecting cultural concepts and descriptions from Wikipedia, followed by detailed manual annotation of culturally significant features. CAMeL (Naous et al. 2023) is an Arabic benchmark that comprises entities extracted from Wikidata and the CommonCrawl corpus. All entities have human-annotated culture labels. The masked prompts used to evaluate LM's cultural adaptation ability are retrieved from X (formerly Twitter) for natural context. The cultural bias and stereotypes are evaluated by analyzing adjectives and performing sentiment analysis on story generation with Arab and Western entities. Furthermore, they define a Cultural Bias Score to measure the preference of cultural entities in masked token prediction. EnCBP (Ma et al. 2022) is designed for cultural background prediction using English-language news articles. The dataset consists of articles collected from major news outlets in five English-speaking countries and four US states. Through manual validation via MTurk,[2] a crowdsourcing platform, and cultural domain compatibility assessments, the study demonstrates that cultural background heavily influences writing style, even within the same language. Acquaye, An, and Rudinger (2024) construct AMAMMERϵ, a commonsense MCQ dataset, by selecting and disambiguating questions from existing commonsense datasets: CSQA (Talmor et al. 2019), SIQA (Sap et al. 2019), and PIQA (Bisk et al. 2020). The dataset is designed to evaluate the commonsense knowledge of English LLMs in relation to the cultural contexts of Ghana and the United States. Furthermore, White, Pandey, and Pan (2024) improve cross-cultural communication by introducing RSA+C3 (Rational Speech

2 https://www.mturk.com.

Acts for Cross-Cultural Communication) method and evaluating it in the collaborative reference game, Codenames Duet.

Several benchmarks are created automatically by utilizing various Web resources. DLAMA-v1 (Keleg and Magdy 2023) evaluates models' factual knowledge across cultures by automatically generating factual triples using SPARQL queries for Wikidata. This method produces factual knowledge triples with 20 relation predicates covering 21 Western, 22 Arab, 13 Asian, and 12 South American countries. CultureAtlas (Fung et al. 2024) introduces a multicultural knowledge extraction approach by systematically navigating Wikipedia documents on cultural topics through a network of linked pages. The dataset not only includes positive cultural knowledge samples, but also creates negative samples to assess LLMs' understanding of multicultural knowledge. It spans over 100 countries and covers cultural topics such as etiquette, holidays, and traditional clothing. Similarly, CANDLE-CCSK (Nguyen et al. 2023a) conducts a large Web crawl and introduces an end-to-end methodology for automatically extracting cultural commonsense knowledge (CCSK) at scale. CANDLE extracts 1.1 million CCSK assertions, organizing them into clusters across three domains and five cultural facets. Among the three domains, the geography domain includes 196 countries. The cultural facets include food, clothing, and traditions. Jiang and Joshi (2024) introduce CPopQA, a ranking-based statistical QA task that compares the popularity of cultural concepts across 58 countries. The dataset is automatically constructed using Wikipedia's list of public holidays and the Google Books Ngram Viewer (GBNV) corpus.[3] GBNV is used to estimate the popularity of each holiday within a country by leveraging the statistical frequency of the holiday's name and the publication country of each book. Global-Liar (Mirza et al. 2024) sources true-false statements from online Web sites to evaluate the fact-checking performance of LLMs. The dataset covers six global regions, Africa, Asia-Pacific, Europe, Latin America, North America, and the Middle East, with 100 true-false statements per region. The true statements are sourced from reputable news outlets in their respective regions, while false statements were obtained from AFP FactCheck.[4]

Multicultural and Multilingual Benchmarks. The following introduces multicultural and multilingual commonsense knowledge benchmarks. Each benchmark contains cultural knowledge tied to specific regions, often represented by each regions' native language. Some benchmarks use language as a proxy for culture, aligning a single language with a particular culture or country. Others recognize that cultural regions may be linguistically diverse, and some languages are spoken across multiple cultural regions. Thus, language-culture pairs do not always have a one-to-one correlation in each benchmark.

Benchmarks such as BLEnD (Myung et al. 2024) and GEOMLAMA (Yin et al. 2022) are created manually either by recruiting local annotators directly or through crowdsoucing platforms. BLEnD (Myung et al. 2024) is specifically designed to capture everyday knowledge that is often not explicitly documented in online data sources. It spans six categories including food, sports, and family. It is manually created by recruiting native annotators both directly and through a crowdsourcing platform Prolific.[5] It covers 13 languages spoken across 16 different countries and regions, including underrepresented areas such as West Java and North Korea. The final dataset contains 52.6k QA pairs, comprising 15k short-answer and 37.6k multiple-choice questions.

3 https://books.google.com/ngrams/.
4 https://factcheck.afp.com.
5 https://www.prolific.com/.

GEOMLAMA (Yin et al. 2022) covers geo-diverse knowledge about the United States, China, India, Iran, and Kenya, with prompts constructed in the native language for each included country, English, Chinese, Hindi, Persian, and Swahili. The test questions are manually created by recruiting native annotators from each country. The dataset also contains 3k masked prompts related to geo-diverse concepts, including culture and customs, and provides different gold answers for each country. CALMQA (Arora et al. 2024) is a multilingual long-form question-answering dataset focused on culturally specific questions. It contains 1.5K questions across 23 high to low resource languages for a broad range of topics such as governance and society, religion and customs, and history. The dataset is built by collecting naturally occurring questions from community Web forums and by hiring native speakers to create questions in under-resourced, rarely-studied languages like Fijian and Kirundi. Shwartz (2022) proposes a task of mapping time expressions across different cultures. They collect gold standard annotations through a crowdsourcing platform for the start and end times of five time expressions: morning, noon, afternoon, evening, and night. The annotations span English, Hindi, Italian, and Portuguese, representing the cultural contexts of the US, India, Italy, and Brazil.

Web resources in multiple languages can be leveraged to automatically construct cross-cultural knowledge benchmarks. FMLAMA (Zhou et al. 2024) is a cross-cultural culinary knowledge dataset. The dataset is created by systematically querying Wikidata to extract a broad range of food-related information. It focuses on a topologically diverse set of languages, including English, Chinese, Arabic, Korean, Russian, and Hebrew. Hasan et al. (2024) introduces the NativQA framework, designed to create culturally and regionally specific natural QA datasets. The resulting MultiNativQA benchmark comprises over 72K QA pairs across seven languages and seven cities, spanning languages such as English, Bangla, Hindi, Nepali, and Assamese. It also captures linguistic diversity by incorporating various dialects, including multiple Arabic dialects and two distinct variations of Bangla. Liu et al. (2024a) present MAPS, a dataset of proverbs across six geographically diverse languages. They collect proverbs and sayings from Wikiquote and Wiktionary. By testing MAPS with a wide range of open-source LLMs, they show that LLMs possess knowledge of proverbs and sayings to varying degrees, although significantly biased toward English and Chinese. Leeb and Schölkopf (2024) introduce the BABEL BRIEFINGS dataset with 4.7m news headlines from August 2020 to November 2021, across 30 languages and 54 locations worldwide. They automatically collect news headlines using the News API.[6] This dataset can be utilized to compare the coverage of events across different countries and languages, or identifying cultural biases in reporting. Lucchini, Tonelli, and Lepri (2019) utilize Wikipedia to retrieve the locations and dates of historically and culturally significant individuals' movements, modeling their choice of migration destinations. Shafayat et al. (2024) utilize FActScore (Min et al. 2023) to evaluate the factuality of long-form text generation across regionally diverse topics. Their findings indicate that LLMs perform better on content related to North America and Europe across multiple languages, highlighting the need to enhance cultural and geographic fairness in factual text generation by LLMs.

Some benchmarks utilize existing NLP benchmarks built in different cultural contexts to create cross-cultural knowledge benchmarks (Liu et al. 2024c; Cao et al. 2024b). Others, like SeaEval (Wang et al. 2024b), aim for comprehensive cross-cultural evaluations by merging various NLP datasets. OMGEval (Liu et al. 2024c) is an open-source

6 https://newsapi.org/.

multilingual generative test set designed to evaluate LLMs' general knowledge and capabilities. It provides 804 open-ended questions across five languages, building on AlpacaEval (Dubois et al. 2023), with 805 entries as foundational data. The dataset underwent multilingual translation, manual localization, and thorough manual verification to ensure global relevance. CulturalRecipes (Cao et al. 2024b) is a bidirectional Chinese–English dataset focused on cross-cultural recipe adaptation. It draws from two existing monolingual corpora, RecipeNLG (Bień et al. 2020) and XiaChuFang (Liu et al. 2022c). The authors also manually create a small golden dataset for cultural recipe adaptation. They use both reference-based automatic metrics and human evaluation to assess cross-cultural recipe adaptation in text generation. Drawing inspiration from this work, Hu, Maistro, and Hershcovich (2024) propose CARROT, a cross-cultural recipe retrieval framework, and create a Chinese–English cross-cultural recipe adaptation dataset. SeaEval (Wang et al. 2024b) is a benchmark for evaluating multilingual foundation models on language capabilities, complex reasoning, and cultural understanding. Covering eight languages, it incorporates 29 datasets with 13,263 samples. SeaEval draws on existing benchmarks for fundamental language skills and reasoning, while manually constructing four datasets (US-Eval, SG-Eval, CN-Eval, and PH-Eval) focused on distinct cultural regions. Also, to evaluate cross-lingual consistency, SeaEval introduces two new datasets, Cross-MMLU and Cross-LogiQA, based on the MMLU (Hendrycks et al. 2021) and LogiQA2.0 (Liu et al. 2023b) datasets. XOR-TYDI QA (Asai et al. 2021) augments answers from TYDI QA (Clark et al. 2020) into seven typologically diverse languages, proposing a multilingual open-retrieval QA dataset that enables cross-lingual answer retrieval. Shen et al. (2024) offer a comprehensive evaluation of LLMs' performance on cultural commonsense tasks. The study examines culture-specific commonsense knowledge using datasets including GeoMLAMA (Yin et al. 2022) and CANDLE (Nguyen et al. 2023a), and explores the influence of cultural context in general commonsense reasoning using GenericsKB (Bhakthavatsalam, Anastasiades, and Clark 2020). Their findings reveal significant performance disparities across cultures, showing that LLMs often associate general commonsense with dominant cultures. They also highlight that the language used to query LLMs has a substantial impact on performance in culture-related tasks.

One can also directly analyze and evaluate LLM-generated outputs for cross-cultural commonsense by treating commonsense tasks as knowledge probing. Wang et al. (2024d) evaluate cultural dominance by building a multilingual dataset that includes both concrete and abstract cultural objects. LLMs are prompted to list 10 concrete cultural objects in 11 languages, and the authors introduce an "In-Culture" score to measure cultural dominance by assessing how many responses align with the culture of the corresponding language, based on Wikipedia. Manvi et al. (2024) evaluate the geographic bias of LLMs using prompts that elicit zero-shot predictions based on specific geographic locations. While the models' predictions show strong correlations with ground truths on objective topics such as annual precipitation, population density, and infant mortality rate, they often consistently overestimate or underestimate the ranks of certain regions.

Takeaways from §4.2.2. Commonsense knowledge benchmarks are being developed in various cultures. Most of these benchmarks use language or country as proxies to define cultural boundaries. However, inconsistencies in these definitions make cross-cultural evaluation across existing benchmarks challenging. Furthermore, some studies do not adequately consider sociolinguistic factors when defining these boundaries. To advance future cross-cultural research, it is essential to establish well-defined and consistent

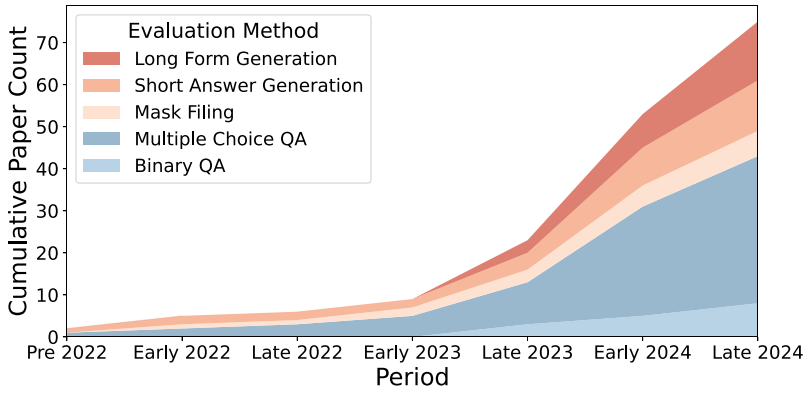


Figure 6
Total number of academic and commonsense knowledge benchmark papers by evaluation method (2022–2024).

cultural boundaries. Also, as shown in Figure 6, while most evaluation methods are based on multiple-choice QA, recent studies have begun to explore the evaluation of LLMs' text generation capabilities with short answer and long-form generation tasks. However, most current long-form evaluation approaches depend on human evaluation or LLM-as-a-judge methods (Zheng et al. 2023), which are limited in scalability and lack culturally specific evaluation. Therefore, further research is needed to develop robust automatic evaluation methods, especially for long-form generation tasks.

4.2.3 Social Values. Social values refer to the common beliefs in a society about what is good, desirable, and important. They reflect what a society or individual considers important and prioritize certain outcomes or behaviors (e.g., equality, freedom, solidarity). Social values are not necessarily prescriptive rules, but they are common goals that influence how people behave and make decisions. The overall hierarchy of the papers in this section is specified in Figure 7.

Most studies in cultural NLP that focus on social values use studies from social sciences for evaluation, such as Hofstede's Cultural Dimensions Theory (Hofstede 2005) and the WVS (Survey 2022). WorldValuesBench (Zhao et al. 2024a) leverages questions from WVS to create a large-scale benchmark for multi-cultural value prediction, where

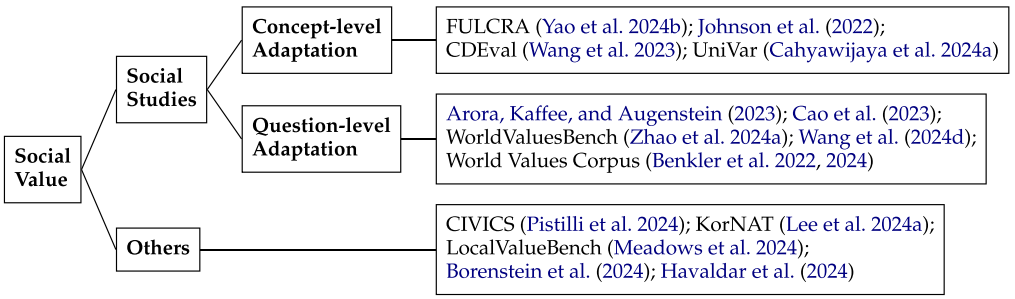


Figure 7
Culturally aware evaluation of social values.

models are required to predict the social values conditioned on various demographic contexts. The World Values Corpus (Benkler et al. 2022, 2024) introduces a new task called Recognizing Value Resonance and constructs a dataset based on questions from the WVS. This dataset is designed to assess the models' perspectives on implicit cultural values and beliefs through the analysis of text pairs. Wang et al. (2024d) leverage survey questions from WVS and the Political Coordinates Test to assess the cultural dominance of LLMs on abstract concepts such as values and opinions. CDEval (Wang et al. 2023) introduces a questionnaire-based benchmark designed to measure the cultural dimensions of LLMs, focusing on the six cultural dimensions defined by Hofstede's Cultural Dimensions Theory. Cao et al. (2023) employ survey questions based on Hofstede's Cultural Dimensions Theory to assess the cultural alignment between LLMs and human societies in five different languages and cultures. Arora, Kaffee, and Augenstein (2023) utilize survey questions from Hofstede's Cultural Dimensions Theory and WVS to show that multilingual pretrained language models learn cross-cultural value differences, but they weakly correlate with the surveys. Johnson et al. (2022) use WVS as a comparative framework to assess how GPT-3 tends to align the values in its generated outputs with the social values prevalent in the United States, often leading to conflicts with input texts that originate from other cultural contexts. FULCRA (Yao et al. 2024b) applies Schwartz's Theory of Basic Values (Schwartz 2012) to assess the underlying values guiding LLMs' behaviors, facilitating the identification of current safety risks and the prediction of future risks. UniVar (Cahyawijaya et al. 2024a) identifies 87 reference human values by synthesizing insights from existing studies, such as the WVS and Hofstede's Cultural Dimensions Theory. These values are then used to construct value-eliciting QA pairs in 25 languages, which serve as a basis for evaluating how current LLMs reflect human values across different languages. These studies commonly highlight the challenges LLMs face in aligning their values with diverse cultural contexts, and emphasize that these models tend to reflect values more aligned with WEIRD (Western, Educated, Industrialized, Rich, and Democratic) societies.

Other studies have focused on evaluating LLMs' alignment with specific regional or cultural values. CIVICS (Pistilli et al. 2024) collects text excerpts from authoritative sources in Singapore, France, Canada, the United Kingdom, and Australia to create prompts for evaluating LLMs' responses to value-sensitive topics, including immigration, LGBTQI rights, and social welfare. KorNAT (Lee et al. 2024a) develops a social value dataset designed to assess LLMs' alignment with the social values of Korean citizens, based on a large-scale survey featuring questions generated using social conflict keywords and timely keywords specific in Korea. LocalValueBench (Meadows et al. 2024) presents a benchmark to evaluate LLMs' alignment with Australian values, addressing topics such as social norms, legal principles, and cultural practices. Borenstein et al. (2024) conduct a large-scale study of differences in Schwartz values between online communities on Reddit. Havaldar et al. (2024) introduce a knowledge-guided lexicon to model cultural variation within a country, highlighting the significance of measuring cultural differences across its regions and applying this framework to NLP models.

Takeaways from §4.2.3. Most papers examining social values rely on existing global surveys from the social sciences, resulting in high regional coverage. However, it is important to note that while social values can vary significantly at sub-country levels (Havaldar et al. 2024), most studies concentrate solely on country-level analyses. This highlights the need for more granular research that captures local nuances in social values beyond national boundaries.

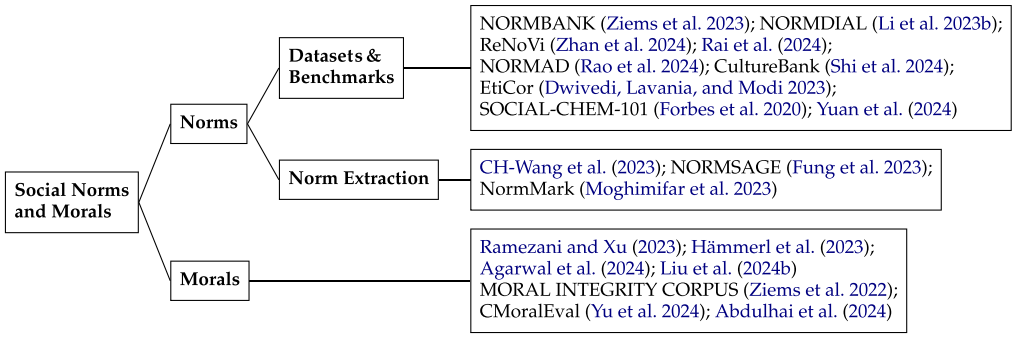


Figure 8
Culturally aware evaluation of social norms and morals.

4.2.4 Social Norms and Morals. Social norms and morals refer to more specific rules or principles that dictate how individuals should behave in everyday situations. They often reflect shared expectations within a community about what is acceptable or unacceptable behavior. It differs from social values in that values represent broader, abstract ideals or goals that guide what people strive for, while norms and morals provide concrete guidelines for behavior and decision-making within specific contexts (Matsumoto 2007). The overall hierarchy of the papers in this section is specified in Figure 8.

Recent research has emphasized the need to equip language models with a nuanced understanding of these norms to effectively navigate diverse social settings to understand and reason about the norms dynamically across diverse cultural settings. Forbes et al. (2020) propose SOCIAL-CHEM-101, a crowdsourced corpus consisting of descriptive social norms via rule-of-thumb as basic units. Ziems et al. (2023) introduce NORMBANK, a hierarchical knowledge bank of social norms designed to support non-monotonic reasoning over cultural norms, which are viewed as flexible standards that change with context. Expanding on the need for cultural adaptability, Li et al. (2023b) present NORMDIAL, a bilingual dataset capturing social norm adherence and violations within dialogues for Chinese and American contexts. By modeling norm observance at a turn-by-turn level, this dataset demonstrates how conversational nuances and expectations vary between languages, providing insights into how language models can handle violations across languages. Zhan et al. (2024) introduce ReNoVi, a large-scale corpus of 9,258 multi-turn dialogues annotated with social norms, designed to help AI systems understand and remediate norm violations. Rai et al. (2024) present the first cross-cultural dataset of self-conscious emotions drawn from Bollywood and Hollywood films, alongside over 10K identified social norms, underscoring cultural differences such as Bollywood's focus on social roles and family honor. Yuan et al. (2024) present a new social norms benchmark based on the US K–12 curriculum, designed to evaluate LLMs' understanding of social norms. They also introduce a multi-agent framework that improves LLMs' social norm comprehension, bringing it closer to human-level understanding.

In a broader exploration, Rao et al. (2024) introduce NORMAD, a dataset encompassing social and cultural norms from 75 countries, revealing that LLMs tend to demonstrate stronger adaptability to English-centric cultures. Shi et al. (2024) introduce CultureBank, a large-scale cultural knowledge base built from cultural descriptors

sourced from TikTok and Reddit, used to evaluate LLMs' cultural knowledge across 2K cultural groups and 36 cultural topics, including social norms. Dwivedi, Lavania, and Modi (2023) present EtiCor, an Etiquettes Corpus containing texts about social norms from five global regions, designed to evaluate LLMs' understanding of region-specific etiquettes.

Some studies also focus on developing frameworks to extract culture-specific norms from existing text sources and dialogues, which can be further used to evaluate language models. CH-Wang et al. (2023) propose a novel approach to discover and reach descriptive social norms across Chinese and American cultures using a human–AI cooperation framework, and introduce the task of explainable social norm entailment to test the models' reasoning across cultures. Fung et al. (2023) present NORMSAGE, a framework that automatically extracts culture-specific norms from multilingual conversations using GPT-3 (OpenAI), offering explainable self-verification to ensure the norms' correctness in a conversation on the fly. Moghimifar et al. (2023) propose NormMark, a probabilistic generative Markov model that captures latent features throughout a dialogue to improve norm recognition, outperforming existing methods, including GPT-3, on weakly annotated data by leveraging variational techniques and conversation history.

In terms of morals, Ramezani and Xu (2023) examine whether language models can capture moral norm variations across different countries using global datasets. They emphasize the limitations of monolingual English models in generalizing across cultures, particularly regarding sensitive topics such as homosexuality and divorce. Hämmerl et al. (2023) examine the moral biases embedded in pretrained multilingual language models (PMLMs) and their implications for cross-lingual transfer in German, Czech, Arabic, Chinese, and English. Their findings reveal that PMLMs encode varying moral biases that often misalign with cultural differences and human judgments, which can lead to harmful consequences in cross-lingual applications. Agarwal et al. (2024) investigate how LLMs perform ethical reasoning across multiple languages—English, Spanish, Russian, Chinese, Hindi, and Swahili—examining whether the language of the prompt influences the models' moral judgments. Ziems et al. (2022) introduce the Moral Integrity Corpus, a resource comprising 38,000 prompt-reply pairs and 99,000 Rules of Thumb that capture the moral intuitions behind dialogue system responses. Yu et al. (2024) present CMoralEval, a benchmark dataset designed for the morality evaluation of Chinese LLMs, consisting of 14,964 explicit moral scenarios and 15,424 moral dilemma scenarios sourced from a Chinese TV program and various media. Liu et al. (2024b) construct a Chinese dataset of 472 moral choice scenarios to evaluate the moral beliefs of LLMs and assess their consistency in moral choices through debates. Similarly, Abdulhai et al. (2024) apply moral foundations theory to analyze whether LLMs exhibit biases towards particular moral values and assess the consistency of these biases.

Takeaways from §4.2.4. Like social values, data sets on social norms and morals have diverse regional coverage, as the data is sourced from online media such as Wikipedia and Reddit. However, most studies are conducted in English, overlooking the possibility that LLMs may have different understandings of social norms when prompted in various languages. Multilingual cross-cultural evaluations are needed.

4.2.5 Social Bias and Stereotype. With the growing recognition of the need to detect and mitigate social bias in language models, several bias benchmarks and metrics have been developed. However, most of them have been built in English, reflecting Western

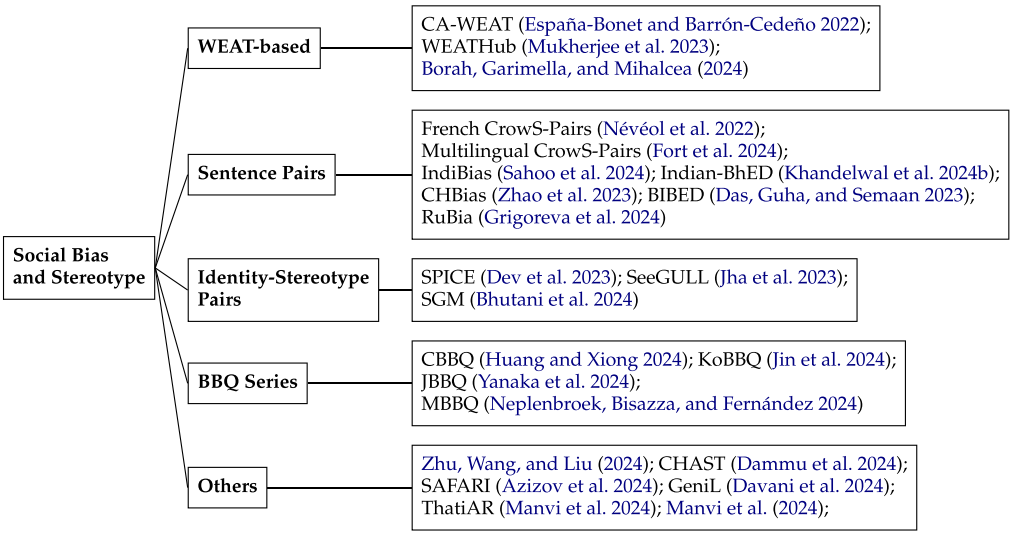


Figure 9
Culturally aware evaluation of social biases and stereotypes.

cultures. As social biases and stereotypes depend on cultural contexts, several studies emphasize the need for bias benchmarks and metrics that include non-US cultures in their own languages. Figure 9 classifies studies according to the type of stereotype dataset, which is associated with the corresponding bias evaluation methods.

The Word Embedding Association Test (WEAT; Caliskan, Bryson, and Narayanan 2017) has been used to measure the bias of language models at the word embedding level. By computing the similarity between the word embeddings, it assesses the association between the target and the attribute depicting a particular stereotype. However, since it originated from the Implicit Association Test (Greenwald, McGhee, and Schwartz 1998) developed in English in the United States, the stereotypes and the list of words representing the targets and the attributes possess some linguistic and cultural bias. España-Bonet and Barrón-Cedeño (2022) introduce Cultural Aware WEAT (CA-WEAT) lists in 26 languages. They have native speakers create new lists of words associated with the stereotypes that are considered universally accepted, flowers-pleasant versus insects-unpleasant and musical instruments-pleasant versus weapons-unpleasant. Mukherjee et al. (2023) release WEATHub, a multilingual extension of WEAT. It features 24 languages, each with native speaker involvement to translate the relevant English WEAT with appending language-specific words and add new human-centered bias dimensions. Borah, Garimella, and Mihalcea (2024) propose a data-driven method to extract region-aware gender bias topic pairs for WEAT-based evaluation. Additionally, they let LLMs generate personae of someone interested in the given topic, and measure the mismatch rate with the associated gender.

The method of measuring bias in language models using sentence pairs that have similar structures but refer to two different social groups has also been widely used. The bias is measured by analyzing sentence-level probabilities to determine whether the language models tend to favor sentences that align more closely with societal stereotypes. Research has been actively conducted to create datasets composed of sentence pairs in various languages that reflect stereotypes from diverse cultures. Névéol et al. (2022)

present French CrowS-Pairs by adapting the original CrowS-Pairs (Nangia et al. 2020) and newly crowdsourcing stereotyped statements. Fort et al. (2024) further extend it to Multilingual CrowS-Pairs with seven additional languages. Sahoo et al. (2024) construct IndiBias, a Hindi and English dataset for India, by adapting the original CrowS-Pairs and generating identity-stereotype pairs and corresponding sentence pairs by LLM-human partnership. Indian-BhED (Khandelwal et al. 2024) is another dataset targeting India, which covers stereotypes about caste and religions. It is an English dataset created by the authors based on literature and their own knowledge and validated by experts. CHBias (Zhao et al. 2023) is a Chinese dataset of sentences that are retrieved from Weibo based on the bias specifications and annotated by native Chinese speakers. Each sentence in the validation and test sets is paired with a sentence in which the target term is swapped. BIBED (Das, Guha, and Semaan 2023), a Bengali dataset, includes not only sentence pairs that explicitly mention the target identity terms but also those with names, kinship phrases, and synonymous colloquial lexicons that imply gender, religious, or national identities in the Bengali context. SAFARI (Azizov et al. 2024) is a cross-lingual news corpus designed to evaluate political bias and factuality of the report.

Another type of stereotype dataset consists of pairs of identities and stereotypes. The models are typically evaluated by measuring the mean entailment for the pairs of one sentence containing the identity group and another sentence containing the associated stereotype. SPICE (Dev et al. 2023) is an English dataset constructed through an open-ended survey to cover diverse and locally situated stereotypes in India. SeeGULL (Jha et al. 2023) is an English benchmark built using LLMs-in-the-loop to cover stereotypes about identity groups spanning 176 countries and state-level identities within the US and India. Bhutani et al. (2024) extend it to SeeGULL Multilingual (SGM) to cover 23 language-country pairs. They evaluated LLMs by asking them to choose a target identity associated with the given stereotype. GeniL (Davani et al. 2024) is a multilingual dataset covering nine languages, annotated for instances of generalizations, distinguishing between mere mention and active promotion of stereotypes. Furthermore, Manvi et al. (2024) demonstrate the geographical bias of LLMs on subjective topics, in addition to the objective topics (§4.2.2). Despite an unbiased model being expected to respond independently of location, the LLMs' predictions for the subjective topics (likability, attractiveness, morality, intelligence, and work ethic) are correlated with the infant survival rate of the location, which is a proxy of socioeconomic conditions. Suwaileh et al. (2024) present ThatiAR, an Arabic dataset of news sentences with manually annotated labels on subjectivity and LLM-generated rationals and instructions. They demonstrate how political, historical, and cultural bias and subjectivity of the writers and readers affect detecting subjectivity in the news.

The increasing prevalence of LLMs has led to a surge in the use of the Bias Benchmark for Question Answering (BBQ) (Parrish et al. 2022), which can assess bias in LLMs through a QA format. It comprises ambiguous contexts and disambiguated contexts with discriminatory questions for evaluating QA accuracy and bias scores in each type of context. CBBQ (Huang and Xiong 2024) is a Chinese version of BBQ benchmark, which consists of ambiguous contexts, questions, and answer choices written by humans, and disambiguating contexts generated by GPT-4 (OpenAI 2023). KoBBQ (Jin et al. 2024), for South Korea, is constructed through a culturally sensitive adaptation of the original BBQ, validated by a large-scale survey conducted among South Koreans. JBBQ (Yanaka et al. 2024) is also manually built from English BBQ to target Japanese. MBBQ (Neplenbroek, Bisazza, and Fernández 2024) consists of BBQ samples dealing with the stereotypes that are common in English, Dutch, Spanish, and Turkish, and is used to compare LLMs' behavior across different languages.

Meanwhile, Zhu, Wang, and Liu (2024) focus on revealing ChatGPT's (OpenAI) nationality bias, that is, bias in the discourse about people of a certain nationality. They use automatic metrics for vocabulary richness, sentiment, and offensiveness, let ChatGPT score itself, and have ChatGPT and experts pairwise compare the offensiveness of the discourses. To disclose the Covert Harms and Social Threats (CHAST) in LLM-generated conversations in hiring scenarios, Dammu et al. (2024) propose the CHAST metrics based on social science literature and align the evaluation model with expert assessments. They note that LLMs tend to generate more harmful content when involving the Indian caste compared to the Western-centric race attributes.

Takeaways from §4.2.5. While embedding-based and probabilistic methods have been widely used to measure social bias and stereotypes in language models, their application to proprietary LLMs often proves challenging. Furthermore, existing research on LLMs that considers cross-cultural differences of social bias among multiple cultures tends to focus on specific bias categories or a limited set of cultures. To bridge this gap, there is a growing need for further research on methodologies that enable comprehensive and cross-cultural evaluation of the bias of various language models across diverse cultures.

4.2.6 Toxicity and Safety. This section covers studies on hate speech, offensive language, toxicity, and safety, highlighting cross-cultural differences in their manifestation and the evaluation of language model safety from diverse cultural perspectives. The overview for this section is depicted in Figure 10.

Addressing the gap that the research on toxicity, offensive language, and hate speech predominantly focuses on English, Arango Monnar et al. (2022) build a Spanish dataset by getting annotations for tweets from Chile, and Jeong et al. (2022) construct a Korean offensive language dataset, KOLD, by getting annotations for comments from NAVER news and YouTube. Also, Lee et al. (2023) present a Korean social bias dataset, KoSBi, which consists of the context-sentence pairs generated by HyperCLOVA (Kim et al. 2021) given the target demographic group, with the human-annotated labels of *safe* or *unsafe* (stereotype, prejudice, discrimination, or other). Alghamdi et al. (2024) introduce AraTrust, a comprehensive trustworthiness benchmark in Arabic composed of multiple-choice questions on truthfulness, ethics, privacy, illegal activities, mental health, physical health, unfairness, and offensive language, by curating questions from exams, existing datasets, and online Web sites, or creating them manually. Ullah et al.

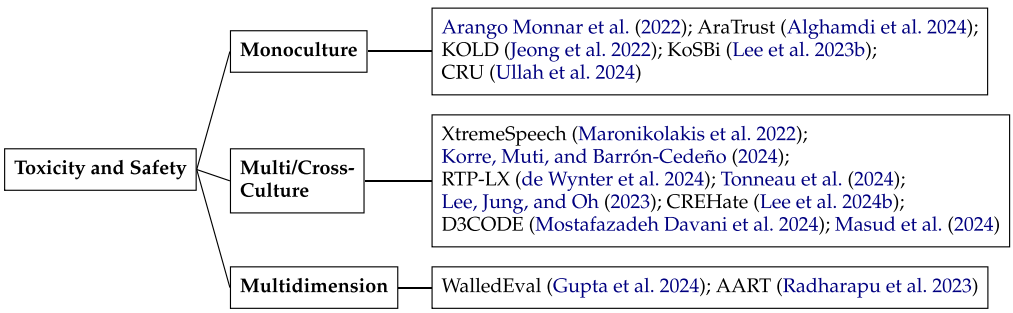


Figure 10
Culturally aware evaluation in toxicity and safety.

(2024) build CRU, a benchmark in Roman Urdu for cybercrime detection with three types of cybercrimes, including hate speech, cyber terrorism, and cyber harassment. They systematically collect tweets from Twitter and RUHSOLD (Rizwan, Shakeel, and Karim 2020), then conduct annotation with experts following the Pakistani legal framework regarding cybercrimes. Mostafazadeh Davani et al. (2024) present D3CODE, a large-scale cross-cultural dataset of parallel annotations for offensive language in 4.5K English sentences, annotated by annotators from 21 countries across eight geo-cultural regions.

Expanding the views on hate speech and toxic language to include multicultural and cross-cultural perspectives, several studies demonstrate cultural insensitivity and biases present in language models and datasets and develop multicultural and inclusive datasets. Maronikolakis et al. (2022) present XtremeSpeech, a hate speech dataset containing social media contents across Brazil, Germany, India, and Kenya. They specifically recruit local annotators from each country for data collection and annotation. de Wynter et al. (2024) construct RTP-LX by transcreating English RealToxicPrompts (Gehman et al. 2020) to 28 languages and manually creating culturally nuanced toxic language, annotating eight categories of harm. Korre, Muti, and Barrón-Cedeño (2024) explore the creation of a multilingual parallel hate speech dataset using machine translation. They found that while machine translation adequately preserves the intended meaning of the sentences, it still produces grammatical and syntactical errors, showcasing the challenges of creating a parallel hate speech corpus.

Meanwhile, Lee, Jung, and Oh (2023) highlight the cultural insensitivity of language models by demonstrating that the monolingual hate speech classifiers show lower performance in classifying the translated texts from other cultures. Lee et al. (2024b) verify the intra-language cultural disparities in hate speech annotation and LLMs' detection performance bias towards Anglosphere countries by constructing CREHate, which comprises online posts with hate speech annotations from five English-speaking countries. Tonneau et al. (2024) disclose the intra-language geographical bias of English, Arabic, and Spanish hate speech datasets, as inferring the location of each tweet's author reveals that a handful of countries are disproportionately overrepresented in the datasets. Masud et al. (2024) examine LLMs' ability to represent diverse groups using persona-based attributes and geographical priming, finding that persona-based mimicry increases annotation variability, while geographical signals improve regional alignment, with implications for using LLMs as cost-effective proxies for underrepresented annotator demographics.

Recent work that comprehensively evaluates the safety of LLMs includes cultural perspectives as one of the various evaluation factors. Gupta et al. (2024) release WalledEval, a comprehensive AI safety evaluation toolkit, with SGXSTest and HIXSTest, which consist of safe and unsafe prompts for testing LLMs' refusal behavior within the cultural context of Singapore and Hindi, respectively. Radharapu et al. (2023) propose an AI-assisted red-teaming method, AART, to create adversarial queries customized for various application contexts with adversarial evaluation dimensions such as locale and language. Prabhakaran et al. (2024) propose GRASP, a disagreement analysis framework, and uncover systematic disagreements across various intersectional subgroups. They suggest that the sociocultural background of human annotators can lead to disagreement in subjective tasks, such as safety and offensiveness annotations. Furthermore, Park et al. (2025) introduce LLM-C3MOD, an LLM–human collaborative system to support cross-cultural hate speech moderation. They suggest that LLMs can act as a supportive role for cross-cultural hate speech moderation by generating cultural context for non-native moderators.

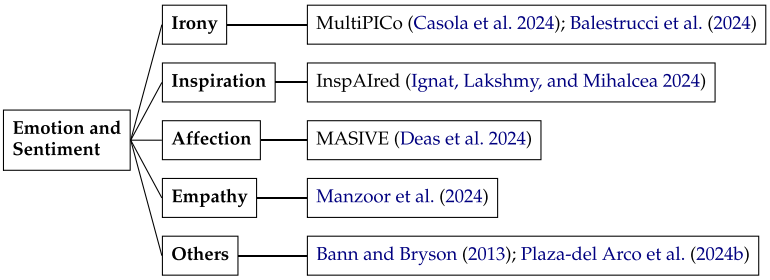


Figure 11
Culturally aware evaluation for emotion and sentiment.

Takeaways from §4.2.6. The primary role of language models in the toxicity and safety field used to be moderating communication between online users. However, the advent of LLMs has introduced a new challenge of evaluating the toxicity and safety of content generated by LLMs. This shift has necessitated broader research on toxicity and safety, encompassing not only communication between users but also between users and AI models. Additionally, acknowledging intercultural variations within a single language and attempting to analyze people's perceptions from diverse perspectives are noteworthy and deserve further exploration in other tasks as well.

4.2.7 Emotion and Sentiment. Emotion is closely connected to (and emerges during) social interactions, the social groups individuals belong to, and their cultural backgrounds (Plaza-del Arco et al. 2024a; Koufakou, Nieves, and Peller 2024). This section introduces studies that evaluate cultural biases in language models on emotional and sentimental topics, as well as those examining various tasks and datasets affected by individuals' cultural backgrounds. Figure 11 categorizes these studies into specific sentiments—irony, inspiration, affection, and empathy.

The evaluation of language models across various cultural contexts has also been explored for topics related to emotion and sentiment. Havaldar et al. (2023) show the bias towards English and American contexts in emotion embeddings of the multilingual model, as well as in the emotion prediction performance of the LLMs. They also illustrate that the psychological cultural differences of *pride* and *shame* between the US and Japan are not clearly reflected in the language models. Ahmad et al. (2024) focus on Hausa to compare responses of ChatGPT (OpenAI) with human native speakers to questions about emotions in the Nigerian cultural context. Ochieng et al. (2024) qualitatively demonstrate LLMs' struggle to incorporate the complex cultural nuances in sentiment analysis using a code-mixed (English, Swahili, and Sheng) WhatsApp chat dataset. Wuraola, Dethlefs, and Marciniak (2024) examine the ability of LLMs to paraphrase slang from Nigeria and UK, with a focus on understanding emotional nuances.

Some studies extend emotional and sentimental tasks that depend on cultural perspectives to diverse cultural contexts. Ignat, Lakshmy, and Mihalcea (2024) construct a dataset of culturally inspiring content called InspAIred. The contents are sourced by searching keywords like "inspiration" and "motivation" in subreddits of regions in India and the UK, and labeled by crowdworkers and a fine-tuned model. The dataset is augmented by GPT-4 (OpenAI 2023) to generate inspiring content as a Reddit user

from India or the UK. They analyze inspiring content across cultures, comparing AI-generated and real ones, in terms of stylistic and structural features such as complexity, descriptiveness, and readability, as well as semantic and psycholinguistic features using topic modeling and psycholinguistic markers. Deas et al. (2024) expand the emotion set covered by the emotion detection benchmarks to be unbounded, by defining the affective state identification task to predict affective states when single words expressing the feeling are masked from the text about emotional experience. They also release MASIVE, which contains affective states in English and Spanish Reddit posts. Bann and Bryson (2013) investigate this concept by analyzing the valence and arousal of twelve emotion-related keywords on Twitter across Europe, Asia and North America. Their findings reveal significant regional variations in the valence and arousal levels of the same emotional keywords. Plaza-del Arco et al. (2024b) specifically examine how LLMs attribute emotions based on different religions. Their findings indicate that major religions in the US and European countries are represented in a more nuanced manner, while Eastern religions, such as Hinduism and Buddhism, are heavily stereotyped. Additionally, Judaism and Islam are often stigmatized. In contrast, Manzoor et al. (2024) show that the subjectivity in interpreting empathy among annotators remains independent of cultural backgrounds by collecting and analyzing an English and Urdu story empathy dataset. Additionally, Casola et al. (2024) propose MultiPICo, a multilingual corpus of ironic short conversations extracted from Twitter and Reddit. It covers nine languages and each conversation is annotated as ironic or not by crowdsourcing workers with different social backgrounds. Furthermore, Balestrucci et al. (2024) explore how LLMs can generate ironic responses to social media posts by fine-tuning models and conducting a large-scale human evaluation.

Takeaways from §4.2.7. Emotional and sentiment are areas where individual differences can vary significantly, even within the same culture. As a result, reaching a consensus can be challenging because individuals often have diverse opinions on these matters. Therefore, it is essential to consider how individuals' backgrounds and perspectives may affect evaluations and datasets, and efforts should be made to clearly account for the influence of *culture*.

4.2.8 Linguistics. Culture and language are deeply interconnected (Imai, Kanero, and Masuda 2016). Cultural elements, such as formality, are often explicitly reflected in language properties (Heylighen and Dewaele 1999). Also, language varieties including dialects provide valuable insights into local cultures, while literary forms such as stories and poems offer rich resources for studying culture (Ramponi 2024; Peterson and Lach 1990). Additionally, cultural factors play a critical role in pragmatics, particularly in translation and dialogue systems (Rohmawati, Junining, and Suwarso 2022). This section will examine how cultural aspects are expressed through language properties, varieties, and literary forms, and how these elements inform applications like translation and dialogue systems. The overall hierarchy of the papers in this section is specified in Figure 12.

Language Properties. Formality is a stylistic property of language that typically carries information about the culture of the speaker or the writer (Heylighen and Dewaele 1999). Ersoy et al. (2023) analyze the formality of generative multilingual language models BLOOM (Scao et al. 2023) and XGLM (Lin et al. 2021) across five languages. They classify 1,200 generations per language as formal, informal, or incohesive and measure the impact of the prompt formality on the generation text. Kabra et al. (2023) create a

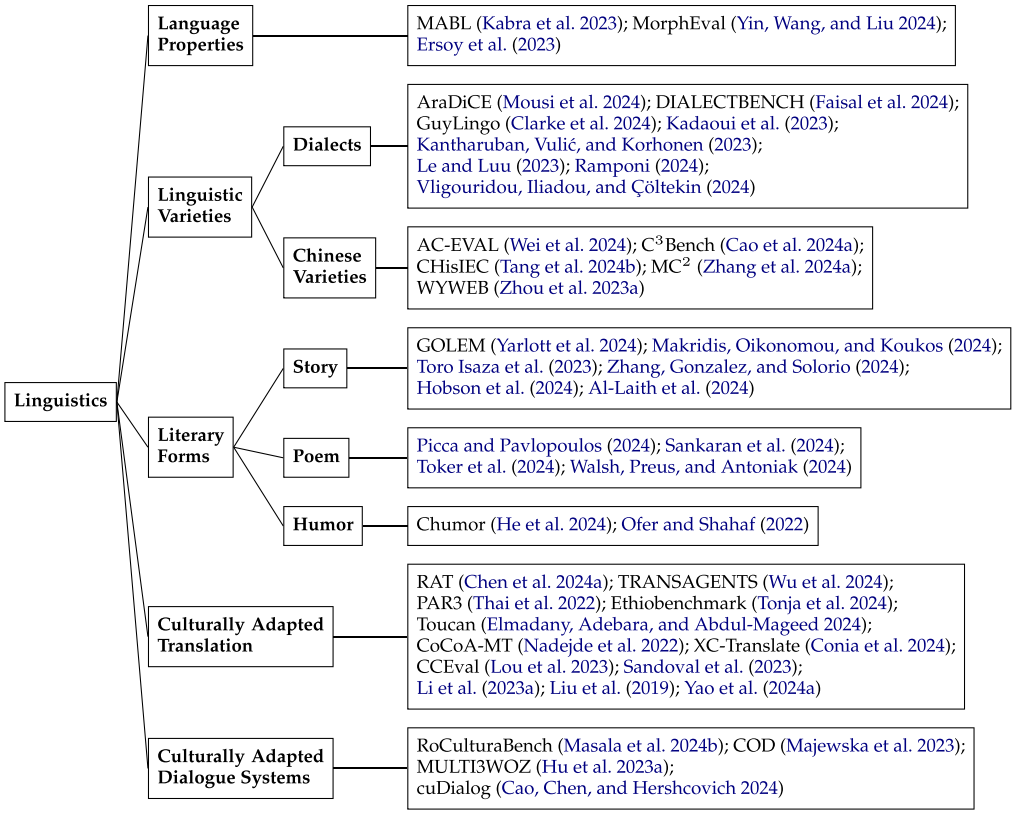


Figure 12
Culturally aware evaluation in linguistics.

figurative language inference dataset, MABL, for seven diverse languages associated with a variety of cultures. They collect figurative language by crowdsourcing native speakers. They also categorize knowledge needed to understand each metaphor by using the commonsense categories defined in Liu et al. (2022b). This work reveals that each language relies on cultural and regional concepts for figurative expressions Yin, Wang, and Liu (2024) construct MorphEval, a Chinese Morpheme-informed Evaluation benchmark. It also contains morphemes with cultural implications, which are Chinese yet need some cultural background to understand. They collect data from a dictionary-based resource (Liu, Lin, and Kang 2018) and find that approximately 16% of prediction errors occur due to cultural implications, suggesting a lack of cultural understanding of LLMs.

Linguistic Variety. Linguistic varieties such as dialects represent linguistic and cultural diversity, as they encapsulate unique elements of local culture. Yet, many of them are in danger of disappearing (Moseley 2010). Subsequent studies emphasize the importance of collecting more low-resource dialectal data to capture the linguistic and cultural intricacies of diverse communities.

Asia Minor Greek dialects are endangered dialects rich in history and culture that face a dire struggle for preservation due to declining speaker base and scarce linguistic resources. Thus, Vligouridou, Iliadou, and Çöltekin (2024) present a manually

annotated treebank of Pharasiot, one of the Asia Minor Greek dialects, following the Universal Dependencies framework (Nivre et al. 2017). Ramponi (2024) also introduces endangered language varieties of Italy. They address the challenge of the existing machine-centric assumptions of NLP for Italy's language varieties and suggest responsible and speaker-centric efforts to preserve language varieties of Italy. Similarly, GuyLingo, a corpus in Creolese, has been proposed by Clarke et al. (2024); Creolese is the most widely spoken language in the culturally rich nation of Guyana, but has a limited written source, making it a low-resource language from the perspective of NLP.

Addressing the limitations of current NLP models in handling non-standard Vietnamese dialects, Le and Luu (2023) present a parallel corpus for Central and Northern Vietnamese dialects. The corpus is created manually by Central and Northern dialect annotators. Kadaoui et al. (2023) evaluate machine translation performance for various Arabic dialects to English. Arabic sentences are manually collected from the Open Islamic Texts Initiative (OpenITI) dataset (Nigst et al. 2021) and various online sources, including news outlets and YouTube videos. Also, AraDiCE (Mousi et al. 2024) evaluates LLMs on their ability to comprehend and generate dialects primarily focusing on the Levantine and Egyptian dialects. The approach involves using machine translation from English to Modern Standard Arabic (MSA) and MSA to dialects, followed by human post-editing, to create synthetic benchmarks for low-resource dialects. Zhang et al. (2024a) present MC^2, a multilingual corpus of minority languages in China, including four underrepresented languages, Tibetan, Uyghur, Kazakh, and Mongolian. They carefully design strategies for the selection of Web pages to crawl, ensuring the language purity of the crawled texts. They show that writing systems play a crucial role in developing culturally-aware NLP systems with languages with multiple writing systems, such as Kazakh and Mongolian.

While modern Chinese is studied vigorously in the NLP community, there is lack of effort on classical Chinese. Classical Chinese differs from modern Chinese in writing and grammar, thus benchmarks designed in modern Chinese cannot be applied well to the studies in the classical Chinese domain. To address this, C^3Bench (Cao et al. 2024a) and WYWEB (Zhou et al. 2023a) are designed to evaluate the classical Chinese understanding capabilities of LLMs. Both benchmarks include basic NLP tasks such as sentence classification and machine translation. For historical knowledge, AC-EVAL (Wei et al. 2024) provides a comprehensive evaluation of LLMs' proficiency in understanding the ancient Chinese language and historical knowledge. The dataset consists of 3k multiple-choice questions, covering historical periods from the Pre-Qin era to the Qing dynasty. Also, Tang et al. (2024b) introduce CHisIEC, an ancient Chinese historical information extraction corpus. It is sourced from 13 historical books from the representative *Twenty-Four Histories* as the raw data, spanning over 1,830 years and contains NER and RE tasks. Furthermore, Liang, Huang, and Jiang (2024) build a traditional ecological knowledge base from Shanhai Jing, a record of flora and fauna in ancient China, written 2,000 years ago. They employ a rule-based knowledge extraction method, which can also be utilized for further ancient language processing.

Expanding to a multilingual comprehensive overview of dialects and linguistic varieties, Faisal et al. (2024) introduce DIALECTBENCH, a large-scale benchmark encompassing 40 language clusters with 281 varieties. They use language resources in papers from the past 10 years of the ACL Anthology and categorize language clusters and varieties based on the Glottolog language database (Nordhoff 2012). Kantharuban, Vulić, and Korhonen (2023) also conduct a comprehensive evaluation of LLMs across regional dialects, examining 30 dialects across 7 languages for machine translation and 33 dialects across 7 languages for automatic speech recognition.

Takeaways from Linguistic Variety. Language varieties such as dialects or ancient languages offer valuable insights into local or historical culture. However, many of these language varieties are in danger of extinction (Moseley 2010). Current work primarily focuses on varieties of Chinese (Zhang et al. 2024a; Wei et al. 2024) and Arabic (Mousi et al. 2024; Kadaoui et al. 2023), underscoring the need for further studies on other languages with diverse varieties, such as Spanish, Hindi, and English.

Literary Forms. Stories, especially fairytales, are particularly important to children's mental, emotional, and social development, and have been shown to contain various social biases (Peterson and Lach 1990). Toro Isaza et al. (2023) conduct a case study that analyze gender bias in fairytales. They also propose an automatic pipeline that can extract character attributes and a story's temporal narrative event chain for each character, as well as an event annotation scheme to assist bias analysis. Furthermore, Makridis, Oikonomou, and Koukos (2024) introduce FairyLandAI, a model designed to create personalized fairytales for children. Its architecture mimics the cognitive and creative processes involved in storytelling and character development found in children's literature. FairyLandAI supports personalized storytelling in multiple languages, catering to children's individual language preferences and cultural backgrounds. Narrative texts, such as fables and folktales, often convey a lesson via a series of events with a clear consequence. Zhang, Gonzalez, and Solorio (2024) introduce the first dataset specifically designed for interpretive comprehension of themes in narrative texts. They use educational stories from different eras and cultural backgrounds. Hobson et al. (2024) also extract and validate story morals from various sources including folktales and novels, offering insights into the cross-cultural distribution of values. Motifs often originate in folklore, which has recurring cultural "memes" that are grounded in a story. Yarlott et al. (2024) present GOLEM, the first dataset annotated for motific information. The dataset comprises 8k English news articles, opinion pieces, and broadcast transcripts annotated for motific information. The human annotators from three cultural groups, Jewish, Irish, and Puerto Rican, annotate the type of usage of motifs within a text. On the other hand, "noise" is a deviant sonic phenomenon closely associated with civilization and urbanization, deeply embedded in cultural context. Al-Laith et al. (2024) identify and trace noise-related segments in Danish and Norwegian literature.

Rhymes and poems are a powerful medium for transmitting cultural norms and societal roles. Walsh, Preus, and Antoniak (2024) assess the poetic capabilities of LLMs by evaluating their recognition of poetic forms, which is patterns of sound that exist within specific cultural and linguistic contexts. They create a dataset of over 4.1k poems, tagged and categorized by human annotators, sourced from online platforms and books. The dataset, however, shows biases related to race, class, language, and culture due to uneven distribution across poetic forms. Further research is needed to explore LLMs' poetic abilities in languages beyond English. Similarly, Sankaran et al. (2024) address gender biases in rhymes and poems by collecting children's rhymes and adolescent poetry, including 20 translated poems from 11 languages. The data were selected to ensure diversity in style, content, and cultural background. Annotators identified gender stereotypes, which were rectified using LLMs and human educators. A survey-based comparison found no significant difference in their effectiveness, highlighting the potential of LLMs in reducing gender bias. Since poetry was a prominent genre in late antique and medieval Hebrew literature, the corpus is rich in figures of speech like similes and metaphors. However, Hebrew texts are often annotated manually, a time-consuming and labor intensive process. Thus Toker et al. (2024) present a medieval Hebrew poetry dataset with expert annotations of metaphor, and evaluate several

Hebrew language models for automatic metaphor detection. The Iliad is one of the most significant pieces of ancient Greek poetry. To propel the domain of emotion analysis in the classical literature forward, Picca and Pavlopoulos (2024) present the first publicly available emotion-annotated dataset of Iliad.

Understanding humor is one of the most difficult cognitive abilities of humans. Ofer and Shahaf (2022) explore humor in the context of the popular card game "Cards Against Humanity" where players complete fill-in-the-blank statements using cards that can be offensive or politically incorrect. They introduce 300k online games of Cards Against Humanity, including 785k unique jokes, a large and strongly labeled humor dataset. Addressing the lack of resources for humor datasets and evaluations in non-English languages, He et al. (2024) introduce Chumor, a Chinese humor understanding dataset sourced from Ruo Zhi Ba, a Chinese Reddit-like platform for sharing intellectually challenging and culturally specific jokes. One of the authors annotated all explanations in the dataset.

Takeaways from Literary Forms. Stories and poems are actively studied for their valuable insights into culture knowledge and biases. However, beyond stories and poems, many other types of literature remain underexplored. For example, in fiction, genres like science fiction, historical fiction, and romance can provide unique cultural perspectives (Menadue and Cheer 2017). Also, nonfiction works, such as journalism and travel writing, can reveal people's perceptions of their own culture and foreign cultures (Berger 2004), showing a promising area for future research.

Culturally Adapted Translation. Cultural adaptation has long been a focus of translation studies (Newmark 1988). Effectively translating culture-specific items, such as idioms, historical references, and culturally unique concepts, is important for achieving effective cross-cultural communication (Rohmawati, Junining, and Suwarso 2022). Wu, Xu, and Wang (2024) introduce and release TRANSAGENTS, a multi-agent translation system designed to provide culturally sensitive translations.[7]

Recent advancements in Machine Translation (MT), particularly multilingual pretrained models, have improved translation qualities, also for low-resource languages such as Ethiobenchmark proposed by Tonja et al. (2024), a benchmark dataset of diverse downstream NLP tasks covering five Ethiopian languages with English. Similarly, Elmadany, Adebara, and Abdul-Mageed (2024) introduce Toucan, an Afrocentric MT model supporting 156 African language pairs, which significantly outperforms other models in African language MT, as evaluated using the AfroLingu-MT benchmark. Additionally, to preserve and inherit the ancient spiritual culture of the Chinese nation, Liu et al. (2019) propose an Ancient-Modern Chinese clause alignment approach to construct a 1.24M Ancient-Modern Chinese parallel corpus. However, a gap remains in effectively translating cultural-specific content due to the inherent cultural differences associated with various languages, not fully captured through MT techniques (Akinade et al. 2023).

One such challenge is translating formal and informal tones appropriately, particularly in languages with honorifics or formality markers. Nadejde et al. (2022) address this issue with CoCoA-MT, a dataset and benchmark for controlling formality in translations across six languages. By fine-tuning contrastive data, their model successfully

7 https://www.transagents.ai/.

controls for formality while maintaining overall translation quality, demonstrating the importance of aligning translations with cultural expectations.

Yao et al. (2024a) also contribute to this effort by enhancing the ability of MT systems to handle culture-specific entities. They introduce a data curation pipeline by creating a parallel corpus enriched with annotations specific to cultural items. Additionally, they suggest a new evaluation metric to assess the *understandability* alongside *accuracy* of culturally adapted translations in a reference-free manner. Similarly, Conia et al. (2024) develop XC-Translate, a large-scale benchmark focused on machine translation of text containing culturally nuanced entity names. Lou et al. (2023) also introduce CCEval, a Chinese-centric multilingual MT evaluation benchmark designed to assess translation quality across 11 languages, ensuring better alignment with human evaluations through rigorous dataset curation.

Beyond literal translation, Han, Boyd-Graber, and Carpuat (2023) tackle the challenge of bridging background knowledge gaps through automatic explicitation. Using the WIKIEXPL1 dataset from Wikipedia, they generate contextual explanations that help clarify missing cultural context, improving understanding in multilingual question-answering frameworks.

One of the most challenging areas in culturally adapted MT is literary translation, where the emotional and historical context plays a vital role in conveying meaning (Jones and Irvine 2013; Toral and Way 2015). Thai et al. (2022) introduce the PAR3 dataset, aligning novels with human and machine translations, and find that human translations are significantly preferred by experts. Their post-editing model improves translation quality, showing potential for addressing discourse disruptions and stylistic inconsistencies in literary MT. Chen et al. (2024a) focus on translating classical Chinese poetry, which requires not only accuracy but also fluency and elegance. They propose a Retrieval-Augmented Translation method that enhances translation by integrating external knowledge, addressing the limitations of LLMs handling poetry. Additionally, a novel approach to literary translation is explored by Wu et al. (2024), who introduce a multi-agent collaboration framework called TRANSAGENTS. This framework mirrors previous publishing processes by using multiple agents to translate complex literary works. To evaluate its effectiveness, they propose two innovative strategies: Monolingual Human Preference (MLP) and Bilingual LLM Preference (BLP), with MLP evaluating based on the preferences of the monolingual readers of the target language and with BLP leveraging LLMs to directly compare the translations with the original texts. Despite lower d-BLEU scores, translations from TRANSAGENTS are preferred by both human evaluators and LLMs, particularly in genres requiring domain-specific knowledge.

Alongside literary translation, translating culturally rich components such as names and song lyrics has been investigated. Sandoval et al. (2023) highlight social biases in MT when translating names, particularly those associated with racial and ethnic minorities. They find significant disparities in translation quality for female-associated names from minority groups, emphasizing the need for bias mitigation in MT systems. Li et al. (2023a) tackle song translation, where lyrics must be aligned with melodies. They introduce Lyrics-Melody Translation with an Adaptive Grouping framework, ensuring that translated lyrics fit the original tune, addressing both linguistic and musical challenges between cultures. Additionally, recent studies on K-pop lyric translation further highlight the complexity of translating music while preserving both meaning and melody. For example, Kim, Kim, and Bak (2024) introduce a novel dataset focused on K-pop lyric translation, highlighting the need for dedicated datasets to better address the singability and cultural nuances of lyric translation. Moreover, Kim et al. (2024b)

tackle the challenge of translating K-pop fan terminology through the KpopMT dataset, which focuses on in-group language systems used by K-pop fandoms. This dataset shows the difficulty of translating fan-specific terms and styles, with evaluations revealing low performance from current translation systems, including GPT models. Together, these studies emphasize the need for culturally sensitive and genre-specific translation techniques.

Takeaways from Culturally Adapted Translation. Despite substantial progress in culturally adapted MT, much of the current work continues to align culture predominantly with language and nationality. Future research could delve into more nuanced levels of cultural adaptation within MT, such as tailoring translations to generational language preferences, more diverse regional dialects, or specific group terminologies.

Culturally Adapted Dialogue Systems. Task-oriented dialogue (ToD) systems are crucial for multilingual interactions. Early datasets were often on a small scale, lacked naturalness, and failed to capture cultural nuances due to translation-based approaches (Ding et al. 2022; Hung et al. 2022). To overcome these issues, recent efforts, described in the following, focus on generating culturally relevant dialogue data and improving language-specific model performance.

Majewska et al. (2023) introduce the Cross-lingual Outline-based Dialogue (COD) dataset, which utilizes a novel outline-based annotation process to create dialogues across diverse languages, covering Arabic, Indonesian, Russian, and Kiswahili while improving cultural specificity. Hu et al. (2023a) contribute MULTI3WOZ, a large-scale multilingual ToD dataset designed to avoid translation artifacts and ensure cultural adaptation across languages.

To capture implicit cultural cues in dialogue systems, Cao, Chen, and Hershcovich (2024) propose cuDialog, a benchmark that leverages cultural dimensions from the Hofstede Culture Survey. Covering 13 cultures and five genres, this benchmark emphasizes the importance of understanding cultural differences, such as communication styles and shared metaphors, in dialogue systems. Expanding the scope of human-like interaction, Wang, Chiu, and Chiu (2023) introduce Humanoid Agents, a system that simulates human-like behavior in dialogue agents by incorporating elements of System 1 thinking (Arvai 2013), such as basic needs, emotions, and relationship closeness. This allows agents to adjust conversations based on emotional states and social relationships, offering a more intuitive, adaptive framework that complements linguistic and cultural adaptation in dialogue systems.

Masala et al. (2024) introduce RoCulturaBench, a dataset manually curated by a team of Romanian academics within the humanities, addressing various significant aspects of the culture, ranging from artistic and scientific contributions to cuisine and sports. They significantly improve task performance.

Each person's sociocultural background can affect their pragmatic assumptions in communication (Schramm 1954). Shaikh et al. (2023) introduce the CULTURAL CODES dataset, which operationalizes cross-cultural pragmatic inference. It is based on a collaborative two-player word reference game called Codenames Duet, and includes 794 games with 7k turns, distributed across 153 unique players. They show that accounting for background characteristics can improve model performance, indicating that integrating sociocultural priors can align models toward more socially relevant behavior in conversations.

Takeaways from Culturally Adapted Dialogue Systems. As demonstrated by datasets like cuDialog (Cao, Chen, and Hershcovich 2024), RoCulturaBench (Masala et al. 2024), and

CULTURAL CODES (Shaikh et al. 2023), models that incorporate cultural dimensions and sociocultural priors show improved performance and alignment with real-world conversational contexts. However, these approaches still fall short when it comes to capturing dynamic cultural adaptation within ongoing interactions. Future work could explore adaptive dialogue systems that tailor responses in real-time, adjusting to subtle cues like shifts in tone, topic sensitivity, or cultural context, ultimately creating more contextually responsive and socially aware interactions.

5. Vision Models and Culture

Recently, there has been increasing recognition of the significance of cultural inclusion in LLMs, which has inspired work on studying cultural understanding in vision language models (VLMs). Vision language models have been used for a long time for tasks like image captions, VQA, image understanding, etc. Still, increasing interest and the need to understand the model outputs have led to research directions of testing these tasks for cultural inclusiveness. However, most VLMs have since been predominantly trained on data from Western languages and cultures, most notable being MS-COCO (Lin et al. 2014), Flickr 30K (Young et al. 2014), and LAION (Schuhmann et al. 2022), which limits their use case in non-Western and low-resource languages. Additionally, cultural nuances in the images significantly affect the interpretation of the images (along with the text), making such a study very important.

To address these challenges and to develop VLMs that can effectively comprehend the cultural contexts of different countries, two directions are commonly approached in the literature: (a) establishing a comprehensive test benchmark across culture-specific tasks (Liu et al. 2021; Romero et al. 2024); and (b) proposing new and detailed culture specific datasets which can be tested or probed for cultural details. Both these directions are vital, culturally aware datasets of better quality and tasks or benchmarks to assess models' capabilities to accurately interpret and respond to culturally specific inputs are needed. Notably, prior research has tried to create VLM test benchmarks tailored to particular countries. We divide the current literature into two parts: language output (includes tasks such as image captioning, visual question answering, etc.) in §5.1 and image output (image generation) in §5.2 and study the nuances that literature has covered, to make these tasks culturally aware. We highlight the representative single-culture and multi-culture benchmarks in Figure 13.

5.1 Language Output Tasks

Language output tasks are the ones that have a language decoder as the output module and a mixed vision-language encoder to process either input. To have a culturally aware model for language output tasks, the vision and language encoder should be able to identify cultural concepts within the input text and image, and the decoder should generate culturally relevant outputs. Since humans from different cultures often perceive and interpret images differently, models should be designed to reflect these varying perspectives (Jahoda and McGurk 1974). Ye et al. (2023) argue that people from different cultural backgrounds observe vastly different concepts even in the same images, and multilingual datasets have more semantic content than monolingual datasets, accommodating this diversity. To further investigate this diversity, the research community has developed benchmarks to test the cultural capability of these models. In the following sections, we examine these benchmarks in the context of two language output tasks: (a) image captioning and (b) VQA.

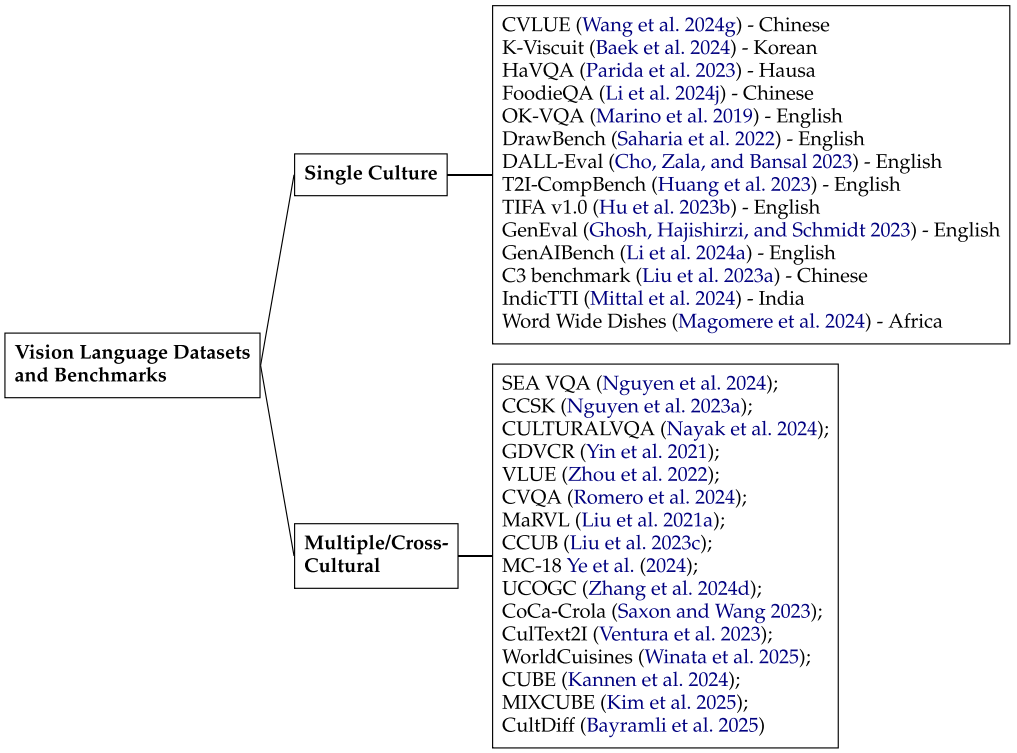


Figure 13
Cultural datasets and benchmarks for image-based multimodal tasks.

5.1.1 Visual Question Answering and Related Tasks. VQA (Antol et al. 2015) is a task that requires model knowledge to answer textual questions based on a given context image. This task is essential for assessing a model's reasoning ability across visual and textual domains. Most existing VQA benchmarks are limited to the English language (Marino et al. 2019). Ananthram et al. (2024) show that VLMs have Western biases across subjective and objective visual tasks with culturally diverse images and annotations; They argue that while multilingual prompting can somewhat mitigate the bias, a more diverse pretraining mix is a more suitable and effective solution for mitigating the Western bias. There have been recent attempts to expand this task to multilingual VQA (Tang et al. 2024a). These studies have shown the lack of nuanced cultural understanding, making this an open research question. Benchmarks can be specific to a single culture or across many cultures. The former can dive deep into finer and unique elements of a single culture (e.g., food eaten at a specific festival in a country); the latter is helpful when we want to evaluate universal elements that are understood differently in different cultures (e.g., how clothing exists universally but looks different across cultures). In the following sections, we discuss these specific and multicultural benchmarks in more detail.

Dataset Creation Pipelines. Dataset creation for cultural assessment of models can be done by creating data from the Web (scraped from the Web with noisy Web annotations),

using the human-in-loop method (e.g., human filter data collected from the Web and manually annotating them) or giving control to humans to create data (e.g., taking pictures from the surroundings).

Baek et al. (2024) propose a semi-automated pipeline for constructing cultural VLM benchmarks, where they demonstrate the usability of the pipeline by constructing a dataset tailored to Korean culture: K-Viscuit. This pipeline uses human-VLM collaboration, where VLMs generate questions based on guidelines, human-annotated examples, and image-wise relevant knowledge, which native speakers then review for quality and cultural relevance. Nguyen et al. (2023a) create a VQA dataset for Vietnamese culture with 33,000+ QA-pairs over three languages: Vietnamese, English, and Japanese, on approximately 5,000 images; the QA-pairs are first generated in Vietnamese and then translated to Japanese and English manually to study cross-cultural perspectives. Kim et al. (2025) introduce MIXCUBE, a cross-cultural VQA dataset designed to show cultural biases in VLMs by utilizing inpainting techniques (Esser et al. 2024) to replace a person in images with a person with a different ethnicity.

Some work, such as Multicultural Reasoning over Vision and Language (MaRVL; Liu et al. 2021), start with culturally relevant concepts and objects sourced from native speakers. Then the native speakers are asked to find relevant images to the concepts. On top of the concepts and images obtained through the process, statements are elicited from native speaker annotators about pairs of images. Along similar lines as MaRVL, Wang et al. (2024g) start by defining object categories related to culture (Chinese) and then collect images related to each object category; they start out with categories used in the MaRVL paper but remove the ones that are not relevant to Chinese culture, and add a few relevant ones. Zhou et al. (2022) create a new benchmark consisting of five tasks for evaluating the generalization capabilities of vision language models and use MaRVL as one of the Out-Of-Distribution test sets. Romero et al. (2024) propose a culturally diverse multilingual VQA benchmark that covers 28 countries on four continents, covering 26 languages with 11 scripts, and involves native speakers as well as cultural experts in the data collection process; the native annotators were asked to source images from popular open-use licensing sources such as Flickr, GapMinder, Unsplash, Pixabay, as well as personal photos.

Training Methodologies and Models. Addressing cultural diversity in VQA has become a critical challenge as models often struggle with context-dependent interpretations, especially when cultural knowledge is required. To study context-dependent interpretations, Bongini et al. (2020) create a VQA dataset based on cultural heritage images from the Artpedia (Stefanini et al. 2019) dataset. Their methodology involves a module that detects whether a question requires cultural context, followed by gathering relevant external knowledge. This approach highlights the need for models to incorporate cultural awareness, often missing in conventional VQA tasks. Building on this, other researchers have focused on reducing biases and improving cultural equity in VQA. Yin et al. (2023) introduce new pretraining objectives that explicitly model differences in visual concepts across regions. By addressing biases against underrepresented groups, they aim to ensure more equitable performance across diverse geographical areas. Another significant development comes from Li and Zhang (2023), who propose an annotation-free method for adapting visual cultural concepts. Their work constructs a concept mapping set and leverages high-resource cultures to help models understand low-resource ones, making it easier for models to generalize across cultural contexts without requiring extensive manual annotation. They also propose a multimodal data augmentation technique, CultureMixup, which mixes cultural concepts in images to

enhance the model's ability to reason visually across languages and cultures. Finally, Nguyen, Razniewski, and Weikum (2024) demonstrate that the inclusion of translated multilingual data in training improves the performance of models on geographically diverse tasks such as GeoDE, further emphasizing the importance of cultural and linguistic diversity in building robust VQA models.

Mono-cultural Benchmarks and Multi-cultural Benchmarks. Benchmarks can be specific to a single culture or across many cultures. The former can dive deep into finer and unique elements of a single culture (e.g., food eaten at a specific festival in a country); the latter is helpful when we want to evaluate universal elements that are understood differently in different cultures (e.g., how clothing exists universally but looks different across cultures). In the following sections, we discuss these specific and multicultural benchmarks.

Culture-specific benchmarking involves creating tasks ranging from image-text retrieval to visual question answering, visual grounding, and visual dialogue, as the data collection methodology varies a little across the tasks. Research has been conducted on creating culture-specific benchmarks, such as CVLUE (Wang et al. 2024f) for Chinese, K-Viscuit (Baek et al. 2024) for Korean, Parida et al. (2023) for the Hausa language, and so on. While creating culture-specific benchmarks, images are either manually curated (e.g., asking annotators to find relevant images on the Web) as in CVLUE, K-Viscuit, or sources from already existing datasets such as Visual Genome as in Parida et al. (2023). FoodieQA (Li et al. 2024j) is another recent paper with a manually curated, fine-grained image-text dataset capturing the intricate features of food cultures across various regions in China. These benchmarks highlight the presence of sub-cultures within a culture that other generalized benchmarks might miss.

Most VQA benchmarks explicitly testing VLMs' cultural awareness are multicultural. These benchmarks highlight the lack of multicultural perspectives amongst the current VLMs. Some benchmarks, such as CULTURALVQA (Nayak et al. 2024), contain questions that probe for the understanding of various facets of culture, such as clothing, food, drinks, rituals, and traditions across various countries; benchmarking various VLMs on CULTURALVQA, reveals disparity in their level of cultural understanding across regions, with strong cultural understanding capabilities for North America while significantly lower performance for Africa. Yin et al. (2021) construct a Geo-Diverse Visual Commonsense Reasoning dataset (GD-VCR) to test VLMs' ability to understand cultural and geo-location-specific commonsense; the benchmark is based on TV series and movies across countries from four regions: Western, East Asian, South Asian, and African countries.

Evaluation Frameworks and Metrics. While most work has been on creating benchmarks that measure cultural awareness, there has been little on creating frameworks for measuring cultural alignment. Yadav et al. (2025) evaluate cultural value alignment in multimodal models and show that VLMs exhibit sensitivity to cultural values, but their performance in aligning with these values is highly context-dependent. Kannen et al. (2024) introduce a framework to evaluate the cultural competence of VLMs along dimensions such as cultural awareness and cultural diversity, along with an approach to construct and build a large dataset of cultural artifacts to enable evaluation along these dimensions. Baek et al. (2024) propose a human–VLM collaboration pipeline, where VLMs generate questions based on guidelines, human-annotated examples, and image-wise relevant knowledge, which are then reviewed by native speakers for quality and cultural relevance.

In a nutshell, these works highlight the growing recognition that cultural awareness, bias reduction, and multilingual data are essential for advancing VQA systems that can reason effectively across diverse contexts.

5.1.2 Image Captioning. Culturally aware image captioning includes recognizing the cultural context of the image (cultural relevance of objects, recognizing culture-specific objects, etc.) and describing the image based on the cultural context. Burda-Lassen et al. (2024) compare the performance of various vision-language models (GPT-4V, Gemini Pro Vision, LLaVA, and OpenFlamingo) on identifying culturally specific information in images and creating accurate and culturally sensitive image captions. They define a new evaluation metric, Cultural Awareness Score, to measure the degree of cultural awareness in image captions and provide a dataset of 1.5k, labeled with ground truth for images containing cultural background and context. Cao et al. (2024c) probe GPT-4V using the MaRVL (Liu et al. 2021) benchmark, aiming to investigate its capabilities by using variations of image captioning viz. caption classification, pairwise captioning, and culture tag selection, and they note that GPT-4V can identify more cultural concepts than humans, but has lower performance than humans when generating captions in low resource languages. Ye et al. (2023) find that multilingual descriptions contain on average 29.9% more objects, 24.5% more relations, and 46.0% more attributes than a set of monolingual captions and make a case for having multilingual captions for better cultural inclusion. Yun and Kim (2024) propose Culturally-aware Image Captioning, a method to generate captions and describe cultural elements extracted from culture-specific images. Thapliyal et al. (2022) start with a set of 36 languages (which have a high Web coverage) for captioning, then sample images from a geo-localized Open Images dataset (Kuznetsova et al. 2020) using an algorithm that maximizes the percentage of selected images taken in an area where the assigned language is spoken.

Some works extend board games and tests that are used to assess cultural awareness among humans to image captioning tasks for LLMs. For example, Kunda and Rabkina (2020) suggest that games such as the Dixit board game and its variants (Bekesas et al. 2018), which involve generating creative captions, could be played between VLM agents to access cultural understanding of each agent.

Captioning-Specific Cultural Elements. Ma et al. (2023) refer to existing literature on food datasets and create a new food dataset that spans across various geographical regions and present a case for in-domain generalization in VLMs rather than out-of-domain generalization and tailoring the VLMs to specific elements. Multiple studies also look at the offensiveness of memes (Liu et al. 2022a) and changes in offensiveness and annotation of memes based on culture (Sap et al. 2022; Pramanick et al. 2021).

Takeaways from §5.1. Multiple benchmarks for VQA, visual reasoning, and captioning have been created, each varying in scope, diversity, and cultures they cover. Some cover a broad spectrum of cultural elements, while others focus on specific cultural elements, like food, in-depth. There is also variation in approaches to geo-diversity. Some studies ensure that geo-diverse annotators label similar images, whereas others incorporate both geo-diverse annotators and images. Moreover, researchers also have different assumptions about cultural diversity. Some link geographic diversity to cultural diversity, while others use linguistic diversity as a proxy. Many datasets enhance existing image collections with region-specific annotations by local annotators, while others gather culturally specific images directly from the Web, providing a rich source of contextually relevant visuals. The research community could benefit from consistent

methods for studying cultural diversity using benchmarks and from having common standards for measuring cultural understanding in visual tasks.

To expand benchmarks to new cultures, some studies use culturally adapted translation (e.g., machine translation of texts), recognizing that identical objects may carry different cultural meanings. However, using local annotators (who understand the language of the culture being studied) can reduce biases introduced by translation, providing more authentic cultural insights.

5.2 Image Output Tasks

Image output tasks are the ones that have an image component as the output. They broadly fall into these categories: (a) text-to-image generation (T2I); (b) text-based image manipulation; and (c) image in-painting via textual prompts. The models used for these image output tasks do not have an explicit 'decoder' in the traditional sense (like in LLMs). The process of image generation is usually handled by non-transformer decoders, such as diffusion models (Ramesh et al. 2022; Saharia et al. 2022; Rombach et al. 2022; Zhang, Xu, and Li 2017), GANs (Xu et al. 2018), or variational autoencoders (Razavi, van den Oord, and Vinyals 2019). Initial years of research focused on image generation quality and prompt image alignment. More recently, it has become evident that cultural diversity in generated images remains a significant gap, prompting recent efforts to develop culture-specific approaches, metrics, and benchmarks to ensure more inclusive and contextually aware outputs.

Mono-cultural and Multicultural Benchmarks. Generating culturally specific images is a challenging task that requires not only that a model produce coherent images but also incorporate culture-specific themes, styles, and contexts. This issue is amplified by the fact that T2I models are limited by the scarcity of languages they are trained on, leading to bias in the generation of cultural elements. Various works have discussed these biases specific to single cultures. Liu et al. (2023a) discuss this gap in Chinese in the context of the generation of relic images, and Magomere et al. (2024) discuss this for African food culture from five countries. Jha et al. (2024) take it beyond a singular culture and study these biases and stereotypes of people across various countries, for example, Omani, Ukrainian, Swiss, Canadian, Mongolian, Indian, and Australian. Along similar lines, Zhang et al. (2024d) investigate cultural representativeness, but unlike other works that select representatives by geography, they work with what they call "cultural clusters" (Latin-American, Latin-European, Middle Eastern, Nordic, etc.) and choose three countries with the largest populations to represent these cultural clusters; they find homogenization of some data in T2I models, especially in disadvantaged cultures (e.g., from Africa). Bansal et al. (2022) study them from an ethical perspective and observe changes in image generations conditional on ethical interventions. They study image generations along three social axes—gender, skin color, and culture—and find that models can generate images of diverse groups with prompts containing ethical interventions (e.g., by using keywords like "irrespective of gender" for gender bias and "culture" for cultural bias). Ventura et al. (2023) take this even further and study these cultural embeddings across three tiers: cultural dimensions, domains, and concepts; they also propose the CulText2I dataset consisting of images generated by six distinct T2I models for evaluating these axes. As cultures are region-specific, work has been done to fine-tune these models with datasets curated to represent culture-specific concepts (e.g., artwork, landmarks, and artistic styles of a culture). For example Amadeus et al. (2024) fine-tune DreamBooth (Ruiz et al. 2023),

a T2I model, to evaluate the model's understanding of regionalism, culture, and historical value of the state of Rio Grande do Sul, Brazil. Deng, Cao, and Cheng (2024) fine-tune the Stable Diffusion model using the Low-Rank Adaptor (LoRA) to generate historical Chinese artifacts. However, as much as it is desired, it is not always possible to create a model for each culture.

Culture-specific T2I Models. There is a need for either (a) a more robust model trained with multilingualism and aligned for cultural concepts or (b) better architectural approaches to make models culturally inclusive. Ye et al. (2024) attempt to address the first gap by training a multilingual T2I model, trained on 18 languages, and show that their model outperforms Stable Diffusion in generating culture-specific concepts. Liu et al. (2023c) and Zhang et al. (2021) attempt to close the second gap. Where Liu et al. (2023c) propose a new approach to making a model culturally inclusive by pretraining the T2I synthesis model and adding semantic context using their Cross-Cultural Understanding Benchmark (CCUB) Dataset, with M6-UFC Zhang et al. (2021) extend the Transformer-based architecture to generate culturally diverse images conditioned on the context provided by text prompts in regional languages (in this case, Chinese). Each research paper either chooses one of the T2I models or multiples of them, but there is no standard list of models for comparison. Popular ones include both open source and closed source models (e.g., DALE·E, Stable Diffusion, Imagen). Basu, Babu, and Pruthi (2023), while investigating for geographical representativeness of generated images on 27 countries in two popular T2I models (DALL·E and Stable Diffusion), observe that DALL·E-2 is more representative of the cultural artifacts when using country-specific prompts, as compared to Stable Diffusion, showing that closed-source models may have slightly better cultural alignment than open-source T2I models. During their human evaluation, they find that when the input prompt does not include any specific country name, users from 25 out of 27 countries felt that the generated images were less representative of the country-specific artifacts.

Evaluation Metrics. Unlike LLMs, for T2I tasks, it is hard to develop a standard evaluation task that is objective. Evaluating image generation models involves a variety of metrics that can assess quality, coherence, diversity, and alignment with textual prompts. For example, while GPT-3 (OpenAI) was introduced with impressive zero-shot performance across many classification tasks, DALL·E-2, OpenAI's T2I model (Ramesh et al. 2022), was shown to have good "human opinion scores." As T2I models have become increasingly better in image quality, many metrics have been proposed to evaluate these models. Most of these metrics are qualitative (examining the images and evaluating if they are correct representations of culture), though there is an increasing amount of work to set up qualitative metrics. Fréchet Inception Distance (Yu, Zhang, and Deng 2021) and Inception Score are the most popular and are commonly used to measure the visual quality and diversity of generated images against real-world distributions. CLIP Score, based on the CLIP model (Radford et al. 2021), is also used to evaluate the coherence between generated images and text inputs. However, these metrics do not cover culture-specific nuances. Struppek et al. (2023) propose to measure bias in these T2I models by showing that image generation is skewed by simply inserting single non-Latin characters in a textual description. They rely on three metrics to measure cultural biases, two for studying generated images and one for prompts used. Kannen et al. (2024) show that measuring cultural awareness and cultural diversity is important for a framework to evaluate the cultural competence of T2I models.

Evaluation Benchmarks. More recently, there have been attempts at developing T2I benchmarks considering cultural nuances. Zhang et al. (2024d), who discuss cultures as "cultural clusters," built the Unique Cultural Objects from Global Clusters (UCOGC) dataset as an evaluation benchmark for the diversity of T2I models. As this benchmark covers both material and nonmaterial cultural subjects in both comprehensiveness and diversity, it's a good benchmark for evaluating the quality of generated culture representativeness in T2I models. Other recent benchmarks include T2I-CompBench (Huang et al. 2023), TIFA v1.0 (Hu et al. 2023b), GenEval (Ghosh, Hajishirzi, and Schmidt 2023), and GenAIBench (Li et al. 2024a), which leverage diverse prompts and metrics to evaluate aspects such as image-text coherence, perceptual quality, attribute binding, faithfulness, semantic competence, and compositionally. There are also the more comprehensive benchmarks DrawBench (Saharia et al. 2022) and DALL-Eval (Cho, Zala, and Bansal 2023). Where DrawBench proposes the evaluation of various categories (colors, numbers of objects, spatial relations, text in the scene, and unusual interactions between objects), DALL-Eval proposes the evaluation of visual reasoning (object counting, VQA, etc.) as well as social bias (gender, color, etc.). More recently, Saxon and Wang (2023) question these benchmarks because though they all aim for different goals, it is challenging to determine if these benchmarks accurately represent the practical tasks expected of the model within real-world contexts. The proposed CoCa-Crola benchmark uses three distinct metrics, Distinctiveness, Self-Consistency, and Correctness, as a technique for benchmarking the degree to which any generative text-to-image system provides multilingual parity to its training language in terms of tangible nouns. Liu et al. (2023a) propose the C3 benchmark to study cultural relevance and image quality and propose evaluation on six metrics: cultural appropriateness, object presence, object localization, semantic consistency, visual aesthetics, and cohesion. Unlike Saxon and Wang (2023), who focus on generating simple concepts through translation, Mittal et al. (2024) focus on prompts describing multiple elements in the generated image. They investigate bias in T2I models in 30 Indic languages on Stable Diffusion, Alt Diffusion, Midjourney, and Dalle3 and evaluate on four proposed metrics: Cyclic Language-Grounded Correctness, Language-Grounded Correctness, Image-Grounded Correctness, and Self-Consistency Across Languages. Bayramli et al. (2025) introduce the CultDiff benchmark, and show that diffusion models often fail to acurately generate cultural artifacts in architecture, clothing, and food.

Takeaways from §5.2. Although there has been increasing attention on evaluating large multimodal models for cultural awareness, there is a greater need for models trained with balanced multilingual and culture-specific data to ensure solid multilingual and cultural capabilities. Additionally, there is no standardized list of evaluation methods, with each study selecting the methods independently. Metrics for image output tasks, such as measuring cultural awareness in VQA and T2I models, are also not standardized and remain a significant future direction to explore. Overall, developing consistent metrics to test cultural awareness in text-to-image models remains a significant future direction that could be explored.

5.3 Art Forms Related Tasks

Art forms and paintings evoke different emotions across different cultures and have been used by the community to study the expression of emotions across cultures (Mohammad and Kiritchenko 2018; Achlioptas et al. 2021). There have been multiple

studies on art and generating art using Vision models recently. One of the main goals when studying and examining art forms is to match the objects in an image to their symbolic meaning. Zhang et al. (2023) create a dataset for art understanding deeply rooted in traditional Chinese culture; they address three tasks: identifying salient visual elements, matching elements with their symbolic meanings, and explanations for the conveyed message. Hamilton et al. (2021) create a Web application named MosAIc, which allows users to find pairs of semantically related artworks that span different cultures, media, and millennia; they use Conditional Image Retrieval, which combines visual similarity search with user-supplied filters or "conditions." To study the similarity between arts across cultures and evaluate cultural-transfer performance, Mohamed et al. (2022) creates a dataset of 80k artworks, with many artworks being annotated by multiple people in three languages. Fan, Wang, and Hodel (2023) create a multimodal knowledge graph linking visual entities and concepts associated with the entities. Zhang et al. (2024g) addresses the challenge of translating the nuanced symbolism in art, which involves interpreting complex cultural contexts, aligning cross-cultural symbols, and validating cultural acceptance. Ozaki et al. (2024) create a dataset of artworks and explanations in multiple languages with nuances and country-specific phrases.

5.4 Miscellaneous Tasks

The following papers introduce new tasks, including variants of image captioning, image classification, or somewhere in between, to better access the cultural understanding in Vision language models. Buettner et al. (2024) attempt to improve object recognition models to be more robust to objects from geographically diverse regions. M5 (Schneider and Sitaram 2024), for example, collects data for 12 languages to pair it with photos from the regions that speak those languages. The authors then create a benchmark for tasks such as visually grounded reasoning, VQA, visual natural Language inference, visio-linguistic outlier detection (VLOD), and captioning. They introduced new novel benchmarks, such as M5-VGR and M5-VLOD, including a new visio-linguistic outlier detection task. The images for M5 are sourced from the Dollar Street dataset (Rojas et al. 2022), comprising around 38K photos taken in 63 different regions or countries. These photos depict the lives of families, including their homes, neighborhoods, or everyday objects, in a culturally diverse way. There is no explicit paring between languages/cultures and images. Pappas et al. (2016) conduct a crowdsourcing experiment to annotate the sentiment score of visual concepts from 11 languages associated with 16,000 multilingual visual concepts. The MVSO dataset (Jou et al. 2015) is used as the source of visual concepts, and the photo-sharing service Flickr is used as the source of images. Zhang et al. (2024c) create a dataset that spans 30 countries and almost two centuries; their goal is to test if VLMs can identify cultural markers required to determine the time and place a photo was taken. On similar lines, Hsiao and Grauman (2021) provide a data-driven approach to identify specific cultural factors affecting the clothes people wear where they use news articles and vintage photos spanning a century to create a model that detects influence relationships between happenings in the world and people's choice of clothing. Li et al. (2022) construct a multimodal knowledge graph for classical Chinese poetry (PKG), in which the visual information of words in the poetry are incorporated for the task of poverty image retrieval. Zhang et al. (2024f) propose a preference-based reinforcement learning method that fine-tunes the vision models to distill the knowledge from both LLMs' reasoning and the aesthetic models to

better align the vision models with human aesthetic standards, which vary with culture. Khanuja et al. (2024) introduce a new task of translating images to make them culturally relevant by changing concepts in an image that varies with culture.

Takeaways from §5. Apart from takeaways mentioned in specific subsections, some papers model cultural change in images across time, an important aspect missed when using images directly from the Internet as a source. Most of the vision papers do assume that a language implies a culture, but they use the assumption that a geographical region corresponds to a culture. Automatic data scraping methods that rely on getting culture-specific images based on the language of the caption (from sources like Wikipedia) can lead to language culture bias, where multiple cultures sharing the same language may be merged into a single, undifferentiated culture.

6. Other Modalities and Culture

In this section, we include papers that look at cultural adaptation in other modalities, such as videos, audio, and so forth, tasks that do not fall under the text-only or vision-language (i.e., text+image) tasks, but have text (semantic content) as one of the components. We highlight major areas and representative papers in Figure 14.

6.1 Audio and Speech

While understanding the cultural context in music and speech may not always involve text (semantic content) as one of the component major components, music recommendations require understanding the culture-specific preferences of the user as well as the cultural context in the query (Weck et al. 2024). Moradi, Neophytou, and Farnadi (2024) study cultural biases in music recommendation systems and provide a method to improve fairness in music recommendation systems. Li et al. (2024h) argue that understanding the cultural context should be one of the goals that should be prioritized while building foundation models for music.

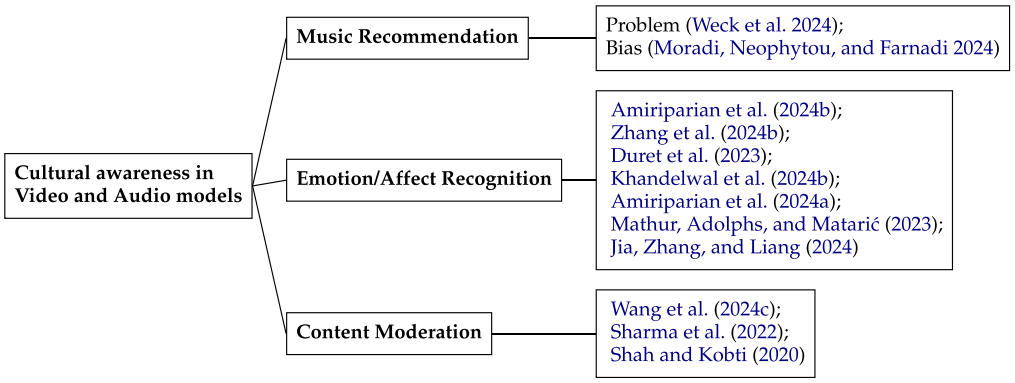


Figure 14
Areas explored for cultural adaptation in video and audio modalities, with representative examples. Music recommendation involves only the audio modality, emotion/affect recognition has been explored in both audio and video modalities, and content moderation has been explored for videos and memes.

One of the applications where understanding data streams such as speech along with semantic content becomes important is emotion recognition, as the expression of emotions varies across cultures. Belani and Flanigan (2022) study the relation between emotional expression and code-switching for Spanish. Amiriparian et al. (2024b) gather a comprehensive multilingual, multicultural speech emotion corpus with 37 datasets, 150,907 samples, and a total duration of 119.5 hours. Tran, Yin, and Soleymani (2023) demonstrate personalized and adapted speech encoders for continuous emotion recognition. Sapinski and Kaminska (2015) investigate emotion detection based on audio characteristics and the semantic content in tandem. Zhang et al. (2024b) propose a multimodal LLM-based multi-agent system designed for simulating human communication along with rich emotions expressed through speech and semantic content (text). As the research on textless speech-to-speech translation continues to grow, it is important to ensure that expressions and emotions are translated (and mapped) correctly across languages. Duret et al. (2023) propose a method to enhance expressivity transfer in textless speech-to-speech translation. Wunarso and Soelistio (2017) create a dataset for speech-emotion detection for Indonesian.

There have been a few papers that examine the presence of subculture within a border culture and collect data to understand cultural nuances. Javed et al. (2024) collect a dataset (INDICVOICES) of natural and spontaneous speech covering 16,237 speakers covering 145 Indian districts and 22 languages to capture the cultural, linguistic, and demographic diversity of India. SEACrowd (Lovenia et al. 2024) carries similar efforts and collects data for 1,000 Southeast Asia (SEA) languages spanning three modalities, with one of the goals being reducing cultural misrepresentation and flattening.

Takeaways from §6.1. While most relevant work in the audio domain focuses on emotion recognition and music recommendations, there has been a lack of studies that simultaneously model audio and text (semantic content) to understand the cultural context part from emotions. This capability could be useful for applications such as voice assistants. Work such as IndicVoices (Javed et al. 2024) and SEACrowd (Lovenia et al. 2024) are some initial efforts in the direction of collecting culturally diverse speech-text data. One of the limitations of datasets such as SEACrowd and IndicVoices is that they collect speech data in a controlled setup, typically by asking questions and recording responses, which may not accurately capture the nuances of everyday conversations.

6.2 Video

Concerning the video modality, most research is on task-specific cultural adaptation, focused mainly on emotion detection and content moderation. Cultural factors also affect personality (Walker et al. 2011) and how people interact in certain situations. Khan et al. (2020) create a multimodal dataset of peer-to-peer Hindi conversations to study the variance of personality with factors such as income and cultural orientation. Funk, Okada, and André (2024) study how culture affects the non-verbal features (such as facial expressions and tone) of the speakers during conversations. Amiriparian et al. (2024a) create a dataset for the cultural humor detection challenge, which focuses on cross-lingual and cross-cultural multimodal humor detection, as humor detection depends not only on words but also on gestures and facial expressions. SEWA DB (Kossaifi et al. 2019) is a dataset of conversations between people coming from different cultures during various social situations; the dataset contains videos annotated with facial landmarks, facial action units, various vocalizations, mirroring, and continuously valued valence, arousal, liking, agreement, and prototypic examples of (dis)liking. The

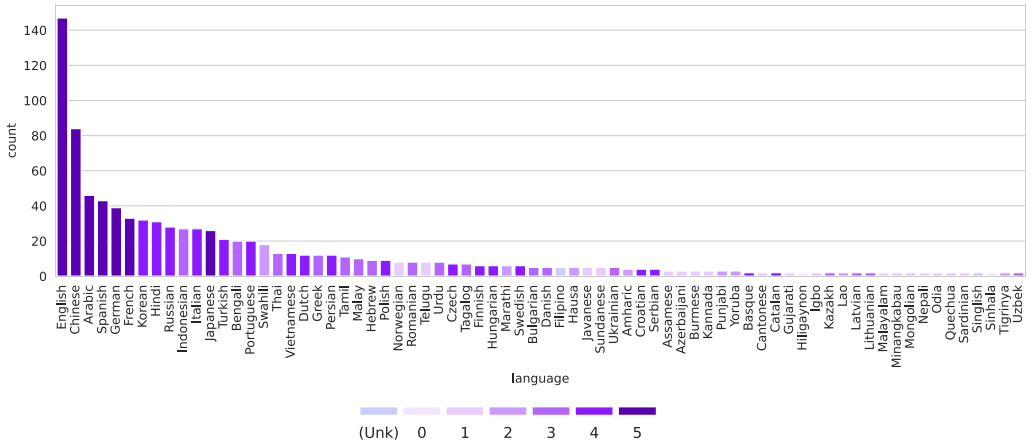
AVEC challenge through the years has looked at (and created datasets) for cross-cultural emotion detection (Ringeval et al. 2019). Detecting emotional cues (affect recognition) is an important part of HCI systems; Mathur, Adolphs, and Matarić (2023) study inter-cultural affect recognition models using videos of real-world dyadic interactions from six cultures. Zhao et al. (2022) create a multi-modal, multi-scene, multi-label emotional dialogue dataset from 56 television series capturing Chinese culture. Migon Favaretto et al. (2019) present a video analysis application to detect personality, emotion, and cultural aspects of pedestrians in video sequences, along with a visualizer of features (the features include elements of well-known frameworks such as Hofstede Cultural Dimensions). The work studying the effect of culture on emotional (body) gestures and the speech uttered during the gestures has been reviewed in Noroozi et al. (2018). Jia, Zhang, and Liang (2024) propose a multimodal strategy for emotion recognition based on facial expressions, voice tones, and transcripts from video clips. Liu, Courant, and Kalogeiton (2024) propose a multimodal approach based on transcripts, video-frames, and audio for detecting funny moments in video clips of television series. Bruno et al. (2019) present a case for embedding cultural knowledge into personal robots, as home activity recognition can be improved using cultural knowledge (Menicatti, Bruno, and Sgorbissa 2017). Rehm et al. (2009) provide guidelines for creating video recordings of multimodal interactions across cultures.

With the rise in online video-sharing platforms, multimodal hate speech detection has become an integral part of content moderation (Hee et al. 2024). Wang et al. (2024c) create a multilingual dataset of videos annotated for hatefulness, offensiveness, and normalcy and argue that the dataset provides a cross-cultural perspective on gender-based hate speech. The rise in memes as a source of information sharing (Shifman 2013) has also fueled interest in automatically detecting harmful and biased memes (Sharma et al. 2022). As memes marked as normal by one culture can be offensive to others, understanding the cultural context in the memes becomes necessary (Hegde et al. 2021). Shah and Kobti (2020) propose a methodology that uses situational and normative knowledge to detect fake news using text and images. Lyu et al. (2025) use GPT-4V for hate-speech detection in multimedia using cultural insights; they also study multimodal sentiment analysis in cultural context using GPT-4V.

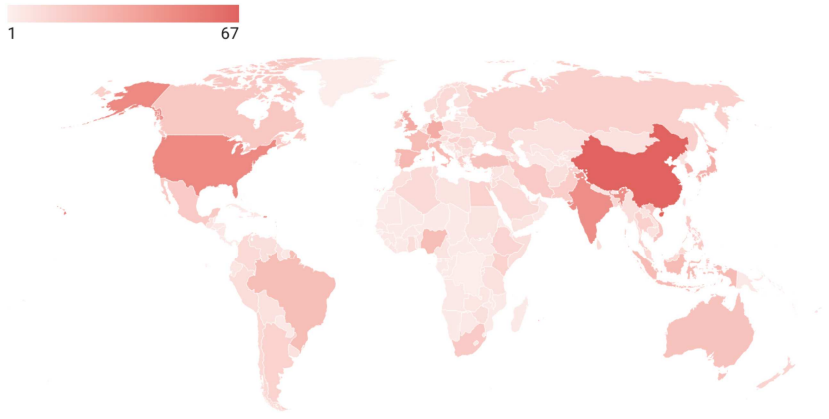
Takeaways from §6.2. The analysis of cultural nuances in day-to-day interactions between people and videos of day-to-day activities is often missing, while the major focus is memes, personality, and hate speech. Videos can be an important source of cultural information. Although there have been studies looking at information extraction from videos (An et al. 2023), extracting culture-specific information from videos can be an important next step. There has been a rise in video generation models (Ho et al. 2022), and how people perceive a video also depends on cultural background (Scott et al. 2015), so the video generation models should consider cultural context as one of the input features for the video generation model.

7. Language and Region Coverage

We manually annotate the languages and regions covered by benchmarks and evaluations presented in §4.2 and benchmarks in §5 and §6. For languages with multiple names (e.g., endonyms) or variations (e.g., dialects), we standardize them based on the conventions in Joshi et al. (2020). For regions, we visualize the distribution at the country level, including only countries explicitly mentioned by the authors in their papers. We choose country-level analysis since most papers emphasize national cultures.



(a) Distribution across languages



(b) Distribution across countries

Figure 15
Distribution of languages and the countries covered in model evaluations. (a) The colors represent the language resource classes from Joshi et al. (2020). The plot includes only languages that appear two or more times.

Sub-regional cultures are aggregated under their respective countries, and broader regions are excluded from our analysis.

Figure 15(a) presents the frequency distribution of languages used in the evaluations, showing only languages with a frequency of two or more. The colors indicate the language resource classes from Joshi et al. (2020), with darker colors (with higher numbers) representing higher-resource languages. As discussed in previous surveys (Liu, Gurevych, and Korhonen 2024; Adilazuarda et al. 2024), most studies collect data in English. Chinese, Spanish, and German are other high-resource languages observed frequently in cultural studies. Also, research on Korean, Indonesian, Bengali, and Swahili has been relatively active compared with other languages of their resource levels. Notably, for most languages classified as level 2 or below, there are at most seven studies, with the only exception of Swahili, underscoring the gap in research.

Figure 15(b) visualizes the target countries of the research on cultural language models.[8] Most of the studies focus on WEIRD (Western, Educated, Industrialized, Rich, and Democratic) countries (Henrich, Heine, and Norenzayan 2010), along with regions such as East Asia, Indonesia, and India. In contrast, countries in Africa, Central and South America, Eastern Europe, and Central Asia are significantly underrepresented. Even for studies including underrepresented nations, the diversity of data sources tends to be limited, often relying on global surveys like the WVS (Survey 2022) or publicly available online platforms such as Wikipedia. Additionally, the volume of data points across different regions frequently varies, further contributing to an uneven representation.

The scope of our analysis of cultural coverage is limited by the inability to account for macro-level regions such as continents and broad cultural groups with boundaries that do not align with national borders. In particular, while Arabic is one of the languages studied extensively, many of those studies broadly lump the cultural sphere together under the label of the "Arab world." Recognizing the challenges of defining the exact boundaries of cultures, researchers should nonetheless strive for the most accurate representation of the cultures they examine. Additionally, our analysis lacks finer-grained sub-regional or historical district breakdowns; this represents another notable gap in current cultural NLP research.

8. Broader Implications

8.1 Ethics of Cultural Alignment

The incorporation of cultural information into general-purpose foundational models is an important research direction. Whether the style or content of outputs of these models is intentional or accidental, they co-create our sense of meaning and identity and have an impact on shaping our collective knowledge (Lu, Kay, and McKee 2022). Lack of appropriate cultural representation can lead to several harms, including disparate access due to performance gaps, imposition of hegemonic classifications, violation of cultural values, misinformation, or stereotypes about cultures, misrepresentation of cultural experiences, and outright erasure of cultures (Prabhakaran, Qadri, and Hutchinson 2022; Kay, Kasirzadeh, and Mohamed 2024).

Communication across different cultures differs substantially; some are low context while others are highly contextual in the way they communicate, both in the real world (Liu 2016) and online (Würtz 2017). Imposition of one culture's communication style on another can lead to erasure and flattening of cultures. A mismatch or misalignment in these styles can lead to problems pertaining to intercultural communication like misunderstanding, distrust, and conflict. Therefore, it is crucial to be careful in approaches towards cultural alignment. Further, since using generative models for writing can impact the opinions of the users themselves (Jakesch et al. 2023), there are also broader questions about the systemic impact on users and society at large (Burton et al. 2024) that need to be taken into account during this process.

To tackle the problem of lack of cultural knowledge, several papers have tried to adapt existing models and incorporate features for other cultures across different tasks, hoping to improve performance on tasks requiring cultural knowledge. These

8 We generate the choropleth map with Datawrapper at `https://www.datawrapper.de/maps/choropleth-map`.

include tasks like Affect Detection (Neiberg, Laukka, and Elfenbein 2011), Offensive Language Detection (Zhou et al. 2023b), Humor Detection (Xie et al. 2023), and Recipe Adaptation (Cao et al. 2024b), among others. Across these efforts, there is an underlying assumption that since cultural knowledge is required for these tasks, cultural alignment will improve cross-cultural performance on them. However, while there are improvements on the performance on the datasets, it is unclear whether the improvement is due to the actual incorporation of cultural knowledge or due to surface level features in the datasets that the models are picking up. When evaluated with domain experts, models often fail to appropriately use cultural information. For instance, in experiments with comedians for co-creating humor, LLMs fail to produce non-bland or generic outputs, especially when text is about cultures other than the dominant ones embedded into the model (Mirowski et al. 2024) or when used for mental health support, LLMs fail to adapt based on the cultural background of the users and provide misaligned recommendations (Song et al. 2024).

There are also issues concerning what cultural information is available to encode. Since data for inclusion is primarily scraped from the Internet, which is a biased sample of what cultural information exists, it only captures some aspects of knowledge (Bender et al. 2021). The long tail of cultural information, which pertains to everyday tasks, is unspecified or not recorded and hence does not make it to the datasets. There are efforts to address this and benchmark the performance of models on everyday knowledge (Myung et al. 2024), but the area is largely ignored, making it a core limitation for both model designers and practitioners. Further, since large language models are increasingly being used as writing assistants and sources of synthetic data, which are biased samples in regards to the diversity of the content (Padmakumar and He 2023) and the values embedded therein (Wright et al. 2024; Santurkar et al. 2023), the data that is being used to train these models will constantly reinforce one set of values and lead to more biased models. This will lead to poorer representation of diverse cultural representation in model outputs, resulting in a potential cultural model collapse (Shumailov et al. 2024).

Beyond harms associated with non-inclusion or simplistic inclusion of cultures, there are also harms associated with explicit inclusion of a culture. Kirk et al. (2024) outline this by creating a taxonomy discussing the benefits and harms of personalizing language models. They show that while there are clear use cases of aligning language models such as the increased autonomy, empathy, and usefulness, one should be considerate of the often overlooked harms that such alignment can bring. The study shows how each benefit that personalization brings has a potential harm resulting from it, recommending that model designers and practitioners have to take these trade-offs into account when creating or deploying these models. For instance, the increased usefulness or empathy in models can lead to dependency on these models and contribute to their anthropomorphism. Similarly, at a societal level, adaptation to each culture can contribute to increased polarization and labor displacement.

Further, it is unclear what the correct approach is when the culture that needs to be included has fundamentally opposing values to the ones where models are usually created. When NLP research suggests alignment, it is typically associated with cultures that are not at odds with the value systems of Western nations. Further, there is also cultural information that is unsafe. For instance, medical advice from certain cultures challenges Western notions of medicine and advises against relying on it, instead promoting local forms of medicine, which can at times be harmful. Another related example is alignment to fringe or extremist communities. With adapted generative models, the potential for harm that they can cause would also be much higher when used for nefarious purposes

by malicious agents (Kaffee et al. 2023). Alignment in such scenarios can lead to unsafe behavior, thus bringing forward this trade-off between two desirable characteristics of model behavior.

Finally, there are also questions about whether building general-purpose systems suitable for all audiences is the right way forward (Gabriel 2020). In their work, Zhi-Xuan et al. (2024) highlight some assumptions made by current AI alignment efforts, namely, that human preferences (which is the dominant method of encoding values) are an adequate representation of human values and AI systems should be aligned with preferences of one or more humans to ensure that they behave safely and in alignment with our values. They challenge these assumptions, critiquing the normativity of expected utility theory as the dominant method for alignment of AI assistants. They argue that these systems should instead be domain-specific and aligned based on standards negotiated upon by the corresponding users and stakeholders, allowing for meeting diverse needs and co-existing in the presence of pluralistic values. In a cultural context, such a system would benefit multicultural societies where certain normative standards have been established.

8.2 Accelerating Social Science Research

There has been a lot of optimism and uptick in adoption from the social sciences towards AI systems aiding in research. They have been used in several fields, including psychology, sociology, political science, history, and many others, for different tasks. While the appropriateness and motivation for using them as an accurate representation of society is an ongoing discussion (Grossmann et al. 2023; Agnew et al. 2024), they certainly aid in performing several subtasks relevant for social science research (Ziems et al. 2024; Bail 2024). Across these fields, they serve a variety of different purposes. For some, they aid in the analysis of large volumes of content (Törnberg 2024), extracting dispositions from social media (Peters and Matz 2024) to make inferences about users' preferences, while for others it is simulation of responses of human samples for surveys (Argyle et al. 2023b; Manvi et al. 2024) or serving as agents for agent-based modeling methods for predicting hypothetical behavior (Horton 2023; Grossmann et al. 2023). They have also been used for interventions to existing ecosystems (Yang et al. 2024; Argyle et al. 2023a), trying to address existing issues like misinformation or polarity. Since most of these tasks rely on models appropriately reflecting people from different cultures faithfully, careful cultural alignment is a crucial part in this process. Machine learning models are trained to generalize and learn from abstractions in data. This can lead to flattening of identity (Wang, Morgenstern, and Dickerson 2024) or poor performance on and misrepresentation of non-majority cultures. Such an effect can cast doubt on inferences made on top of results extracted from these systems. Thus, cross-cultural alignment with robust human evaluation is imperative for reliable inferences made in the social science.

For achieving this alignment and embedding cultural information and knowledge, however, practitioners often use sources and literature from the social sciences (§2). Simultaneously, generative models are proposed as a means to replace human participants in surveys (Argyle et al. 2023b; Manvi et al. 2024). Such a loop can lead to a vicious circle, where cultures are misrepresented and cultural change is not incorporated. It has also been found that while group level can be simulated by these models to some extent, they fail to capture the diversity of human behavior (Cao et al. 2025), which is a core motivation behind survey approaches, and are prone to biases (Qu and Wang 2024). In light of these findings, social scientists need to be aware of sources of data used for

training these models, and the biases that may be embedded in them corresponding to the populations they are studying. Further, generative model designers need to be careful while culturally aligning these models to not over-rely on social science survey data, when not directly extracted from human participants from different cultures but rather, synthetically generated.

8.3 Human–Computer Interaction and Cultural Alignment

Another important aspect of cultural alignment is how people interact with the culturally aligned LLMs and the corresponding interaction patterns. Understanding these interaction patterns includes studying how cultural alignment for models affects their use in applications such as creating generative art (Zhang et al. 2024e), generating culturally relevant stories (Toro Isaza et al. 2023), professional communication, language acquisition (Zhao et al. 2024b), cross-cultural communication, and so forth. Aesthetic standards, expression, and the emotions invoked during interaction are different in different cultures (Section 5.3). Further, how people from different cultures perceive these models also differs (Liu et al. 2024d). As such, it is imperative to have faithful representation of cultural artifacts during model interactions and to measure them holistically.

Weidinger et al. (2023) propose a three-layered approach to evaluating this effect in AI systems: The first layer is their capability; the second is how the system affects human interaction with the system; and the last layer is understanding the impact of the system on the broader context in which it is embedded, such as society, the economy, and the natural environment. Using this framework to understand the embedding of cultural information in generative models from an HCI perspective, we find that most research has focused on building and evaluating the capability (cultural understanding and awareness). On the other hand, ensuring cultural inclusion during human interaction with the system and studying broader systemic impact has received very little attention. For instance, generative models are tested for the understanding of culture-specific references in a conversation and the ability of a model to produce (culturally) relevant responses, testing their capability. However, the risk of people being deceived, misled, or enraged by that output because of them being misaligned with cultures depends on factors such as the context in which an AI system is used, who uses it, and the features of an application. Such evaluation is rarely performed before deployment of current generative systems, and is imperative for safe deployment in cultures distant from the ones that models are biased towards. For instance, when conducting an ethnographic study of use of generative AI based text-to-image tools, Mim et al. (2024) find that these tools, while boosting speed, limited the creative explorations of artists. They fail to effectively process cultural references, linguistic nuances, or generate images representative of local styles, art, or architecture. This raises concerns about the imposition of the Western perspective, potential erasure, and possible colonization of the imagination of Global South practitioners. Further, since the usage of these models can impact the users themselves, these factors have to be studied in the context of population-level effects. For instance, political values and opinions from biased models affecting the opinions of users (Jakesch et al. 2023; Fisher et al. 2024) can lead to a shift in norms and values that a culture identifies itself with (Wagner et al. 2021).

An HCI-based perspective on cultural inclusivity in generative models would include adapting the models to the needs and expectations of culture and the intended applications (e.g., high risk vs. low risk). As the models become more general purpose, there needs to be a distinction between the tasks and applications that would require

culturally agnostic capabilities vs. those that would require culture-specific capabilities (Cetinic 2022). The HCI component should drive the data-collection for cultural alignment, as some cultures might be over-represented and while others might be misrepresented due to variance of technology access and expectations across different cultures. Participatory frameworks for co-designing these models and the data used for training them, involving stakeholders from the corresponding cultures, is one effective approach of addressing current gaps in culturally misaligned models (Birhane et al. 2022).

9. Pointers to Future Research

Expand Research on Low-resource Cultures and Languages. Research on low- and mid-resource cultures and languages has progressed but remains limited compared with high-resource counterparts, as discussed in sections 4.2.8 and 7. Thus, more efforts are needed to collect data and evaluate the language models on low-resource cultures and languages. For example, creating benchmarks for dialects—which capture unique aspects of local culture—is important, especially as many of these dialects are at risk of disappearing (Moseley 2010). In regions with low technology access (e.g., Sudan), involving native annotators becomes essential due to the limited available Web data. However, these methods are not scalable, underscoring the need for more research into scalable data collection for low-resource languages and cultures. Additionally, it is necessary to develop alignment methodologies that can perform effectively with relatively small datasets.

Vary Data Collection Strategies According to Target Culture. While collecting data for low-resource languages and cultures, considerable emphasis should be given to technology access of the cultures. Technology access determines how the data collection strategy should be varied. For example, although cultures like Estonian and Finnish are low-resource cultures, data collection strategies for that culture would involve scraping region-specific Web data, recruiting annotators, and running crowdsourced experiments to understand cultural preferences due to high penetration of technology (mobiles, Internet, etc.) in those cultures.[9] On the other hand, for cultures with low technology penetration (e.g., Sudan), the data collection would involve on-ground annotators talking to native people to collect data. The population of people following the culture would affect data collection, as some languages and cultural practices are spoken and followed by people in a restricted domain (Liu et al. 2022d). While the data collection strategies would vary across cultures, care must be taken to standardize the data (e.g., by having humans in the loop) to ensure equity of model performance across cultures, as both methods would lead to differences in the quality of data.

Approach Defining Cultural Boundaries with Caution. In language and vision research, culture is often represented through language or geographical regions, typically at the country level. However, countries do not always align with cultural boundaries (Bashkow 2004). For instance, Indonesia—one of the most ethnographically diverse nations globally—contains a wide array of local cultures not captured by a single national identity. Recent efforts aim to incorporate these local cultures into cultural benchmarks

9 Internet penetration statistics: `https://worldpopulationreview.com/country-rankings/internet-penetration-by-country`.

(Putri et al. 2024; Koto et al. 2024b, 2023), though such attempts remain limited to certain regions. Using language as a cultural proxy also presents challenges, as languages like English, Spanish, or Arabic can span multiple cultural contexts (Lee, Jung, and Oh 2023). Therefore, it is crucial to carefully define cultural boundaries when conducting cross-cultural research or developing culture-specific benchmarks. One effective approach to address this challenge is to engage in interdisciplinary collaboration with sociolinguistic researchers, who can provide deeper insights into the nuances of cultural and linguistic diversity.

Ensure Inclusive Cultural Representations. Even within the same region or cultural group, social values and norms can vary significantly based on demographics such as age, gender, and race (Weber and Urick 2017). Therefore, when constructing cultural datasets or benchmarks, it is essential to involve annotators with diverse demographic backgrounds, even within a single cultural group. Moreover, as cultural values and norms can vary between individuals, using annotators from a specific demographic group might not be fully representative of the culture. For instance, when gathering responses to commonsense questions like "What is a common school cafeteria food in your country?", relying on a small, homogeneous group of annotators can lead to incomplete or biased representations. A diverse and sizable pool of annotators is essential to capture a full range of perspectives. Additionally, evaluating the level of agreement among annotators can help determine if the *gold* answer truly reflects the culture context (Havaldar et al. 2024).

Develop LLMs That Can Adapt and Evolve with Cultural Change Over Time. Another important factor to consider is the dynamic nature of culture. As Naylor (1996) noted, no culture is static; people continually adapt to changes in their physical and sociocultural environment. Especially in today's globalized world, interactions between different cultural groups can quickly lead to the emergence and transformation of new cultural identities (Holton 2000). While some studies have examined historical cultures (Wei et al. 2024; Tang et al. 2024b), there remains a notable gap in research on how to adapt LLMs as cultures evolve. Addressing this challenge requires moving beyond static LLMs that only align with current cultural norms. Instead, LLMs should act as repositories for cultural preservation and adaptable systems that can respond to ongoing cultural transformations.

Alternative Image Data Collection to Mitigate Biases in Web Images. Most vision benchmarks rely on images sourced from the Web, as discussed in Section 2. However, Web images are susceptible to various biases like availability bias (different subjects, light conditions, locations, camera settings, and other features may be more likely to be uploaded on the Web than others), apprehension bias (people may pose and look differently when they know that they are being photographed), and negative set bias (Goldman and Tsotsos 2024). For example, certain subjects and locations are more likely to be uploaded online, and people may pose differently when photographed. These biases could result in omitting everyday objects and cultural concepts on the Web. To mitigate these biases, we could actively photograph culturally relevant concepts and objects with guidance from local residents and anthropology experts. Additionally, using frames from videos that document themes such as "a typical day in the life of a person with a specific identity" or content from regional television shows can help capture more realistic and broader cultural images.

Expand Cultural Evaluation Methods to Diverse Interaction Settings. As discussed in Section 4.2, culturally aware LLMs are mostly evaluated using MCQs. However, this approach has limitations, as it cannot fully capture the complexities of real-life human-AI interactions. MCQs primarily evaluate a predefined set of cultural knowledge and focus on explicit cultural norms. However, in real-world scenarios, human–AI communication involves natural dialogue, where LLMs need to interpret implicit cultural cues and generate culturally sensitive responses. One promising research direction is evaluating the long-form generation of LLMs, which recent studies have started to explore, as shown in Figure 6. However, most evaluations depend on human judgment or LLM-as-a-judge (Zheng et al. 2023) methods, underscoring a gap in culturally specific and robust automatic evaluation techniques. Therefore, further research is needed to develop reliable evaluation methods for assessing LLMs in natural, conversational settings, including long-form generation.

Balance Development of Culturally Specific LLMs and Comprehensive Universal LLMs. Currently, various techniques are used to culturally align LLMs, as discussed in Sections 4.1 and 5. Most works have focused on training and developing culture-specific LLMs, particularly for non-Western local cultures. However, there has been comparatively less emphasis on creating cross-cultural models capable of reasoning across diverse cultural contexts. Given the diverse cultural backgrounds of users, it is essential for LLMs to possess a comprehensive cultural knowledge encompassing all high- and low-resource cultures. Therefore, it is essential to seek a balance between developing culture-specific LLMs tailored to local needs and creating comprehensive cross-cultural LLMs that can serve a global audience.

Develop Culturally Aware LLMs from User Perspectives. Section 8.3 discusses current applications of culturally aware LLMs, such as generating culturally relevant art, storytelling, and facilitating cross-cultural interactions. However, there remains a gap in understanding how users interact with culturally aware LLMs from the user's perspective. This gap could be addressed through observational studies of user behavior in real-world scenarios. For instance, by observing when users are offended by an LLM's lack of cultural knowledge, we could gather insights for building safer, more culturally sensitive models. Additionally, studying interactions between multiple LLM agents and humans could reveal new applications, such as LLMs facilitating communication between individuals from diverse cultural backgrounds who speak different languages. Thus, observing real-world use cases from the user's perspective is important for developing practical, culturally aware LLMs.

10. Conclusion

This survey presented a comprehensive review of papers studying cultural inclusion in text-based and multimodal models. We surveyed recent research efforts toward cultural awareness in LLMs and have consolidated the efforts under various themes. We have defined cultural awareness in LLMs by leveraging definitions of culture in psychology and anthropology. We then discussed methodologies adopted for creating cross-cultural datasets, strategies for cultural inclusion in downstream tasks, and methodologies that have been used for benchmarking cultural awareness in LLMs. We also discussed several important topics, such as the role of HCI in cultural inclusion, the role of cultural alignment in accelerating social science research, and ethical issues related to cultural

inclusion. We hope this survey will serve as a useful reference for future research on cultural alignment in AI systems.

Acknowledgments

We want to thank the members of the CopeNLU lab for providing feedback on the survey. We would also like to thank the anonymous reviewers for providing constructive feedback on improving the paper's structure, content, and presentation. This research is partially funded by a Carlsberg Foundation Semper Ardens Accelerate grant under grant agreement no. CF22-1461, a DFF Sapere Aude research leader grant under grant agreement no. 0171-00034B, a DFF Research Project 1 under grant agreement No 9130-00092B, and is supported by the Pioneer Center for AI, DNRF grant no. P1. This work was supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT, no. RS-2024-00509258 and no. RS-2024-00469482, Global AI Frontier Lab).

References

A, Pranav and Isabelle Augenstein. 2020. 2kenize: Tying subword sequences for Chinese script conversion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7257–7272. https://doi.org/10.18653/v1/2020.acl-main.648

Abbasi, Mohammad Amin, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. 2023. PersianLLaMA: Towards building first Persian large language model. *ArXiv preprint*, abs/2312.15713. https://doi.org/10.21203/rs.3.rs-3789059/v1

Abdulhai, Marwa, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752. https://doi.org/10.18653/v1/2024.emnlp-main.982

Achlioptas, Panos, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J. Guibas. 2021. ArtEmis: Affective language for visual art. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, pages 11569–11579. https://doi.org/10.1109/CVPR46437.2021.01140

Acquaye, Christabel, 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. https://doi.org/10.18653/v1/2024.emnlp-main.532

Adilazuarda, Muhammad Farid, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. *ArXiv preprint*, abs/2403.15412. https://doi.org/10.18653/v1/2024.emnlp-main.882

Agarwal, Utkarsh, 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.

Agnew, William, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024*, pages 286:1–286:12. https://doi.org/10.1145/3613904.3642703

Ahmad, Ibrahim, 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. https://doi.org/10.18653/v1/2024.c3nlp-1.8

Akinade, Idris, Jesujoba Alabi, David 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. https://doi.org/10.18653/v1/2023.c3nlp-1.1

Al-Laith, Ali, Daniel Hershcovich, Jens Bjerring-Hansen, Jakob Ingemann Parby, Alexander Conroy, and Timothy R. Tangherlini. 2024. Noise, novels, numbers. A framework for detecting and categorizing noise in Danish and Norwegian literature. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3344–3354. https://doi.org/10.18653/v1/2024.emnlp-main.196

Alghamdi, Emad A., Reem I. Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. 2024. AraTrust: An evaluation of trustworthiness for LLMs in Arabic. *ArXiv preprint*, abs/2403.09017.

AlKhamissi, Badr, 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. https://doi.org/10.18653/v1/2024.acl-long.671

Almeida, Thales Sales, Hugo Abonizio, Rodrigo Nogueira, and Ramon Pires. 2024. Sabia-2: A new generation of Portuguese large language models. *ArXiv preprint*, abs/2403.09887.

Aloui, Manel, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. 101 billion Arabic words dataset. *ArXiv preprint*, abs/2405.01590.

Alyafeai, Zaid, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged Al-shaibani. 2024. CIDAR: Culturally relevant instruction dataset for Arabic. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12878–12901. https://doi.org/10.18653/v1/2024.findings-acl.764

Amadeus, Marcellus, William Alberto Cruz Castañeda, André Felipe Zanella, and Felipe Rodrigues Perche Mahlow. 2024. From pampas to pixels: Fine-tuning diffusion models for Gaúcho heritage. *ArXiv preprint*, abs/2401.05520.

Amiriparian, Shahin, Lukas Christ, Alexander Kathan, Maurice Gerczuk, Niklas Müller, Steffen Klug, Lukas Stappen, Andreas König, Erik Cambria, Björn Schuller, et al. 2024a. The MuSe 2024 multimodal sentiment analysis challenge: Social perception and humor recognition. *ArXiv preprint*, abs/2406.07753. https://doi.org/10.1145/3689062.3689088

Amiriparian, Shahin, Filip Packań, Maurice Gerczuk, and Björn W. Schuller. 2024b. ExHuBERT: Enhancing HuBERT through block extension and fine-tuning on 37 emotion datasets. *ArXiv preprint*, abs/2406.10275. https://doi.org/10.21437/Interspeech.2024-280

An, Siyu, Ye Liu, Haoyuan Peng, and Di Yin. 2023. VKIE: The application of key information extraction on video text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 532–540. https://doi.org/10.33745/ijzi.2023.v09i02.093

Ananthram, Amith, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: Diagnosing the Western cultural bias of large vision-language models in image understanding. *ArXiv preprint*, abs/2406.11665.

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*, pages 2425–2433. https://doi.org/10.1109/ICCV.2015.279

Arango Monnar, Ayme, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130. https://doi.org/10.18653/v1/2022.woah-1.12

Argyle, Lisa P., Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023a. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120. https://doi.org/10.1073/pnas.2311627120, PubMed: 37788311

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023b. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351. https://doi.org/10.1017/pan.2023.2

Arora, Arnav, 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. https://doi.org/10.18653/v1/2023.c3nlp-1.12

Arora, Shane, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2024. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. *ArXiv preprint*, abs/2406.17761.

Arvai, Joseph. 2013. Thinking, fast and slow, Daniel Kahneman, Farrar, Straus & Giroux. *Journal of Risk Research*, 16(10):1322–1324. https://doi.org/10.1080/13669877.2013.766389

Asai, Akari, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564. https://doi.org/10.18653/v1/2021.naacl-main.46

Azizov, Dilshod, Zain Muhammad Mujahid, Hilal AlQuabeh, Preslav Nakov, and Shangsong Liang. 2024. SAFARI: Cross-lingual bias and factuality detection in news media and news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12217–12231. https://doi.org/10.18653/v1/2024.findings-emnlp.712

Baek, Yujin, chaeHun Park, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Evaluating visual and cultural interpretation: The K-Viscuit benchmark with human-VLM collaboration. *ArXiv preprint*, abs/2406.16469.

Bai, Yuelin, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. 2024. COIG-CQIA: Quality is all you need for Chinese instruction fine-tuning. *ArXiv preprint*, abs/2403.18058.

Bail, Christopher A. 2024. Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121. https://doi.org/10.1073/pnas.2314021121, PubMed: 38722813

Balestrucci, Pier Felice, Silvia Casola, Soda Marem Lo, Valerio Basile, and Alessandro Mazzei. 2024. I'm sure you're a real scholar yourself: Exploring ironic content generation by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14480–14494. https://doi.org/10.18653/v1/2024.findings-emnlp.847

Bann, Eugene Y. and Joanna J. Bryson. 2013. Measuring cultural relativity of emotional valence and arousal using semantic clustering and Twitter. In *COGSCI 2013*, pages 1809–1814.

Bansal, Hritik, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1370. https://doi.org/10.18653/v1/2022.emnlp-main.88

Bao, Keqin, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yancheng Luo, Chong Chen, Fuli Feng, and Qi Tian. 2023. A bi-step grounding paradigm for large language models in recommendation systems. *ArXiv preprint*, abs/2308.08434.

Bashkow, Ira. 2004. A neo-Boasian conception of cultural boundaries. *American Anthropologist*, 106(3):443–458.

Basu, Abhipsa, R. Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 5113–5124. https://doi.org/10.1109/ICCV51070.2023.00474

Bayramli, Zahra, 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? *arXiv preprint arXiv:2502.08914*.

Beck, Tilman, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615. https://doi.org/10.18653/v1/2024.eacl-long.159

Bekesas, Wilson Roberto, Mauro Berimbau, Renato Vercesi Mader, Joana Angelica Pellerano, Viviane Riegel, and Joana Pellerano. 2018. CosmoCult Card Game: A methodological tool to understand the hybrid and peripheral cultural consumption of young people. *Open*

Library of Humanities, 4(1). https://doi.org/10.16995/olh.167

Belani, Ritu and Jeffrey Flanigan. 2022. Automatic identification of motivation for code-switching in speech transcripts. *ArXiv preprint*, abs/2212.08565.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. https://doi.org/10.1145/3442188.3445922

Benkler, Noam, Scott Friedman, Sonja Schmer-Galunder, Drisana Mosaphir, Vasanth Sarathy, Pavan Kantharaju, Matthew D. McLure, and Robert P. Goldman. 2022. Cultural value resonance in folktales: A transformer-based analysis with the world value corpus. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 209–218. https://doi.org/10.1007/978-3-031-17114-7_20

Benkler, Noam K., 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.

Bennett III, Robert H., Paul A. Fadil, and Robin T. Greenwood. 1994. Cultural alignment in response to strategic organizational change: New considerations for a change framework. *Journal of Managerial Issues*, pages 474–490.

Berger, Arthur Asa. 2004. *Deconstructing Travel: Cultural Perspectives on Tourism*. Rowman Altamira.

Bhakthavatsalam, Sumithra, Chloe Anastasiades, and Peter Clark. 2020. GenericsKB: A knowledge base of generic statements. *ArXiv preprint*, abs/2005.00660.

Bhatia, Mehar and Vered Shwartz. 2023. GD-COMET: A geo-diverse commonsense inference model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7993–8001. https://doi.org/10.18653/v1/2023.emnlp-main.496

Bhatt, Shaily 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. https://doi.org/10.18653/v1/2024.findings-emnlp.942

Bhutani, Mukul, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: A dataset of geo-culturally situated stereotypes. *ArXiv preprint*, abs/2403.05696. https://doi.org/10.18653/v1/2024.acl-short.75

Bień, Michał, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. 2020. RecipeNLG: A cooking recipes dataset for semi-structured text generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 22–28. https://doi.org/10.18653/v1/2020.inlg-1.4

Billah Nagoudi, El Moatez, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Inciarte, and Md Tawkat Islam Khondaker. 2023. JASMINE: Arabic GPT models for few-shot learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16721–16744. https://doi.org/10.18653/v1/2023.emnlp-main.1040

Birhane, Abeba, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8. https://doi.org/10.1145/3551624.3555290

Bisk, Yonatan, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):7432–7439. https://doi.org/10.1609/aaai.v34i05.6239

Bloom, B. S. 1956. *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*, 2nd edition. David McKay Company.

Bongini, Pietro, Federico Becattini, Andrew D. Bagdanov, and Alberto Del Bimbo. 2020. Visual question answering for cultural heritage. In *IOP Conference Series: Materials*

Science and Engineering, volume 949, page 012074. https://doi.org/10.1088/1757-899X/949/1/012074

Borah, Angana, Aparna Garimella, and Rada Mihalcea. 2024. Towards region-aware bias evaluation metrics. *ArXiv preprint*, abs/2406.16152.

Borenstein, Nadav, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2024. Investigating human values in online communities. 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 1607–1627. https://doi.org/10.18653/v1/2025.naacl-long.77

Bruno, Barbara, Carmine Tommaso Recchiuto, Irena Papadopoulos, Alessandro Saffiotti, Christina Koulouglioti, Roberto Menicatti, Fulvio Mastrogiovanni, Renato Zaccaria, and Antonio Sgorbissa. 2019. Knowledge representation for culturally competent personal robots: Requirements, design principles, implementation, and assessment. *International Journal of Social Robotics*, 11:515–538. https://doi.org/10.1007/s12369-019-00519-w

Buettner, Kyle, Sina Malakouti, Xiang Lorraine Li, and Adriana Kovashka. 2024. Incorporating geo-diverse knowledge into prompting for increased geographical robustness in object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13515–13524. https://doi.org/10.1109/CVPR52733.2024.01283

Burda-Lassen, Olena, Aman Chadha, Shashank Goswami, and Vinija Jain. 2024. How culturally aware are vision-language models? *ArXiv preprint*, abs/2405.17475.

Burton, Jason W., Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A. Bakker, Joshua A. Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, et al. 2024. How large language models can reshape collective intelligence. *Nature Human Behaviour*, 8(9):1643–1655. https://doi.org/10.1038/s41562-024-01959-9, PubMed: 39304760

Cahyawijaya, Samuel, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2024a. High-dimension human value representation in large language models. *ArXiv preprint*, abs/2404.07900.

Cahyawijaya, Samuel, Holy Lovenia, Fajri Koto, Rifki Afina Putri, Emmanuel Dave, Jhonson Lee, Nuur Shadieq, Wawan Cenggoro, Salsabil Maulana Akbar, Muhammad Ihza Mahendra, et al. 2024b. Cendol: Open instruction-tuned generative large language models for Indonesian languages. *ArXiv preprint*, abs/2404.06138. https://doi.org/10.18653/v1/2024.acl-long.796

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. https://doi.org/10.1126/science.aal4230, PubMed: 28408601

Cao, Jiahuan, Yongxin Shi, Dezhi Peng, Yang Liu, and Lianwen Jin. 2024a. C^3Bench: A comprehensive classical Chinese understanding benchmark for large language models. *ArXiv preprint*, abs/2405.17732.

Cao, Yong, 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.

Cao, Yong, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024b. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99. https://doi.org/10.1162/tacl_a_00634

Cao, Yong, Wenyan Li, Jiaang Li, Yifei Yuan, and Daniel Hershcovich. 2024c. Exploring visual culture awareness in GPT-4V: A comprehensive probing. *ArXiv preprint*, abs/2402.06015.

Cao, Yong, 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. https://doi.org/10.18653/v1/2025.naacl-long.162, PubMed: 39841050

Cao, Yong, 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. https://doi.org/10.18653/v1/2023.c3nlp-1.7

Cao, Yang Trista, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295. https://doi.org/10.18653/v1/2022.naacl-main.92

Casola, Silvia, Simona Frenda, Soda 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. https://doi.org/10.18653/v1/2024.acl-long.849

Center, Pew Research. 2022. *Pew Global Attitudes Survey*. Accessed: 2022.

Cetinic, Eva. 2022. The myth of culturally agnostic AI models. *ArXiv preprint*, abs/2211.15271.

CH-Wang, Sky, 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. https://doi.org/10.18653/v1/2023.emnlp-main.215

Chan, Alex J., José Luis Redondo García, Fabrizio Silvestri, Colm O'Donnel, and Konstantina Palla. 2023. Harmonizing global voices: Culturally-aware models for enhanced content moderation. *ArXiv preprint*, abs/2312.02401.

Chang, Chen-Chi, Ching-Yuan Chen, Hung-Shin Lee, and Chih-Cheng Lee. 2024. Benchmarking cognitive domains for LLMs: Insights from Taiwanese Hakka culture. *ArXiv preprint*, abs/2409.01556. https://doi.org/10.1109/O-COCOSDA64382.2024.10800594

Chen, Andong, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024a. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. *ArXiv preprint*, abs/2408.09945.

Chen, Lichang, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024b. AlpaGasus: Training a better Alpaca with fewer data. In the *Twelfth International Conference on Learning Representations*.

Chen, Po-Heng, Sijia Cheng, Wei-Lin Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024c. Measuring Taiwanese Mandarin language understanding. *arXiv:2403.20180*.

Cheng, Myra, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532. https://doi.org/10.18653/v1/2023.acl-long.84

Chiu, Yu Ying, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. CulturalTeaming: AI-assisted interactive red-teaming for challenging LLMs' (lack of) multicultural knowledge. *ArXiv preprint*, abs/2404.06664.

Cho, Jaemin, Abhay Zala, and Mohit Bansal. 2023. DALL-EVAL: Probing the reasoning skills and social biases of text-to-image generation models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 3020–3031. https://doi.org/10.1109/ICCV51070.2023.00283

Choenni, Rochelle, 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. https://doi.org/10.18653/v1/2024.acl-long.803

Clark, Jonathan H., Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470. https://doi.org/10.1162/tacl_a_00317

Clarke, Christopher, Roland Daynauth, Jason Mars, Charlene Wilkinson, and Hubert Devonish. 2024. GuyLingo: The Republic of Guyana Creole corpora. In *Proceedings of the 2024 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 792–798. https://doi.org/10.18653/v1/2024.naacl-short.70

Conia, Simone, 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. https://doi.org/10.18653/v1/2024.emnlp-main.914

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Dammu, Preetam Prabhu Srikar, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanushree Mitra. 2024. "They are uncultured": Unveiling covert harms and social threats in LLM generated conversations. *ArXiv preprint*, abs/2405.05378. https://doi.org/10.18653/v1/2024.emnlp-main.1134

Das, Dipto, 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. https://doi.org/10.18653/v1/2023.c3nlp-1.8

Davani, Aida Mostafazadeh, Sagar Gubbi, Sunipa Dev, Shachi Dave, and Vinodkumar Prabhakaran. 2024. GeniL: A multilingual dataset on generalizing language. *ArXiv preprint*, abs/2404.05866.

Davis, Ernest and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103. https://doi.org/10.1145/2701413

De Angelis, Luigi, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1166120. https://doi.org/10.3389/fpubh.2023.1166120, PubMed: 37181697

Deas, Nicholas, Elsbeth Turcan, Iván Pérez Mejía, and Kathleen McKeown. 2024. MASIVE: Open-ended affective state identification in English and Spanish. *ArXiv preprint*, abs/2407.12196. https://doi.org/10.18653/v1/2024.emnlp-main.1139

Dehghan, Somaiyeh and Berrin Yanıkoğlu. 2024. Multi-domain hate speech detection using dual contrastive learning and paralinguistic features. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11745–11755.

Deng, Juntao, Xu Cao, and Bingqi Cheng. 2024. Research on generating cultural relic images based on a low-rank adaptive diffusion model. In *Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, pages 629–634. https://doi.org/10.1145/3675417.3675521

Deshpande, Ameet, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in ChatGPT: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270. https://doi.org/10.18653/v1/2023.findings-emnlp.88

Deshpande, Awantee, 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. https://doi.org/10.18653/v1/2022.woah-1.7

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

de Wynter, Adrian, Ishaan Watts, Nektar Ege Altintoprak, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Lena Baur, Samantha Claudet, Pavel Gajdusek, Can Gören, et al. 2024. RTP-LX: Can LLMs

evaluate toxicity in multilingual scenarios? *ArXiv preprint*, abs/2404.14397.

Ding, Bosheng, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1639–1657. https://doi.org/10.18653/v1/2022.acl-long.115

Du, Xinrun, Zhouliang Yu, Songyang Gao, Ding Pan, Yuyang Cheng, Ziyang Ma, Ruibin Yuan, Xingwei Qu, Jiaheng Liu, Tianyu Zheng, et al. 2024. Chinese tiny LLM: Pretraining a Chinese-centric large language model. *ArXiv preprint*, abs/2404.04167.

Dubois, Yann, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaFarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Duret, Jarod, Benjamin O'Brien, Yannick Estève, and Titouan Parcollet. 2023. Enhancing expressivity transfer in textless speech-to-speech translation. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. https://doi.org/10.1109/ASRU57964.2023.10389647

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

Dwivedi, Ashutosh, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931. https://doi.org/10.18653/v1/2023.emnlp-main.428

Elmadany, AbdelRahim, Ife Adebara, and Muhammad Abdul-Mageed. 2024. Toucan: Many-to-many translation for 150 African language pairs. *ArXiv preprint*, abs/2407.04796. https://doi.org/10.18653/v1/2024.findings-acl.781

Ember, Carol R. 2009. *Cross-cultural Research Methods*. Rowman Altamira.

Ennen, Philipp, Po-Chun Hsu, Chan-Jan Hsu, Chang-Le Liu, Yen-Chen Wu, Yin-Hsiang Liao, Chin-Tung Lin, Da-Shan Shiu, and Wei-Yun Ma. 2023. Extending the pre-training of bloom for improved support of traditional Chinese: Models, methods and results. *arXiv preprint arXiv:2303.04715*.

Ersoy, Asım, Gerson Vizcarra, Tahsin Mayeesha, and Benjamin Muller. 2023. In what languages are generative language models the most formal? Analyzing formality distribution across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2650–2666. https://doi.org/10.18653/v1/2023.findings-emnlp.175

España-Bonet, Cristina and Alberto Barrón-Cedeño. 2022. The (undesired) attenuation of human biases by multilinguality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2056–2077. https://doi.org/10.18653/v1/2022.emnlp-main.133

Esser, Patrick, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*.

Etxaniz, Julen, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. BertaQA: How much do language models know about local culture? *ArXiv preprint*, abs/2406.07302.

Faisal, Fahim, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: A NLP benchmark for dialects, varieties, and closely-related languages. *ArXiv preprint*, abs/2403.11009. https://doi.org/10.18653/v1/2024.acl-long.777

Fan, Tao, Hao Wang, and Tobias Hodel. 2023. CICHMKG: A large-scale and comprehensive Chinese intangible cultural heritage multimodal knowledge graph. *Heritage Science*, 11:1–18. https://doi.org/10.1186/s40494-023-00927-2

Fei, Nanyi, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu,

Ruihua Song, Xin Gao, Tao Xiang, et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094. https://doi.org/10.1038/s41467-022-30761-2, PubMed: 35655064

Feng, Shangbin, 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. https://doi.org/10.18653/v1/2024.emnlp-main.239

Feng, Shangbin, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024b. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. *ArXiv preprint*, abs/2406.15951. https://doi.org/10.18653/v1/2024.emnlp-main.240

Fisher, Jillian, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W. Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. 2024. Biased AI can influence political decision-making. *ArXiv preprint*, abs/2410.06415.

Forbes, Maxwell, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670. https://doi.org/10.18653/v1/2020.emnlp-main.48

Fort, Karen, 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, Javier Torroba Marchante, Shilin Xie, Sergio E. Zanotto, and Aurélie Névéol. 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.

Fung, Yi, 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. https://doi.org/10.18653/v1/2023.emnlp-main.941

Fung, Yi, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & LM benchmarking. *ArXiv preprint*, abs/2402.09369.

Funk, Marius, Shogo Okada, and Elisabeth André. 2024. Multilingual dyadic interaction corpus NoXi+ J: Toward understanding Asian-European non-verbal cultural characteristics and their influences on engagement. *ArXiv preprint*, abs/2409.13726. https://doi.org/10.1145/3678957.3685757

Furst, Edward J. 1981. Bloom's taxonomy of educational objectives for the cognitive domain: Philosophical and educational issues. *Review of Educational Research*, 51(4):441–453. https://doi.org/10.3102/00346543051004441

Gabriel, Iason. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437. https://doi.org/10.1007/s11023-020-09539-2

Ganguli, Deep, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *ArXiv preprint*, abs/2209.07858.

Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

Ghahroodi, Omid, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam Challenge (PersianMMLU): Is your LLM truly wise to the Persian language? *ArXiv preprint*, abs/2404.06644.

Ghosh, Dhruba, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in*

Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023.
Goldman, Josh and John K. Tsotsos. 2024. Statistical challenges with dataset construction: Why you will never have enough images. *ArXiv preprint*, abs/2408.11160.
Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464. https://doi.org/10.1037//0022-3514.74.6.1464, PubMed: 9654756
Grossmann, Igor, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. AI and the transformation of social science research. *Science*, 380(6650):1108–1109. https://doi.org/10.1126/science.adi1778, PubMed: 37319216
Gupta, Prannaya, Le Qi Yau, Hao Han Low, I.-Shiang Lee, Hugo Maximus Lim, Yu Xin Teoh, Jia Hng Koh, Dar Win Liew, Rishabh Bhardwaj, Rajat Bhardwaj, and Soujanya Poria. 2024. WalledEval: A comprehensive safety evaluation toolkit for large language models. *ArXiv preprint*, abs/2408.03837. https://doi.org/10.18653/v1/2024.emnlp-demo.42
Haerpfer, Christian W. and Kseniya Kizilova. 2012. The World Values Survey. *The Wiley-Blackwell Encyclopedia of Globalization*, pages 1–5. https://doi.org/10.1002/9780470670590.wbeog954
Hall, Edward T. 1976. *Beyond Culture*. Anchor.
Halpern, Ben. 1955. The dynamic elements of culture. *Ethics*, 65:235–249. https://doi.org/10.1086/291013
Hamilton, Mark, Stephanie Fu, Mindren Lu, Johnny Bui, Darius Bopp, Zhenbang Chen, Felix Tran, Margaret Wang, Marina Rogers, Lei Zhang, et al. 2021. MosAIc: Finding artistic connections across culture with conditional image retrieval. In *NeurIPS 2020 Competition and Demonstration Track*, pages 133–155.
Hämmerl, Katharina, Bjoern Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. 2023. Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2137–2156. https://doi.org/10.18653/v1/2023.findings-acl.134
Han, HyoJung, Jordan Boyd-Graber, and Marine Carpuat. 2023. Bridging background knowledge gaps in translation with automatic explicitation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9718–9735. https://doi.org/10.18653/v1/2023.emnlp-main.603
Han, Xu, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250. https://doi.org/10.1016/j.aiopen.2021.08.002
Hartvigsen, Thomas, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326. https://doi.org/10.18653/v1/2022.acl-long.234
Hasan, Md Arid, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N. Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2024. NativQA: Multilingual culturally-aligned natural query for LLMs. *ArXiv preprint*, abs/2407.09823.
Havaldar, Shreya, 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. https://doi.org/10.18653/v1/2024.naacl-long.12
Havaldar, Shreya, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. 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. https://doi.org/10.18653/v1/2023.wassa-1.19

Hayati, Shirley Anugrah, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2024. How far can we extract diverse perspectives from large language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5336–5366. https://doi.org/10.18653/v1/2024.emnlp-main.306

He, Ruiqi, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, and Naihao Deng. 2024. Chumor 1.0: A truly funny and challenging Chinese humor understanding dataset from Ruo Zhi Ba. *ArXiv preprint*, abs/2406.12754.

Hee, Ming Shan, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Recent advances in hate speech moderation: Multimodality and the role of large models. *ArXiv preprint*, abs/2401.16727. https://doi.org/10.18653/v1/2024.findings-emnlp.254

Hegde, Siddhanth U., Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Sathiyaraj Thangasamy, B. Bharathi, and Bharathi Raja Chakravarthi. 2021. Do images really do the talking? Analysing the significance of images in Tamil troll meme classification. *ArXiv preprint*, abs/2108.03886.

Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021*, OpenReview.net.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3):61–83. https://doi.org/10.1017/S0140525X0999152X, PubMed: 20550733

Hershcovich, Daniel, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013. https://doi.org/10.18653/v1/2022.acl-long.482

Heylighen, Francis and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 4(1).

Ho, Jonathan, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *ArXiv preprint*, abs/2210.02303.

Hobson, David G., 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. https://doi.org/10.18653/v1/2024.emnlp-main.723

Hofstede, Geert. 2005. Culture's recent consequences. In *Designing for Global Markets 7, IWIPS 2005, Bridging Cultural Differences, Proceedings of the Seventh International Workshop on Internationalisation of Products and Systems*, pages 3–4.

Hofstede, Geert, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and Organizations: Software of the Mind*, third edition. McGraw-Hill Professional.

Holton, Robert. 2000. Globalization's cultural consequences. *The Annals of the American Academy of Political and Social Science*, 570(1):140–152. https://doi.org/10.1177/0002716200570001011

Horton, John J. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research. https://doi.org/10.3386/w31122

Hovy, Dirk and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602. https://doi.org/10.18653/v1/2021.naacl-main.49

Hsiao, Wei-Lin and Kristen Grauman. 2021. From culture to clothing: Discovering the world events behind a century of fashion images. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, pages 1046–1055. https://doi.org/10.1109/ICCV48922.2021.00110

Hsu, Chan Jan, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da shan Shiu. 2023. Advancing the evaluation of traditional Chinese language models: Towards a comprehensive benchmark suite. *arXiv preprint arXiv:2309.08448*.

Hu, Songbo, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023a. 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. https://doi.org/10.1162/tacl_a_00609

Hu, Tianyi, 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. https://doi.org/10.18653/v1/2024.emnlp-main.61

Hu, Yushi, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023b. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1–6, 2023*, pages 20349–20360. https://doi.org/10.1109/ICCV51070.2023.01866

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. https://doi.org/10.18653/v1/2024.naacl-long.450

Huang, Kaiyi, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Huang, Yufei 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.

Hung, Chia Chien, Anne Lauscher, Ivan Vulić, Simone Ponzetto, and Goran Glavaš. 2022. Multi2WOZ: A robust multilingual dataset and conversational pretraining for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3687–3703. https://doi.org/10.18653/v1/2022.naacl-main.270

Hwang, EunJeong, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919. https://doi.org/10.18653/v1/2023.findings-emnlp.393

Ignat, Oana, Gayathri Ganesh Lakshmy, and Rada Mihalcea. 2024. Cross-cultural inspiration detection and analysis in real and LLM-generated social media data. *ArXiv preprint*, abs/2404.12933.

Imai, Mutsumi, Junko Kanero, and Takahiko Masuda. 2016. The relation between language, culture, and thought. *Current Opinion in Psychology*, 8:70–77. https://doi.org/10.1016/j.copsyc.2015.10.011, PubMed: 29506807

ImaniGooghari, Ayyoob, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117. https://doi.org/10.18653/v1/2023.acl-long.61

Ito, Eiji, Gordon J. Walker, and Haidong Liang. 2014. A systematic review of non-Western and cross-cultural/national leisure research. *Journal of Leisure Research*, 46(2):226–239. https://doi.org/10.1080/00222216.2014.11950322

Jahoda, Gustav and Harry McGurk. 1974. Pictorial depth perception: A developmental study. *British Journal of Psychology*, 65(1):141–149. https://doi.org/10.1111/j.2044-8295.1974.tb02780.x, PubMed: 4822770

Jakesch, Maurice, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23–28, 2023*, pages 111:1–111:15. https://doi.org/10.1145/3544548.3581196

Jang, Dongyeop, Tae-Rim Yun, Choong-Yeol Lee, Young-Kyu Kwon, and Chang-Eop Kim. 2023. GPT-4 can pass the Korean national licensing examination for Korean medicine doctors. *PLOS Digital Health*, 2(12):e0000416. https://doi.org/10.1371/journal.pdig.0000416, PubMed: 38100393

Javed, Tahir, 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, Pratyush Kumar, et al. 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. https://doi.org/10.18653/v1/2024.findings-acl.639

Jeong, Younghoon, 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. https://doi.org/10.18653/v1/2022.emnlp-main.744

Jha, Akshita, Aida Mostafazadeh Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870. https://doi.org/10.18653/v1/2023.acl-long.548

Jha, Akshita, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. 2024. ViSAGe: A global-scale analysis of visual stereotypes in text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12333–12347. https://doi.org/10.18653/v1/2024.acl-long.667

Jia, Jiehui, Huan Zhang, and Jinhua Liang. 2024. Bridging discrete and continuous: A multimodal strategy for complex emotion detection. *ArXiv preprint*, abs/2409.07901.

Jiang, Ming and Mansi Joshi. 2024. CPopQA: Ranking cultural concept popularity by LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 615–630. https://doi.org/10.18653/v1/2024.naacl-short.52

Jin, Jiho, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524. https://doi.org/10.1162/tacl_a_00661

Jinnai, Yuu. 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. https://doi.org/10.18653/v1/2024.c3nlp-1.5

Johnson, Rebecca L., Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an American accent: Value conflict in GPT-3. *ArXiv preprint*, abs/2203.07785.

Jones, Ruth and Ann Irvine. 2013. The (un)faithful machine translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–101.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Jou, Brendan, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. 2015. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 159–168. https://doi.org/10.1145/2733373.2806246

Kabra, Anubha, 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. https://doi.org/10.18653/v1/2023.findings-acl.525

Kadaoui, Karima, Samar Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. TARJAMAT: Evaluation of Bard and ChatGPT on machine translation of ten Arabic varieties. In *Proceedings of ArabicNLP 2023*, pages 52–75. https://doi.org/10.18653/v1/2023.arabicnlp-1.6

Kaffee, Lucie Aimée, Arnav Arora, Zeerak Talat, and Isabelle Augenstein. 2023. Thorny roses: Investigating the dual use dilemma in natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13977–13998. https://doi.org/10.18653/v1/2023.findings-emnlp.932

Kannen, Nithish, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. *ArXiv preprint*, abs/2407.06863.

Kantharuban, Anjali, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245. https://doi.org/10.18653/v1/2023.findings-emnlp.481

Kasneci, Enkelejda, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274. https://doi.org/10.1016/j.lindif.2023.102274

Kay, Jackie, Atoosa Kasirzadeh, and Shakir Mohamed. 2024. Epistemic injustice in generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:684–697. https://doi.org/10.1609/aies.v7i1.31671

Keleg, Amr 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. https://doi.org/10.18653/v1/2023.findings-acl.389

Khan, Shahid Nawaz, Maitree Leekha, Jainendra Shukla, and Rajiv Ratn Shah. 2020. Vyaktitv: A multimodal peer-to-peer Hindi conversations based dataset for personality assessment. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, pages 103–111. https://doi.org/10.1109/BigMM50055.2020.00024

Khandelwal, Khyati, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. 2024. Indian-BhED: A dataset for measuring India-centric biases in large language models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 231–239. https://doi.org/10.1145/3677525.3678666

Khanuja, Simran, 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. https://doi.org/10.18653/v1/2024.emnlp-main.573

Kim, Boseop, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, et al. 2021. What changes can large-scale language models bring? Intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424. https://doi.org/10.18653/v1/2021.emnlp-main.274

Kim, Eunsu, 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.

Kim, Haven, Jongmin Jung, Dasaem Jeong, and Juhan Nam. 2024b. K-pop lyric translation: Dataset, analysis, and neural-modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9974–9987.

Kim, Jaehong, Chaeyoon Jeong, Seongchan Park, Meeyoung Cha, and Wonjae Lee. 2024c. 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. https://doi.org/10.18653/v1/2024.findings-acl.963

Kim, JiWoo, Yunsu Kim, and JinYeong Bak. 2024. KpopMT: Translation dataset with terminology for Kpop fandom. *ArXiv preprint*, abs/2407.07413. https://doi.org/10.18653/v1/2024.loresmt-1.3

Kim, Jun Seong, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. 2025. When Tom eats kimchi: Evaluating cultural bias 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. https://doi.org/10.18653/v1/2025.c3nlp-1.11

Kirk, Hannah Rose, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392. https://doi.org/10.1038/s42256-024-00820-y

Kiulian, Artur, 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.

Köpf, Andreas, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. OpenAssistant conversations - Democratizing large language model alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Korre, Katerina, Arianna Muti, and Alberto Barrón-Cedeño. 2024. The challenges of creating a parallel multilingual hate speech corpus: An exploration. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15842–15853.

Kossaifi, Jean, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, et al. 2019. SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040. https://doi.org/10.1109/TPAMI.2019.2944808, PubMed: 31581074

Koto, Fajri, 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. https://doi.org/10.18653/v1/2023.emnlp-main.760

Koto, Fajri, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024a. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5622–5640. https://doi.org/10.18653/v1/2024.findings-acl.334

Koto, Fajri, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024b. IndoCulture: Exploring geographically-influenced cultural commonsense reasoning across eleven Indonesian provinces. *ArXiv preprint*, abs/2404.01854. https://doi.org/10.1162/tacl_a_00726

Koufakou, Anna, Elijah Nieves, and John Peller. 2024. Towards a new benchmark for emotion detection in NLP: A unifying framework of recent corpora. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 196–206. https://doi.org/10.18653/v1/2024.genbench-1.13

Kunda, Maithilee and Irina Rabkina. 2020. Creative captioning: An AI grand challenge based on the Dixit board game. *ArXiv preprint*, abs/2010.00048.

Kuznetsova, Alina, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981. https://doi.org/10.1007/s11263-020-01316-z

Kwok, Louis, Michal Bravansky, and Lewis D. Griffin. 2024. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *arXiv preprint arXiv:2408.06929*.

Ladhak, Faisal, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? A case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219. https://doi.org/10.18653/v1/2023.eacl-main.234

Lahoti, Preethi, 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. https://doi.org/10.18653/v1/2023.emnlp-main.643

Le, Thang and Anh Luu. 2023. A parallel corpus for Vietnamese central-northern dialect text transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13839–13855. https://doi.org/10.18653/v1/2023.findings-emnlp.925

Lee, Hwaran, 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. https://doi.org/10.18653/v1/2023.acl-industry.21

Lee, Jiyoung, 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. *ArXiv preprint*, abs/2402.13605. https://doi.org/10.18653/v1/2024.findings-acl.666

Lee, Nayeon, 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. https://doi.org/10.18653/v1/2024.naacl-long.236

Lee, Nayeon, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46. https://doi.org/10.18653/v1/2023.c3nlp-1.5

Leeb, Felix and Bernhard Schölkopf. 2024. A diverse multilingual news headlines dataset from around the world. 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 647–652. https://doi.org/10.18653/v1/2024.naacl-short.55

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

Li, Baiqi, Zhiqiu Lin, Deepak Pathak, Jiayao Emily Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024a. GenAI-Bench: A holistic benchmark for compositional text-to-visual generation. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*. https://doi.org/10.1109/CVPRW63382.2024.00538

Li, Cheng, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024b. CultureLLM: Incorporating cultural differences into large language models. *ArXiv preprint*, abs/2402.10946.

Li, Chengxi, Kai Fan, Jiajun Bu, Boxing Chen, Zhongqiang Huang, and Zhi Yu. 2023a.

Translate the beauty in songs: Jointly learning to align melody and translate lyrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 27–39. https://doi.org/10.18653/v1/2023.findings-emnlp.3

Li, Cheng, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024c. CulturePark: Boosting cross-cultural understanding in large language models. *ArXiv preprint*, abs/2405.15145.

Li, Chong, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024d. X-Instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 546–566. https://doi.org/10.18653/v1/2024.findings-acl.30

Li, Huihan, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024e. CULTURE-GEN: Revealing global cultural perception in language models through natural language prompting. *ArXiv preprint*, abs/2404.10199.

Li, Haonan, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024f. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11260–11285. https://doi.org/10.18653/v1/2024.findings-acl.671

Li, Jialin, Junli Wang, Junjie Hu, and Ming Jiang. 2024g. How well do LLMs identify cultural unity in diversity? In *First Conference on Language Modeling*.

Li, Oliver, Mallika Subramanian, Arkadiy Saakyan, Sky C. H.-Wang, and Smaranda Muresan. 2023b. NormDial: A Comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15732–15744. https://doi.org/10.18653/v1/2023.emnlp-main.974

Li, Wei, Shutan Huang, and Yanqiu Shao. 2024. An unsupervised framework for adaptive context-aware simplified-traditional Chinese character conversion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1318–1326.

Li, Wei, Ren Ma, Jiang Wu, Chenya Gu, Jiahui Peng, Jinyang Len, Songyang Zhang, Hang Yan, Dahua Lin, and Conghui He. 2024h. FoundaBench: Evaluating Chinese fundamental knowledge capabilities of large language models. *ArXiv preprint*, abs/2404.18359.

Li, Wenjun, Ying Cai, Ziyang Wu, Wenyi Zhang, Yifan Chen, Rundong Qi, Mengqi Dong, Peigen Chen, Xiao Dong, Fenghao Shi, et al. 2024i. A survey of foundation models for music understanding. *ArXiv preprint*, abs/2409.09601.

Li, Wenyan, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024j. 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. https://doi.org/10.18653/v1/2024.emnlp-main.1063

Li, Yizhi, Ge Zhang, Xingwei Qu, Jiali Li, Zhaoqun Li, Zekun Wang, Hao Li, Ruibin Yuan, Yinghao Ma, Kai Zhang, et al. 2024k. CIF-Bench: A Chinese instruction-following benchmark for evaluating the generalizability of large language models. *ArXiv preprint*, abs/2402.13109. https://doi.org/10.18653/v1/2024.findings-acl.739

Li, Yuqing, Yuxin Zhang, Bin Wu, Ji-Rong Wen, Ruihua Song, and Ting Bai. 2022. A multi-modal knowledge graph for classical Chinese poetry. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2318–2326. https://doi.org/10.18653/v1/2022.findings-emnlp.171

Li, Zhi 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. https://doi.org/10.18653/v1/2023.emnlp-main.18

Liang, Ke, Chu-Ren Huang, and Xin-Lan Jiang. 2024. From text to historical ecological knowledge: The construction and application of the Shan Jing knowledge base. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7521–7530.

Lin, Tsung Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Proceedings, Part V 13*, pages 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Lin, Xi Victoria, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *ArXiv preprint*, abs/2112.10668. https://doi.org/10.18653/v1/2022.emnlp-main.616

Lin, Yen Ting and Yun-Nung Chen. 2023. Taiwan LLM: Bridging the linguistic divide with a culturally aligned language model. *ArXiv preprint*, abs/2311.17487.

Liu, Bingshuai, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. 2023a. On the cultural gap in text-to-image generation. *ArXiv preprint*, abs/2307.02971. https://doi.org/10.3233/FAIA240581

Liu, Chen, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022a. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086. https://doi.org/10.18653/v1/2022.emnlp-main.476

Liu, Chen, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024a. Are multilingual LLMs culturally-diverse reasoners? An investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039. https://doi.org/10.18653/v1/2024.naacl-long.112

Liu, Chen Cecilia, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *ArXiv preprint*, abs/2406.03930.

Liu, Dayiheng, Kexin Yang, Qian Qu, and Jiancheng Lv. 2019. Ancient–modern Chinese translation with a new large training dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(1). https://doi.org/10.1145/3325887

Liu, Emmy, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022b. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452. https://doi.org/10.18653/v1/2022.naacl-main.330

Liu, Fangyu, 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. https://doi.org/10.18653/v1/2021.emnlp-main.818

Liu, Hanmeng, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023b. LogiQA 2.0—An improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962. https://doi.org/10.1109/TASLP.2023.3293046

Liu, Meina. 2016. Verbal communication styles and culture. Oxford Research Encyclopedia of Communication. https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-162. https://doi.org/10.1093/acrefore/9780190228613.013.162

Liu, Xiao, Yansong Feng, Jizhi Tang, Chengang Hu, and Dongyan Zhao. 2022c. Counterfactual recipe generation: Exploring compositional generalization in a realistic scenario. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7354–7370. https://doi.org/10.18653/v1/2022.emnlp-main.497

Liu, Xuelin, Yanfei Zhu, Shucheng Zhu, Pengyuan Liu, Ying Liu, and Dong Yu. 2024b. Evaluating moral beliefs across LLMs through a pluralistic framework. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4740–4760. https://doi.org/10.18653/v1/2024.findings-emnlp.272

Liu, Yang, Zi Lin, and Sichen Kang. 2018. Towards a description of Chinese morphemic concepts and semantic word-formation. *Journal of Chinese Information Processing*, 32(2):11–20.

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

Liu, Zhi Song, Robin Courant, and Vicky Kalogeiton. 2024. FunnyNet-W: Multimodal learning of funny moments in videos in the wild. *International Journal of Computer Vision*, pages 1–22. https://doi.org/10.1007/s11263-024-02000-2

Liu, Zhixuan, Youeun Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. 2023c. Towards equitable representation in text-to-image synthesis models with the cross-cultural understanding benchmark (CCUB) dataset. *ArXiv preprint*, abs/2301.12073.

Liu, Zihan, Han Li, Anfan Chen, Renwen Zhang, and Yi-Chieh Lee. 2024d. Understanding public perceptions of AI conversational agents: A cross-cultural analysis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3642840

Liu, Zoey, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022d. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944. https://doi.org/10.18653/v1/2022.acl-long.272

Lou, Lianzhang, Xi Yin, Yutao Xie, and Yang Xiang. 2023. CCEval: A representative evaluation benchmark for the Chinese-centric multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10176–10184. https://doi.org/10.18653/v1/2023.findings-emnlp.682

Lovenia, Holy, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James Validad Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, et al. 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. https://doi.org/10.18653/v1/2024.emnlp-main.296

Lu, Christina, Jackie Kay, and Kevin McKee. 2022. Subverting machines, fluctuating identities: Re-learning human categorization. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1005–1015. https://doi.org/10.1145/3531146.3533161

Lucchini, Lorenzo, Sara Tonelli, and Bruno Lepri. 2019. Following the footsteps of giants: Modeling the mobility of historically notable individuals using Wikipedia. *EPJ Data Science*, 8(1):36. https://doi.org/10.1140/epjds/s13688-019-0215-7

Lyu, Hanjia, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. 2025. GPT-4V(ision) as a social media analysis engine. *ACM Transactions on Intelligent Systems and Technology*, 16(3). https://doi.org/10.1145/3709005

Ma, Weicheng, Samiha Datta, Lili Wang, and Soroush Vosoughi. 2022. EnCBP: A new benchmark dataset for finer-grained cultural background prediction in English. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2811–2823. https://doi.org/10.18653/v1/2022.findings-acl.221

Ma, Zheng, Mianzhi Pan, Wenhan Wu, Ka Leong Cheng, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2023. Food-500 Cap: A fine-grained food caption benchmark for evaluating vision-language models. *Proceedings of the 31st ACM International Conference on Multimedia*. https://doi.org/10.1145/3581783.3611994

Magomere, Jabez, Shu Ishida, Tejumade Afonja, Aya Salama, Daniel Kochin, Foutse Yuehgoh, Imane Hamzaoui, Raesetje Sefala, Aisha Alaagib, Elizaveta Semenova, et al. 2024. You are what you eat? Feeding foundation models a regionally diverse food dataset of World Wide Dishes. *ArXiv preprint*, abs/2406.09496.

Majewska, Olga, Evgeniia Razumovskaia, Edoardo M. Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Transactions of the Association for Computational Linguistics*, 11:139–156. https://doi.org/10.1162/tacl_a_00539

Makridis, Georgios, Athanasios Oikonomou, and Vasileios Koukos. 2024. FairyLandAI:

Personalized fairy tales utilizing ChatGPT and DALLE-3. *ArXiv preprint*, abs/2407.09467.

Manvi, Rohin, Samar Khanna, Marshall Burke, David B. Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 34654–34669.

Manzoor, Muhammad Arslan, Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. Can machines resonate with humans? Evaluating the emotional and empathic comprehension of LMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14683–14701. https://doi.org/10.18653/v1/2024.findings-emnlp.861

Marino, Kenneth, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 3195–3204. https://doi.org/10.1109/CVPR.2019.00331

Maronikolakis, Antonis, 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. https://doi.org/10.18653/v1/2022.findings-acl.87

Masala, Mihai, Denis C. Ilie-Ablachim, Alexandru Dima, Dragos Corlatescu, Miruna Zavelca, Ovio Olaru, Simina Terian-Dan, Andrei Terian-Dan, Marius Leordeanu, Horia Velicu, et al. 2024. "Vorbeşti Româneşte?" A recipe to train powerful Romanian LLMs with English instructions. *ArXiv preprint*, abs/2406.18266. https://doi.org/10.18653/v1/2024.findings-emnlp.681

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

Masud, Sarah, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of LLMs in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863. https://doi.org/10.18653/v1/2024.emnlp-main.886

Mathur, Leena, Ralph Adolphs, and Maja J. Matarić. 2023. Towards intercultural affect recognition: Audio-visual affect recognition in the wild across six cultures. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. https://doi.org/10.1109/FG57933.2023.10042676

Matsumoto, David. 2007. Culture, context, and behavior. *Journal of personality*, 75(6):1285–1320. https://doi.org/10.1111/j.1467-6494.2007.00476.x, PubMed: 17995466

Mausam. 2016. Open information extraction systems and downstream applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016*, pages 4074–4077.

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

Menadue, Christopher Benjamin and Karen Diane Cheer. 2017. Human culture and science fiction: A review of the literature, 1980–2016. *Sage Open*, 7(3):2158244017723690. https://doi.org/10.1177/2158244017723690

Menicatti, Roberto, Barbara Bruno, and Antonio Sgorbissa. 2017. Modelling the influence of cultural information on vision-based human home activity recognition. In *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 32–38. https://doi.org/10.1109/URAI.2017.7992880

Mercorio, Fabio, Mario Mezzanzanica, Daniele Potertì, Antonio Serino, and Andrea Seveso. 2024. Disce aut deficere: Evaluating LLMs proficiency on the INVALSI Italian benchmark. *ArXiv preprint*, abs/2406.17535.

Migon Favaretto, Rodolfo, Soraia Raupp Musse, Angelo Brandelli Costa, Rodolfo Migon Favaretto, Soraia Raupp Musse, and Angelo Brandelli Costa. 2019. Detecting Hofstede cultural dimensions. *Emotion, Personality and Cultural Aspects in Crowds: Towards a Geometrical Mind*, pages 93–103. https://doi.org/10.1007/978-3-030-22078-5_8

Mim, Nusrat Jahan, Dipannita Nandi, Sadaf Sumyia Khan, Arundhuti Dey, and Syed Ishtiaque Ahmed. 2024. In-between visuals and visible: The impacts of text-to-image generative AI tools on digital image-making practices in the Global South. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3641951

Min, Sewon, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

Min, Weiqing, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2020. ISIA Food-500: A dataset for large-scale food recognition via stacked global-local attention network. In *MM '20: The 28th ACM International Conference on Multimedia*, pages 393–401. https://doi.org/10.1145/3394171.3414031

Mirowski, Piotr, Juliette Love, Kory Mathewson, and Shakir Mohamed. 2024. A robot walks into a bar: Can language models serve as creativity support tools for comedy? An evaluation of LLMs' humour alignment with comedians. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1622–1636. https://doi.org/10.1145/3630106.3658993

Mirza, Shujaat, Bruno Coelho, Yuyuan Cui, Christina Pöpper, and Damon McCoy. 2024. Global-Liar: Factuality of LLMs over time and geographic regions. *ArXiv preprint*, abs/2401.17839.

Mittal, Surbhi, Arnav Sudan, Mayank Vatsa, Richa Singh, Tamar Glaser, and Tal Hassner. 2024. Navigating text-to-image generative bias across Indic languages. *ArXiv preprint*, abs/2408.00283. https://doi.org/10.1007/978-3-031-73223-2_4

Moghimifar, Farhad, Shilin Qu, Tongtong Wu, Yuan-Fang Li, and Gholamreza Haffari. 2023. NormMark: A weakly supervised Markov model for socio-cultural norm discovery. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5081–5089. https://doi.org/10.18653/v1/2023.findings-acl.314

Mohamed, Youssef, 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. https://doi.org/10.18653/v1/2022.emnlp-main.600

Mohammad, Saif and Svetlana Kiritchenko. 2018. WikiArt emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Moradi, Armin, Nicola Neophytou, and Golnoosh Farnadi. 2024. Advancing cultural inclusivity: Optimizing embedding spaces for balanced music recommendations. *ArXiv preprint*, abs/2405.17607.

Moseley, Christopher. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Mostafazadeh Davani, Aida, Mark Diaz, Dylan K. 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. https://doi.org/10.18653/v1/2024.emnlp-main.1029

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

Mukherjee, Anjishnu, Aylin Caliskan, Ziwei Zhu, and Antonios Anastasopoulos. 2024a. 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. https://doi.org/10.18653/v1/2024.naacl-long.355

Mukherjee, Anjishnu, 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. https://doi.org/10.18653/v1/2023.emnlp-main.981

Mukherjee, Sagnik, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024b. Cultural conditioning or placebo? On the effectiveness of socio-demographic prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15811–15837. https://doi.org/10.18653/v1/2024.emnlp-main.884

Myung, Junho, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. BLEnD: A Benchmark for LLMs on everyday knowledge in diverse cultures and languages. *ArXiv preprint*, abs/2406.09948.

Nadejde, Maria, 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. https://doi.org/10.18653/v1/2022.findings-naacl.47

Nangia, Nikita, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

Naous, Tarek, Michael J. Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? Measuring cultural bias in large language models. *ArXiv preprint*, abs/2305.14456.

Narayan, Malur, John Pasmore, Elton Sampaio, Vijay Raghavan, and Gabriella Waters. 2024. Bias neutralization framework: Measuring fairness in large language models with bias intelligence quotient (BiQ). *ArXiv preprint*, abs/2404.18276.

Navigli, Roberto, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality*, 15(2). https://doi.org/10.1145/3597307

Nayak, Shravan, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. *ArXiv preprint*, abs/2407.10920. https://doi.org/10.18653/v1/2024.emnlp-main.329

Naylor, Larry. 1996. *Culture and Change: An Introduction*. Bloomsbury Publishing USA.

Neiberg, Daniel, Petri Laukka, and Hillary Anger Elfenbein. 2011. Intra-, inter-, and cross-cultural classification of vocal affect. In *Interspeech 2011*, pages 1581–1584. https://doi.org/10.21437/Interspeech.2011-475

Neplenbroek, Vera, Arianna Bisazza, and Raquel Fernández. 2024. MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. *ArXiv preprint*, abs/2406.07243.

Névéol, Aurélie, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531. https://doi.org/10.18653/v1/2022.acl-long.583

Newmark, Peter. 1988. *A Textbook of Translation*, volume 66. Prentice Hall New York.

Nguyen, Thuat, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237.

Nguyen, Tuan-Phong, Simon Razniewski, Aparna S. Varde, and Gerhard Weikum. 2023a. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023, WWW 2023*, pages 1907–1917. https://doi.org/10.1145/3543507.3583535

Nguyen, Tuan Phong, Simon Razniewski, and Gerhard Weikum. 2024. Multi-cultural commonsense knowledge distillation. *ArXiv preprint*, abs/2402.10689.

Nguyen, Xuan Phi, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen

Yang, Chaoqun Liu, et al. 2023b. SeaLLMs–Large language models for Southeast Asia. *ArXiv preprint*, abs/2312.00738.
Nigst, Lorenz, Maxim Romanov, Sarah Bowen Savant, Masoumeh Seydi, and Peter Verkinderen. 2021. OpenITI: A machine-readable corpus of Islamicate texts (2021.2. 5)[Data Set].
Nisbett, Richard E. and Takahiko Masuda. 2003. Culture and point of view. *Proceedings of the National Academy of Sciences of the United States of America*, 100:11163–11170. https://doi.org/10.1073/pnas.1934527100, PubMed: 12960375
Nivre, Joakim, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*.
Nordhoff, Sebastian. 2012. Linked data for linguistic diversity research: Glottolog/Langdoc and ASJP online. In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*. Springer, pages 191–200. https://doi.org/10.1007/978-3-642-28249-2_18
Noroozi, Fatemeh, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. 2018. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, 12(2):505–523. https://doi.org/10.1109/TAFFC.2018.2874986
Ochieng, Millicent, Varun Gumma, Sunayana Sitaram, Jindong Wang, Vishrav Chaudhary, Keshet Ronen, Kalika Bali, and Jacki O'Neill. 2024. Beyond metrics: Evaluating LLMs' effectiveness in culturally nuanced, low-resource real-world scenarios. *ArXiv preprint*, abs/2406.00343.
Ofer, Dan and Dafna Shahaf. 2022. Cards against AI: Predicting humor in a fill-in-the-blank party game. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5397–5403. https://doi.org/10.18653/v1/2022.findings-emnlp.394
OpenAI. XXXX. ChatGPT. https://chat.openai.com/
OpenAI. 2023. GPT-4 technical report. *ArXiv preprint*, abs/2303.08774.
Owen, Louis, Vishesh Tripathi, Abhay Kumar, and Biddwan Ahmed. 2024. Komodo: A linguistic expedition into Indonesia's regional languages. *ArXiv preprint*, abs/2403.09362.
Ozaki, Shintaro, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards cross-lingual explanation of artwork in large-scale vision language models. *ArXiv preprint*, abs/2409.01584. https://doi.org/10.18653/v1/2025.findings-naacl.209
Padmakumar, Vishakh and He He. 2023. Does writing with language models reduce content diversity? *ArXiv preprint*, abs/2309.05196.
Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71.
Palta, Shramay 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. https://doi.org/10.18653/v1/2023.findings-acl.631
Pandya, Keivalya and Mehfuza Holia. 2023. Automating customer service using LangChain: Building custom open-source GPT chatbot for organizations. *ArXiv preprint*, abs/2310.05421.
Pappas, Nikolaos, Miriam Redi, Mercan Topkara, Brendan Jou, Hongyi Liu, Tao Chen, and Shih-Fu Chang. 2016. Multilingual visual sentiment concept matching. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 151–158. https://doi.org/10.1145/2911996.2912016
Parida, Shantipriya, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, Aneesh Bose, Guneet Singh Kohli, Ibrahim Said Ahmad, Ketan Kotwal, Sayan Deb Sarkar, Ondřej Bojar, and Habeebah Kakudi. 2023. HaVQA: A dataset for visual question answering and multimodal research in Hausa language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10162–10183. https://doi.org/10.18653/v1/2023.findings-acl.646
Park, Joon Sung, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social

simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22. https://doi.org/10.1145/3526113.3545616

Park, Junyeong, Seogyeong Jeong, Seyoung Song, Yohan Lee, and Alice Oh. 2025. 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. https://doi.org/10.18653/v1/2025.c3nlp-1.7

Parrish, Alicia, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105. https://doi.org/10.18653/v1/2022.findings-acl.165

Pawar, Siddhesh, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2025. Presumed cultural identity: How names shape LLM responses. *CoRR*, abs/2502.11995.

Perez, Ethan, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448. https://doi.org/10.18653/v1/2022.emnlp-main.225

Peters, Heinrich and Sandra C. Matz. 2024. Large language models can infer psychological dispositions of social media users. *PNAS Nexus*, 3(6):231. https://doi.org/10.1093/pnasnexus/pgae231, PubMed: 38948324

Peterson, Sharyl Bender and Mary Alyce Lach. 1990. Gender stereotypes in children's books: Their prevalence and influence on cognitive and affective development. *Gender and Education*, 2(2):185–197. https://doi.org/10.1080/0954025900020204

Picca, Davide and John Pavlopoulos. 2024. Deciphering emotional landscapes in the Iliad: A novel French-annotated dataset for emotion recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4462–4467.

Pipatanakul, Kunat, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *ArXiv preprint*, abs/2312.13951.

Pires, Ramon, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. https://doi.org/10.1007/978-3-031-45392-2_15

Pistilli, Giada, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. CIVICS: Building a dataset for examining culturally-informed values in large language models. *ArXiv preprint*, abs/2405.13974. https://doi.org/10.1609/aies.v7i1.31710

Plaza-del Arco, Flor Miriam, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024a. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696. https://doi.org/10.18653/v1/2024.acl-long.415

Plaza-del Arco, Flor Miriam, Amanda Cercas Curry, Susanna Paoli, Alba Cercas Curry, and Dirk Hovy. 2024b. 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. https://doi.org/10.18653/v1/2024.findings-emnlp.251

Prabhakaran, Vinodkumar, 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. https://doi.org/10.18653/v1/2024.naacl-long.190

Prabhakaran, Vinodkumar, Rida Qadri, and Ben Hutchinson. 2022. Cultural

incongruencies in artificial intelligence. *ArXiv preprint*, abs/2211.13069.

Pramanick, Shraman, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796. https://doi.org/10.18653/v1/2021.findings-acl.246

Putri, Rifki Afina, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM generate culturally relevant commonsense QA data? Case study in Indonesian and Sundanese. *ArXiv preprint*, abs/2402.17302. https://doi.org/10.18653/v1/2024.emnlp-main.1145

Qu, Yao and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13. https://doi.org/10.1057/s41599-024-03609-x

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Radharapu, Bhaktipriya, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. 2023. AART: AI-assisted red-teaming with diverse data generation for new LLM-powered applications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 380–395. https://doi.org/10.18653/v1/2023.emnlp-industry.37

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Rai, Sunny, Khushang Zilesh Zaveri, Shreya Havaldar, Soumna Nema, Lyle Ungar, and Sharath Chandra Guntuku. 2024. A cross-cultural analysis of social norms in bollywood and hollywood movies. *ArXiv preprint*, abs/2402.11333.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *ArXiv preprint*, abs/2204.06125.

Ramezani, Aida 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. https://doi.org/10.18653/v1/2023.acl-long.26

Ramponi, Alan. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38. https://doi.org/10.1162/tacl_a_00631

Rao, Abhinav, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. NORMAD: A benchmark for measuring the cultural adaptability of large language models. *ArXiv preprint*, abs/2404.12464.

Razavi, Ali, Aäron van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 14837–14847.

Rehm, Matthias, Elisabeth André, Nikolaus Bee, Birgit Endrass, Michael Wissner, Yukiko Nakano, Afia Akhter Lipi, Toyoaki Nishida, and Hung-Hsuan Huang. 2009. *Creating Standardized Video Recordings of Multimodal Interactions Across Cultures*. Springer-Verlag. https://doi.org/10.1007/978-3-642-04793-0_9

Ringeval, Fabien, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12. https://doi.org/10.1145/3347320.3357688

Rizwan, Hammad, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522. https://doi.org/10.18653/v1/2020.emnlp-main.197

Roemmele, Melissa, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Rohmawati, Inayah Ahyana, Esti Junining, and Pratnyawati Nuridi Suwarso. 2022. The idioms and culture-specific items translation strategy for a classic novel. *Journey: Journal of English Language and Pedagogy*, 5(2):169–181. https://doi.org/10.33503/journey.v5i2.554

Rojas, William Gaviria, Sudnya Frederick Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The Dollar Street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01042

Romero, David, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. *ArXiv preprint*, abs/2406.05967.

Ruiz, Nataniel, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, pages 22500–22510. https://doi.org/10.1109/CVPR52729.2023.02155

Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*. https://doi.org/10.1145/3528233.3530757

Sahoo, Nihar, 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. https://doi.org/10.18653/v1/2024.naacl-long.487

Sandoval, Sandra, Jieyu Zhao, Marine Carpuat, and Hal Daumé III. 2023. A rose by any other name would not smell as sweet: Social bias in names mistranslation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3945. https://doi.org/10.18653/v1/2023.emnlp-main.239

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

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23–29 July 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004.

Sap, Maarten, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473. https://doi.org/10.18653/v1/D19-1454

Sap, Maarten, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

pages 5884–5906. https://doi.org/10.18653/v1/2022.naacl-main.431
Sapinski, Tomasz and Dorota Kaminska. 2015. In Emotion recognition from natural speech–emotional profiles. *Logistyka*, 4.
Saxon, Michael and William Yang Wang. 2023. Multilingual conceptual coverage in text-to-image models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4831–4848. https://doi.org/10.18653/v1/2023.acl-long.266
Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
Schneider, Florian and Sunayana Sitaram. 2024. M5 –A diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. *ArXiv preprint*, abs/2407.03791. https://doi.org/10.18653/v1/2024.findings-emnlp.250
Schramm, W. 1954. *The Process and Effects of Mass Communication. How Communication Works*, volume 586. University of Illinois Press.
Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*.
Schulhoff, Sander, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.
Schwartz, Shalom H. 2012. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1):11. https://doi.org/10.9707/2307-0919.1116
Scott, Michael 'Adrir', Sharath Chandra Guntuku, Huan Yang, Weisi Lin, and George Ghinea. 2015. Modelling human factors in perceptual multimedia quality: On the role of personality and culture. *Proceedings of the 23rd ACM International Conference on Multimedia*. https://doi.org/10.1145/2733373.2806254
Segall, Marshall H., Donald T. Campbell, and Melville J. Herskovits. 1967. *The Influence of Culture on Visual Perception*. Bobs-Merrill.
Sengupta, Neha, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *ArXiv preprint*, abs/2308.16149.
Seth, Agrima, 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. https://doi.org/10.2139/ssrn.4756716
Shafayat, Sheikh, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-fact: Assessing factuality of multilingual LLMs using FActScore. *arXiv preprint arXiv:2402.1804*.
Shah, Priyanshi and Ziad Kobti. 2020. Multimodal fake news detection using a cultural algorithm with situational and normative knowledge. *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–7. https://doi.org/10.1109/CEC48606.2020.9185643
Shaikh, Omar, 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. https://doi.org/10.18653/v1/2023.findings-acl.410
Sharma, Shivam, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Y. Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 5597–5606. https://doi.org/10.24963/ijcai.2022/781
Shen, Siqi, 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. https://doi.org/10.18653/v1/2024.naacl-long.316

Shen, Tao, Xiubo Geng, and Daxin Jiang. 2022. Social norms-grounded machine ethics in complex narrative situation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1333–1343.

Shi, Weiyan, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *ArXiv preprint*, abs/2404.15238. https://doi.org/10.18653/v1/2024.findings-emnlp.288

Shifman, Limor. 2013. *Memes in Digital Culture*. MIT Press. https://doi.org/10.7551/mitpress/9429.001.0001

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759. https://doi.org/10.1038/s41586-024-07566-y, PubMed: 39048682

Shwartz, Vered. 2022. Good night at 4 pm?! Time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853. https://doi.org/10.18653/v1/2022.findings-acl.224

Singh, Akshay and Rahul Thakur. 2024. Generalizable multilingual hate speech detection on low resource Indian languages using fair selection in federated learning. 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 7211–7221. https://doi.org/10.18653/v1/2024.naacl-long.400

Son, Guijin, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024a. KMMLU: Measuring massive multitask language understanding in Korean. *ArXiv preprint*, abs/2402.11548.

Son, Guijin, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 2024b. 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.

Song, Inhwa, Sachin R. Pendse, Neha Kumar, and Munmun De Choudhury. 2024. The typing cure: Experiences with large language model chatbots for mental health support. *ArXiv preprint*, abs/2401.14362.

Speer, Robyn, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA*, pages 4444–4451. https://doi.org/10.1609/aaai.v31i1.11164

Srinivasan, Krishna, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449. https://doi.org/10.1145/3404835.3463257

Stefanini, Matteo, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *Image Analysis and Processing–ICIAP 2019: 20th International Conference, Proceedings, Part II 20*, pages 729–740. https://doi.org/10.1007/978-3-030-30645-8_66

Struppek, Lukas, Dom Hintersdorf, Felix Friedrich, Patrick Schramowski, Kristian Kersting, et al. 2023. Exploiting cultural biases via homoglyphs in text-to-image synthesis. *Journal of Artificial Intelligence Research*, 78:1017–1068. https://doi.org/10.1613/jair.1.15388

Survey, World Values. 2022. World Values Survey. https://www.worldvaluessurvey.org/wvs.jsp

Suwaileh, Reem, Maram Hasanain, Fatema Hubail, Wajdi Zaghouani, and Firoj Alam. 2024. ThatiAR: Subjectivity detection in Arabic news sentences. *ArXiv preprint*, abs/2406.05559. https://doi.org/10.1609/icwsm.v19i1.35960

Talmor, Alon, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense

knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Tam, Zhi Rui, Ya-Ting Pai, Yen-Wei Lee, Jun-Da Chen, Wei-Min Chu, Sega Cheng, and Hong-Han Shuai. 2024. An improved traditional Chinese evaluation suite for foundation model. *arXiv preprint arXiv:2403.01858*.

Tang, Jingqun, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. 2024a. MTVQA: Benchmarking multilingual text-centric visual question answering. *ArXiv preprint*, abs/2405.11985.

Tang, Xuemei, Qi Su, Jun Wang, and Zekun Deng. 2024b. CHisIEC: An information extraction corpus for ancient Chinese history. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3192–3202.

Tao, Yan, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):346. https://doi.org/10.1093/pnasnexus/pgae346, PubMed: 39290441

Thai, Katherine, 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. https://doi.org/10.18653/v1/2022.emnlp-main.672

Thapliyal, Ashish V., Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729. https://doi.org/10.18653/v1/2022.emnlp-main.45

Toker, Michael, Oren Mishali, Ophir Münz-Manor, Benny Kimelfeld, and Yonatan Belinkov. 2024. A dataset for metaphor detection in early medieval Hebrew poetry. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–453. https://doi.org/10.18653/v1/2024.eacl-short.39

Tomaselli, Alexandra and Alexandra Xanthaki. 2021. The struggle of Indigenous peoples to maintain their spirituality in Latin America: Freedom of and from religion (s), and other threats. *Religions*, 12(10):869. https://doi.org/10.3390/rel12100869

Tonja, Atnafu Lambebo, 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, et al. 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.

Tonneau, Manuel, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott 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. https://doi.org/10.18653/v1/2024.woah-1.23

Toral, Antonio and Andy Way. 2015. Machine-assisted translation of literary text: A case study. *Translation Spaces*, 4(2):240–267. https://doi.org/10.1075/ts.4.2.04tor

Törnberg, Petter. 2024. *How to Use Large-Language Models for Text Analysis*. https://doi.org/10.4135/9781529683707

Toro Isaza, Paulina, Guangxuan Xu, Toye Oloko, Yufang Hou, Nanyun Peng, and Dakuo Wang. 2023. Are fairy tales fair? Analyzing gender bias in temporal narrative event chains of children's fairy tales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6509–6531. https://doi.org/10.18653/v1/2023.acl-long.359

Tran, Khanh Tung, Barry O'Sullivan, and Hoang D. Nguyen. 2024. UCCIX: Irish-eXcellence large language model. *ArXiv preprint*, abs/2405.13010. https://doi.org/10.3233/FAIA241040

Tran, Minh, Yufeng Yin, and Mohammad Soleymani. 2023. Personalized adaptation with pre-trained speech encoders for continuous emotion recognition. In *24th*

Annual Conference of the International Speech Communication Association, Interspeech 2023, pages 636–640. https://doi.org/10.21437/Interspeech.2023-2170

Ullah, Faizad, 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.

Üstün, Ahmet, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *ArXiv preprint*, abs/2402.07827. https://doi.org/10.18653/v1/2024.acl-long.845

Ventura, Mor, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2023. Navigating cultural chasms: Exploring and unlocking the cultural POV of text-to-image models. *ArXiv preprint*, abs/2310.01929.

Vligouridou, Eleni, Inessa Iliadou, and Çağrı Çöltekin. 2024. A treebank of Asia Minor Greek. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1715–1721.

Wagner, Claudia, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. 2021. Measuring algorithmically infused societies. *Nature*, 595(7866):197–204. https://doi.org/10.1038/s41586-021-03666-1, PubMed: 34194046

Walker, Mirella, Fang Jiang, Thomas Vetter, and Sabine Sczesny. 2011. Universals and cultural differences in forming personality trait judgments from faces. *Social Psychological and Personality Science*, 2(6):609–617. https://doi.org/10.1177/1948550611402519

Walsh, Melanie, Anna Preus, and Maria Antoniak. 2024. Sonnet or not, bot? Poetry evaluation for large models and datasets. *ArXiv preprint*, abs/2406.18906. https://doi.org/10.18653/v1/2024.findings-emnlp.914

Wang, Angelina, Jamie Morgenstern, and John P. Dickerson. 2024. Large language models should not replace human participants because they can misportray and flatten identity groups. *ArXiv preprint*, abs/2402.01908.

Wang, Bin, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy F. Chen. 2024a. CRAFT: Extracting and tuning cultural instructions from the wild. *ArXiv preprint*, abs/2405.03138. https://doi.org/10.18653/v1/2024.c3nlp-1.4

Wang, Bin, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024b. 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. https://doi.org/10.18653/v1/2024.naacl-long.22

Wang, Han, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024c. MultiHateClip: A multilingual benchmark dataset for hateful video detection on YouTube and Bilibili. *ArXiv preprint*, abs/2408.03468. https://doi.org/10.1145/3664647.3681521

Wang, Wenxuan, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024d. 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. https://doi.org/10.18653/v1/2024.acl-long.345

Wang, Xidong, 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. https://doi.org/10.18653/v1/2024.naacl-long.343

Wang, Yizhong, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Super-natural instructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference*

on Empirical Methods in Natural Language Processing, pages 5085–5109. https://doi.org/10.18653/v1/2022.emnlp-main.340

Wang, Yuhang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2023. CDEval: A benchmark for measuring the cultural dimensions of large language models. *ArXiv preprint*, abs/2311.16421.

Wang, Yuwei, Enmeng Lu, Zizhe Ruan, Yao Liang, and Yi Zeng. 2024f. Stream: Social data and knowledge collective intelligence platform for TRaining Ethical AI Models. *AI & SOCIETY*, pages 1–9. https://doi.org/10.1007/s00146-023-01851-6

Wang, Yuxuan, Yijun Liu, Fei Yu, Chen Huang, Kexin Li, Zhiguo Wan, and Wanxiang Che. 2024g. CVLUE: A new benchmark dataset for Chinese vision-language understanding evaluation. https://doi.org/10.1609/aaai.v39i8.32884

Wang, Zhilin, Yu Ying Chiu, and Yu Cheung Chiu. 2023. Humanoid agents: Platform for simulating human-like generative agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 167–176. https://doi.org/10.18653/v1/2023.emnlp-demo.15

Watts, Ishaan, 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. https://doi.org/10.18653/v1/2024.emnlp-main.451

Waugh, Linda R. 1982. Marked and unmarked: A choice between unequals in semiotic structure. *Semiotica*, 38(3–4):299–318. https://doi.org/10.1515/semi.1982.38.3-4.299

Weber, James and Michael J. Urick. 2017. Examining the millennials' ethical profile: Assessing demographic variations in their personal value orientations. *Business and Society Review*, 122(4):469–506. https://doi.org/10.1111/basr.12128

Weber, Maurice, Carlo Siebenschuh, Rory Butler, Anton Alexandrov, Valdemar Thanner, Georgios Tsolakis, Haris Jabbar, Ian T. Foster, Bo Li, Rick Stevens, and Ce Zhang. 2023. WordScape: A pipeline to extract multilingual, visually rich documents with layout annotations from Web crawl data. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Weck, Benno, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. MuChoMusic: Evaluating music understanding in multimodal audio-language models. *ArXiv preprint*, abs/2408.01337.

Wei, Yuting, 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. *ArXiv preprint*, abs/2403.06574. https://doi.org/10.18653/v1/2024.findings-emnlp.87

Weidinger, Laura, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative AI systems. *ArXiv preprint*, abs/2310.11986.

White, Isadora, 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. https://doi.org/10.18653/v1/2024.findings-emnlp.711

White, Leslie A. 1959. The concept of culture. *American Anthropologist*, 61(2):227–251. https://doi.org/10.1525/aa.1959.61.2.02a00040

Whorf, Benjamin Lee. 2012. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press.

Wibowo, Haryo, 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. https://doi.org/10.18653/v1/2024.naacl-long.77

Winata, Genta Indra, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, et al. 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. https://doi.org/10.18653/v1/2025.naacl-long.167

Wright, Dustin, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. Revealing fine-grained values and opinions in large language models. *ArXiv preprint*, abs/2406.19238. https://doi.org/10.18653/v1/2024.findings-emnlp.995

Wu, Minghao, 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. https://doi.org/10.18653/v1/2024.emnlp-demo.14

Wu, Minghao, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. (Perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *ArXiv preprint*, abs/2405.11804.

Wunarso, Novita Belinda and Yustinus Eko Soelistio. 2017. Towards Indonesian speech-emotion automatic recognition (I-SpEAR). In *2017 4th International Conference on New Media Studies (CONMEDIA)*, pages 98–101. https://doi.org/10.1109/CONMEDIA.2017.8266038

Wuraola, Ifeoluwa, 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. https://doi.org/10.18653/v1/2024.emnlp-main.869

Würtz, Elizabeth. 2017. Intercultural communication on Web sites: A cross-cultural analysis of Web sites from high-context cultures and low-context cultures. *Journal of Computer-Mediated Communication*, 11(1):274–299. https://doi.org/10.1111/j.1083-6101.2006.tb00313.x

Xie, Heng, Jizhou Cui, Yuhang Cao, Junjie Chen, Jianhua Tao, Cunhang Fan, Xuefei Liu, Zhengqi Wen, Heng Lu, Yuguang Yang, Zhao Lv, and Yongwei Li. 2023. Multimodal cross-lingual features and weight fusion for cross-cultural humor detection. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, MuSe '23, pages 51–57. https://doi.org/10.1145/3606039.3613110

Xiong, Haoyi, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: Visions and challenges. *IEEE Transactions on Services Computing*, 17:4558–4577. https://doi.org/10.1109/TSC.2024.3451185

Xu, Tao, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 1316–1324. https://doi.org/10.1109/CVPR.2018.00143

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41

Yadav, Srishti, 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 7592–7608. https://doi.org/10.18653/v1/2025.findings-naacl.422

Yanaka, Hitomi, Namgi Han, Ryoma Kumon, Jie Lu, Masashi Takeshita, Ryo Sekizawa, Taisei Kato, and Hiromi Arai. 2024. Analyzing social biases in Japanese large language models. *ArXiv preprint*, abs/2406.02050.

Yang, Diyi, Caleb Ziems, William Held, Omar Shaikh, Michael S. Bernstein, and John Mitchell. 2024. Social skill training with large language models. *ArXiv preprint*, abs/2404.04204.

Yao, Binwei, 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. https://doi.org/10.18653/v1/2024.findings-emnlp.765

Yao, Jing, 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. https://doi.org/10.18653/v1/2024.naacl-long.486

Yarlott, W. Victor, 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.

Ye, Andre, Sebastin Santy, Jena D. Hwang, Amy X. Zhang, and Ranjay Krishna. 2023. Computer vision datasets and models exhibit cultural and linguistic diversity in perception. *arXiv preprint arXiv:2310.14356v3*.

Ye, Fulong, Guang Liu, Xinya Wu, and Ledell Wu. 2024. AltDiffusion: A multilingual text-to-image diffusion model. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014*, pages 6648–6656. https://doi.org/10.1609/aaai.v38i7.28487

Yin, Da, 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. https://doi.org/10.18653/v1/2022.emnlp-main.132

Yin, Da, Feng Gao, Govind Thattai, Michael Johnston, and Kai-Wei Chang. 2023. GIVL: Improving geographical inclusivity of vision-language models with pre-training methods. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, pages 10951–10961. https://doi.org/10.1109/CVPR52729.2023.01054

Yin, Da, 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. https://doi.org/10.18653/v1/2021.emnlp-main.162

Yin, Yaqi, Yue Wang, and Yang Liu. 2024. Chinese morpheme-informed evaluation of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3165–3178.

Yin, Ziqi, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect LLMs? A cross-lingual study on the influence of prompt politeness on LLM performance. *ArXiv preprint*, abs/2402.14531. https://doi.org/10.18653/v1/2024.sicon-1.2

Yoo, Kang Min, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. 2024. HyperCLOVA X technical report. *ArXiv preprint*, abs/2404.01954.

Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78. https://doi.org/10.1162/tacl_a_00166

Yu, Linhao, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Liutao Liutao, and Deyi Xiong. 2024. CMoralEval: A moral evaluation benchmark for Chinese large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11817–11837. https://doi.org/10.18653/v1/2024.findings-acl.703

Yu, Yu, Weibin Zhang, and Yun Deng. 2021. Frechet Inception Distance (FID) for evaluating GANs. *China University of Mining Technology Beijing Graduate School*, 3.

Yuan, Ye, 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. https://doi.org/10.18653/v1/2024.findings-naacl.43

Yun, Youngsik and Jihie Kim. 2024. CIC: A framework for culturally-aware image captioning. *ArXiv preprint*, abs/2402.05374.

https://doi.org/10.24963/ijcai.2024/180

Yüksel, Arda, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schütze. 2024. TurkishMMLU: Measuring massive multitask language understanding in Turkish. *ArXiv preprint*, abs/2407.12402. https://doi.org/10.18653/v1/2024.findings-emnlp.413

Zellers, Rowan, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104. https://doi.org/10.18653/v1/D18-1009

Zhan, Haolan, 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. https://doi.org/10.18653/v1/2024.findings-naacl.196

Zhang, Chen, 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. https://doi.org/10.18653/v1/2024.acl-long.479

Zhang, Dong, Zhaowei Li, Pengyu Wang, Xin Zhang, Yaqian Zhou, and Xipeng Qiu. 2024b. SpeechAgents: Human-communication simulation with multi-modal multi-agent systems. *ArXiv preprint*, abs/2401.03945.

Zhang, Gengyuan, Yurui Zhang, Kerui Zhang, and Volker Tresp. 2024c. Can vision-language models be a good guesser? Exploring VLMs for times and location reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 636–645. https://doi.org/10.1109/WACV57701.2024.00069

Zhang, Han, Tao Xu, and Hongsheng Li. 2017. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017*, pages 5908–5916. https://doi.org/10.1109/ICCV.2017.629

Zhang, Lili, Xi Liao, Zaijia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. 2024d. Partiality and misconception: Investigating cultural representativeness in text-to-image models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11–16, 2024*, pages 620:1–620:25. https://doi.org/10.1145/3613904.3642877

Zhang, Lili, Xi Liao, Zaijia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. 2024e. Partiality and misconception: Investigating cultural representativeness in text-to-image models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. https://doi.org/10.1145/3613904.3642877

Zhang, Miaosen, Yixuan Wei, Zhen Xing, Yifei Ma, Zuxuan Wu, Ji Li, Zheng Zhang, Qi Dai, Chong Luo, Xin Geng, et al. 2024f. Aligning vision models with human aesthetics in retrieval: Benchmarks and algorithms. *ArXiv preprint*, abs/2406.09397.

Zhang, Wei, Wong Kam-Kwai, Biying Xu, Yiwen Ren, Yuhuai Li, Minfeng Zhu, Yingchaojie Feng, and Wei Chen. 2024g. CultiVerse: Towards cross-cultural understanding for paintings with large language model. *ArXiv preprint*, abs/2405.00435.

Zhang, Wenjing, Siqi Xiao, Xuejiao Lei, Ning Wang, Huazheng Zhang, Meijuan An, Bikun Yang, Zhaoxiang Liu, Kai Wang, and Shiguo Lian. 2024h. Methodology of adapting large English language models for specific cultural contexts. *ArXiv preprint*, abs/2406.18192.

Zhang, Wenxuan, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3Exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Zhang, Yigeng, Fabio Gonzalez, and Thamar Solorio. 2024. Interpreting themes from educational stories. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9190–9203.

Zhang, Zhu, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. 2021. M6-UFC: Unifying multi-modal controls for conditional image synthesis via non-autoregressive generative transformers. *ArXiv preprint*, abs/2105.14211.

Zhao, Jiaxu, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, and Mykola Pechenizkiy. 2023. CHBias: Bias evaluation and mitigation of Chinese conversational language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13538–13556. https://doi.org/10.18653/v1/2023.acl-long.757

Zhao, Jinming, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710. https://doi.org/10.18653/v1/2022.acl-long.391

Zhao, Wenlong, 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.

Zhao, Yijun, Jiangyu Pan, Yan Dong, Tianshu Dong, Guanyun Wang, Fangtian Ying, Qihang Shen, and Jiacheng Cao. 2024b. Language urban odyssey: A serious game for enhancing second language acquisition through large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. https://doi.org/10.1145/3613905.3651112

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Zhi-Xuan, Tan, Micah Carroll, Matija Franklin, and Hal Ashton. 2024. Beyond preferences in AI alignment. *ArXiv preprint*, abs/2408.16984. https://doi.org/10.1007/s11098-024-02249-w

Zhou, Bo, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023a. WYWEB: A NLP evaluation benchmark for classical Chinese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3294–3319. https://doi.org/10.18653/v1/2023.findings-acl.204

Zhou, Li, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023b. 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. https://doi.org/10.1007/978-3-031-34732-0_36

Zhou, Li, Taelin Karidi, Nicolas Garneau, Yong Cao, Wanlong Liu, Wenyu Chen, and Daniel Hershcovich. 2024. Does mapo tofu contain coffee? Probing LLMs for food-related cultural knowledge. *ArXiv preprint*, abs/2404.06833.

Zhou, Wangchunshu, Yan Zeng, Shizhe Diao, and Xinsong Zhang. 2022. VLUE: A multi-task multi-dimension benchmark for evaluating vision-language pre-training. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 27395–27411.

Zhu, Shucheng, Weikang Wang, and Ying Liu. 2024. Quite good, but not enough: Nationality bias in large language models - a case study of ChatGPT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13489–13502.

Ziems, Caleb, 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. https://doi.org/10.18653/v1/2023.acl-long.429

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

https://doi.org/10.1162/coli_a_00502
Ziems, Caleb, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773. https://doi.org/10.18653/v1/2022.acl-long.261

1. Paper Selection Methodology

Our paper selection process followed a systematic multiround approach aligned with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Page et al. 2021), while incorporating additional steps to collect articles. We primarily focused on three major academic repositories: ACL, arXiv, and Semantic Scholar, which collectively encompass publications from all relevant conferences.

1.1 PRISMA Methodology

Our methodology adheres to PRISMA guidelines through:

- **Information Sources**: Comprehensive coverage using multiple academic repositories (ACL, arXiv, Semantic Scholar)
- **Search Strategy**: Well-defined, multi-round search process with explicit inclusion/exclusion criteria
- **Selection Process**: Systematic filtering through multiple rounds with clear criteria
- **Data Validation**: Two-fold validation process combining automated and manual review

1.2 Initial Broad Search (Round 1)

The first round implemented a two-step filtering process:

1. Initial search using primary cultural keywords: {“cultural,” “culturally,” “culture”}
2. Secondary filtering within the initial results using domain-specific keywords related to artificial intelligence: {“vision,” “language models,” “cultural survey,” “LLM,” “VLM,” “benchmark,” “dataset,” “visual,” “NLP,” “linguistics,” “machine learning,” “image generation,” “multimodal”}

To address potential ambiguities and improve precision, we implemented several refinements:

- Replaced generic term “AI” with “artificial intelligence” to avoid false positives
- Required “vision” to co-occur with “model” or “computer vision”

- Required "image" to co-occur with "model," "computer vision," or "dataset"
- Required "speech" and "audio" to co-occur with "model" or "dataset"
- Excluded papers containing terms related to biological cultures ("agriculture," "cell," "bio-culture")

1.3 Language-specific Search (Round 2)

To capture language-specific cultural papers (e.g., "KoBBQ: Korean Bias Benchmark for Question Answering"), we conducted a three-step search:

1. Initial search using keywords comprising all world languages
2. Filtering using cultural keywords used in step 1 of round 1
3. Final filtering using AI-related keywords from step 2 of round 1

1.4 Culture-specific Search (Round 3)

This round targeted papers focusing on multi-lingual cultures, using:

1. Primary search using country and continent names
2. Secondary filtering with cultural keywords from step 1 of round 1
3. Final filtering with AI-related keywords and additional terms including "captioning" and "alignment"

1.5 Citation Network Analysis (Round 4)

To ensure comprehensive coverage, we analyzed the reference sections of all selected papers to identify relevant works that might have been missed in previous rounds. This step aligns with PRISMA's recommendation for additional search strategies beyond database searches.

1.6 Validation and Manual Filtering

We implemented a two-fold validation approach that extends PRISMA's screening process:

1. Manual review of the abstracts, introduction, and titles to filter-out the papers not relevant to the scope of the survey. The authors filtered out the papers outside the scope of the survey by considering factors such as: mention of cultural inclusion (in broader impact statement) without any

discussion of how to do so, focus on multilingual method without consideration of culture, etc. The authors also had extensive discussions about exclusion/inclusion of papers during the filtering process.

2. Additional validation using automated review by large language models (LLMs) to assess relevance and the prompts for paper filtering were borrowed from Schulhoff et al. (2024). The additional positive (to-include) examples that the LLMs found were reviewed manually and some of them were included after discussions.

The paper selection process can be visualized using the PRISMA flow diagram shown in Figure 16, which tracks the number of papers at each stage of the selection process. This systematic approach ensured comprehensive coverage while maintaining high precision in identifying papers relevant to cultural awareness in AI models and natural language processing. Our methodology not only satisfies PRISMA guidelines but also extends them to address the specific challenges of reviewing cultural aspects in AI research.

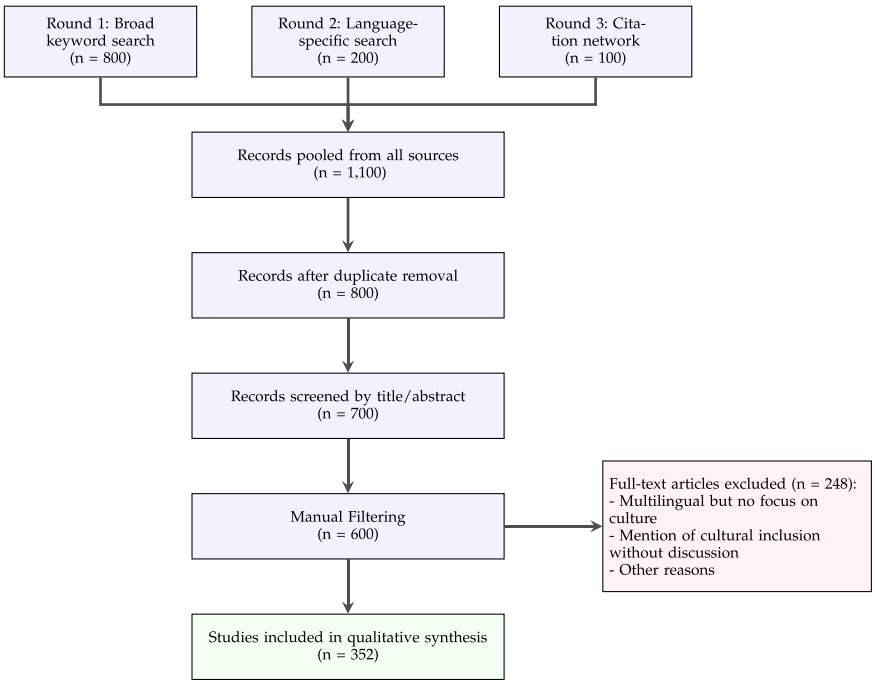


Figure 16
PRISMA flow diagram of the study selection process.