Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art

Chen Cecilia Liu^{1,2,3} and Iryna Gurevych¹ and Anna Korhonen³

¹ Ubiquitous Knowledge Processing Lab,
Department of Computer Science and Hessian Center for AI (hessian.AI),
Technical University of Darmstadt

² Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA),

³ Language Technology Lab, University of Cambridge

Abstract

The surge of interest in culturally aware and adapted Natural Language Processing (NLP) has inspired much recent research. However, the lack of common understanding of the concept of "culture" has made it difficult to evaluate progress in this emerging area. Drawing on prior research in NLP and related fields, we propose an extensive taxonomy of elements of culture that can provide a systematic framework for analyzing and understanding research progress. Using the taxonomy, we survey existing resources and models for culturally aware and adapted NLP, providing an overview of the state of the art and the research gaps that still need to be filled.

1 Introduction

Culture is rapidly becoming an important research topic in Natural Language Processing (NLP), with a significant recent surge in the number of published papers (Figure 1). Given the keen interest in this area and its importance for the safety and fairness of Large Language Models (LLMs), it is now important to consolidate existing research on culturally aware and adapted NLP to take stock of the progress made so far and to identify research gaps. However, this is challenged by the fact that there is no common understanding of the concept of "culture" in NLP.

Prior work such as Hershcovich et al. (2022); Hovy and Yang (2021) laid the vital foundations for understanding ways in which language, culture and society interact in NLP. Hershcovich et al. (2022) proposed a simple taxonomy derived from the interaction between language and culture that captures some elements of culture (linguistic form and style, objectives and values, common ground, and aboutness). Recent papers have adopted a diversity of other definitions (using e.g., "proxies of culture", Adilazuarda et al. 2024), making it challenging to assess progress in this area.

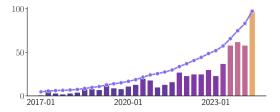


Figure 1: Papers published on arXiv (cs.CL) that contain the keyword "culture", aggregated every 3 months.

Although culture is arguably a complex concept, we can establish a common understanding by drawing on the definitions developed within the humanities and social sciences. In these fields (Tylor, 1871; Kroeber and Kluckhohn, 1952; White, 1959; UNESCO, 1982; Matsumoto and Juang, 1996; Blake, 2000), most definitions of culture encompass *people*, groups of people and the interactions among both individuals and groups.¹

In anthropology, White (1959) provides valuable insights into the *locus of culture*, identifying three key dimensions: 1) "within human" (such as concepts, traditions, beliefs, social practices, etc.), 2) between "social interaction among human beings," and 3) outside of human but "within the patterns of social interaction" (in materialized objects such as tools, arts). An examination of such work reveals that not only people, but also social interaction are critical components of culture.

In NLP, understanding of culture could benefit from the granularity needed to distinguish between more nuanced differences among cultural elements. Additionally, although social interactions are important, they have not been explicitly defined or categorized in prior work on cross-cultural NLP (Hershcovich et al., 2022).

We introduce a new extensive taxonomy of

¹Differences between cultures can exist within groups, such as due to languages; however, variations in language do not imply cultural differences in certain aspects and vice-versa (e.g., United Kingdom versus United States).

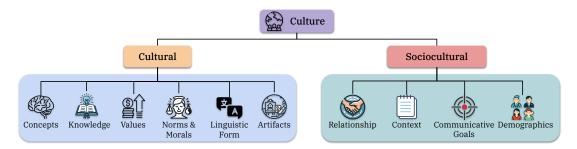


Figure 2: An overview of the taxonomy.

the elements of culture that addresses these problems. Expanding on the basic categories of prior work (Hershcovich et al., 2022; Hovy and Yang, 2021) and grounding the definition of culture in the relevant anthropology literature (Tylor, 1871; Kroeber and Kluckhohn, 1952; White, 1959), we introduce a fine-grained taxonomy that also covers social interactions. We then use this taxonomy to organize and analyze existing works in culturally aware and adapted NLP according to the elements of culture. Our survey of over a hundred publications in NLP provides an up-to-date view of cultural adaptation resources and models and identifies areas of progress as well as research gaps. We hope our taxonomy and analysis can enable and inspire future research in this emerging area.

2 The Taxonomy

In this section, we present our new taxonomy of culture. It differs from earlier NLP works (Hershcovich et al., 2022; Hovy and Yang, 2021) that aimed to define elements of culture in that (i) it is grounded in well-established elements of culture in anthropology literature (Tylor, 1871; Kroeber and Kluckhohn, 1952; White, 1959), (ii) consists of more fine-grained elements than in earlier work, and (iii) allows for a wider consideration of how social factors and variations in humans influence culture.

The taxonomy – shown in Figure 2 has two main branches: **cultural elements** (§3.1) within humans and outside of humans but within the social context (White, 1959) and **sociocultural elements** (§3.2) which affect and are affected by cultural elements through social interactions and communication among humans.²

Cultural Elements. Cultural elements are based

on well-established definitions of culture (Tylor, 1871; Kroeber and Kluckhohn, 1952; White, 1959):

<u>Concepts:</u> basic units of meaning underlying objects, ideas, or beliefs (Jackendoff, 2012).

<u>Knowledge:</u> information that can be acquired through education or practical experience.

<u>Values:</u> beliefs, desirable end states or behaviours ranked by relative importance that can guide evaluations of things (Schwartz, 1992).

Norms and Morals: set of rules or principles that govern people's behaviour and everyday reasoning (Cialdini et al., 1991; Hechter and Opp, 2001; Bicchieri et al., 2018; Gert and Gert, 2002).

Linguistic Form: "how" to construct an utterance (Hershcovich et al., 2022, e.g., dialects).

<u>Artifacts:</u> "materialized" items as the productions of human culture, they can be forms of art, tools, machines (White, 1959) etc.

Sociocultural Elements. "Sociocultural" elements refer to social factors within the scope of NLP that influence and are influenced by previously discussed elements. Leveraging the work of Hovy and Yang (2021), we identify relevant elements:

<u>Relationship:</u> connection between two or more individuals or groups (e.g., father-son, colleagues).

<u>Context:</u> the "containers" of communications (Yang, 2019), which can be linguistic such as surrounding sentences or extra-linguistic (Hovy and Yang, 2021) including social settings (e.g., at a wedding), non-verbal cues (e.g., gesture), or historical contexts (e.g., colonization).

<u>Communicative Goals:</u> the intention behind language use (e.g., requests, apologies).

<u>Demographics</u>: the characteristics of people, such as economic income, education level, nationality, location, political view, family status, etc.

²Note that all elements of culture may interact and/or affect each other depending on context. Our taxonomy abstracts away from this contextual variation.

3 Elements of Culture in Current NLP (Resources) Literature

In this section, we survey and categorize NLP resources published in leading *CL venues since 2022 (details in Appendix A) according to our new taxonomy (§2). For an overview of papers published in each cultural element class of the taxonomy, see Table 1 in the Appendix.

We observe that in resources, culture can be captured in 1) the data itself, or 2) in the labels (e.g., multi-culturally annotated).

3.1 Cultural Elements

3.1.1 Concepts

We can divide concepts into 1) basic concepts that are "configured" differently, reflecting the cultural-specific way of thinking,³ and 2) concepts that are unique to a culture (Wierzbicka, 1992).⁴

Recent NLP works have studied the grounding of time expressions to specific hours across different cultures, as detailed in Shwartz (2022, Time-Expressions, 27 languages). Cao et al. (2024b, CulturalRecipes, 2 languages) study culinary and food-related concepts within the context of recipe adaptations. Majewska et al. (2023); Hu et al. (2023) (4 languages each) culturally adapt names and locations for multilingual Task-Oriented Dialogue (TOD) datasets to enable cultural adaptation of TOD systems. Nevertheless, both training and evaluation datasets are still lacking in this area for diverse cultures, languages, and concept categories (e.g., rituals, aesthetics, kinship, spatial relations).

NLP research has also investigated the usage of concepts spanning multiple categories across cultures, e.g., through metaphors and other figurative expressions (Kabra et al., 2023) or traditional proverbs and sayings (Liu et al., 2023b).

In vision and language (VL) settings, culturally significant or unique concepts have been incorporated into reasoning and captioning tasks such as in MaRVL (Liu et al., 2021), GD-VCR (Yin et al., 2021), and XM3600 (Thapliyal et al., 2022). Additionally, these concepts have been employed to evaluate multimodal content adaptations (Khanuja et al., 2024) or to generate multicultural images (Liu et al., 2023e; Ventura et al., 2023) with

text-to-image models. These datasets are limited in size (e.g., due to the high cost of annotations) and most are available for evaluation only.

3.1.2 Knowledge

Cultural knowledge can be factual or common sense. What weather phenomena can be expected if a rapidly rotating tropical storm forms off the coast of our country? (It's likely called a hurricane in the US and a typhoon in Korea.) Is tofu pudding sweet or savoury by default? (In China, tofu pudding is typically sweet in the south and savoury in the north.)

We identified three major types of resources in NLP literature: 1) probing (by masking entities), 2) multiple choice question answering (MCQA), and 3) knowledge bases.

mLAMA (Kassner et al., 2021), GeoM-LAMA (Yin et al., 2022), DLAMA (Keleg and Magdy, 2023), and FMLAMA (Zhou et al., 2024) probe the diversity of knowledge in LMs.

Recently, MMLU (Hendrycks et al., 2021) style QA benchmarks have greatly aided LLM development, inspiring many cultural variants of the MMLU suite. These suites include ArabicMMLU (Koto et al., 2024a), CMMLU (Li et al., 2023b), IndoMMLU (Koto et al., 2023), SeaEval (Wang et al., 2023a), KMMLU (Son et al., 2024), CLIcK (Kim et al., 2024), just to name a few. The cultural variants of MMLU typically cover some aspects of food, history, literature, geographical knowledge of countries, etc. in respective languages. However, MMLU-style benchmarks are typically constructed using exams and textbooks and cannot combine with other elements in culture. In contrast, common sense datasets (COPAL-ID, Wibowo et al. 2023 or IndoCulture, Koto et al. 2024b) can consider dialects and/or geographical

Lastly, CANDLE (Nguyen et al., 2023), Culture-Atlas (Fung et al., 2024), MANGO (Nguyen et al., 2024) are recent cultural knowledge bases created from either Wikipedia articles or using model distillation. They can be integrated with models to enhance models' cultural awareness. However, diverse sources (i.e., other than Wikipedia, such as from local encyclopedia) should be considered to improve the coverage of knowledge bases.

³For example, one can explore the citizen science project for lexicon associations: https://smallworldofwords.org/en/project/home.

⁴For example, "Kopi Ga Dai" in Singaporean English versus "double-double" in Canada, both referring to coffee with extra sweetness and creaminess, but very different.

⁵Common sense and norms are sometimes used interchangeably in NLP. Norms are acceptable behavioural patterns of a group (§2), which we will discuss in §3.1.4.

3.1.3 Values

Diverse ranking of values among groups can result in differences in aboutness, communication styles, perceptions and multiple other dimensions (Hofstede, 1984, 2011). Such differences in pre-training data can be reflected in LLMs.

Many recent studies on evaluation (Johnson et al., 2022; Ramezani and Xu, 2023; Cao et al., 2023; Wang et al., 2023b; Durmus et al., 2023; Santurkar et al., 2023; Masoud et al., 2023; Havaldar et al., 2023, inter alia) show that LLMs align better with values of WEIRD (Western, Educated, Industrialized, Rich and Democratic, Henrich et al. 2010) people, raising concerns about the fairness and safety of LLM for others. Here, Pew Global Attitudes Survey (PEW)⁶, the World Values Survey (WVS)⁷ and the Hofstede Cultural Dimensions (Hofstede, 1984, 2011) are commonly used for evaluation, along with regional variants like the European Values Survey (EVS, EVS 2011). However, the questions of how to improve the model's value alignment with diverse cultures, what resources to collect and whom to collect from remain unsolved (Kirk et al., 2024).

Biases. In contrast to cultural values, biases have been long-studied in NLP, such as gender bias in machine translation (Stanovsky et al., 2019; Savoldi et al., 2021; Campolungo et al., 2022; Sandoval et al., 2023; Attanasio et al., 2023, inter alia) or bias towards particular social groups. Cultures can have different biases toward the same target groups or have unique biases that are more pronounced in certain cultures (e.g., castes, unnatural beauty standards).

To enable evaluations of cross-cultural variations in biases and develop transferable de-biasing methods, recent work has created culturally aware targets and attribute word sets. WEATHub (Mukherjee et al., 2023) and CA-WEAT (España-Bonet and Barrón-Cedeño, 2022) are multilingual culturally aware extensions based on the standard WEAT (Caliskan et al., 2017) categories. Alternatively, there are other cultural variants such as BIBED (identity-bias in Bengali, Das et al. 2023), CAMeL (Arabic and Western cultural entities, Naous et al. 2023), SeeGULL (regional and national identity-bias, Bhutani et al. 2024), SPICE (stereotypes in India in English, Dev et al. 2023), CHBias (gender/orientation/age/appearance biases,

Chinese, Zhao et al. 2023) among others.

In addition to using word lists, bias evaluation datasets may also consist of sentence pairs (such as French CrowS-Pair, Névéol et al. 2022; WinoQueer, Felkner et al. 2023), biased sentence and context (KosBi, Lee et al. 2023a), conversational (CDIAL-BIAS, Zhou et al. 2022) or question answering (FORK: food-related customs, Palta and Rudinger 2023; SODAPOP, An et al. 2023; KoBBQ, Jin et al. 2024).

In general, there is considerable progress in cultural biases compared to other sub-areas. Many recent surveys on general biases are also available covering critical areas such as evaluations or debiasing methods (Meade et al., 2022; Dev et al., 2022; Sun et al., 2019; Delobelle et al., 2022), we refer readers to them for further details.

Hate. The perception of hatefulness in text varies across cultures, languages and dialects, as recently shown for hate speech classification by Zhou et al. (2023a,b); Lee et al. (2023c); Lwowski et al. (2022); Arango Monnar et al. (2022), among others. Such model disparities may be due to the labelling process, where annotations from diverse cultural groups could serve as a mitigation strategy. CREHate (Lee et al., 2023b) provides hate speech annotations from various English-speaking countries, all applied to the same English dataset. This example is unique in recent research as it explores variations in perceptions within the same language, highlighting the need for further research.

Other Perceptions. The perception of politeness, aesthetic appeal or emotions can also vary across cultures (House and Kasper, 1981; Ringel et al., 2019; Masuda et al., 2008; Mesquita et al., 1997). For example, whether a piece of text is deemed humorous or ironic is culturally dependent. Frenda et al. (2023) tries to address this with a crosscultural annotated irony corpora (EPIC) by people from five English-speaking countries. Similarly, visual elements in arts can elicit different emotions in different cultural groups. ArtELingo (Mohamed et al., 2022) introduces a benchmark with Chinese, Arabic and Spanish captions and emotion labels for artworks, aimed at evaluating models' culturaltransfer performance. This research area is significantly limited.

3.1.4 Norms and Morals

In ethics, a distinction is made between descriptive and normative morality (Gert and Gert, 2002). In NLP, this distinction is often overlooked (Vida

⁶https://www.pewresearch.org/

⁷https://www.worldvaluessurvey.org/

et al., 2023) with a greater emphasis on the "end product", which is the final set of rules or principles and their judgements.

Several norm banks are available, constructed via automatic, semi-automatic or manual norm discovery from conversations, online resources such as Wikipedia, Reddit, government-aided websites or using crowd annotators (Forbes et al., 2020; Fung et al., 2023; Moghimifar et al., 2023; Shi et al., 2024; CH-Wang et al., 2023; Ziems et al., 2023a, 2022b; Rai et al., 2024; Dwivedi et al., 2023). Norm banks have also been automatically adapted to defensible norms in fine-grained situations (Pyatkin et al., 2023; Rao et al., 2023) or culturally aware inference tasks (CH-Wang et al., 2023; Huang and Yang, 2023) for LLM evaluation and adaptation.

Norms and morals can support downstream applications via e.g., enabling the alignment of models. MoralDial (Sun et al., 2023) enables inquisition for moral alignment of LLMs through conversations. PROSOCIALDIALOG (Kim et al., 2022) encourages more socially-aligned responses to problematic content. SocialDial (Zhan et al., 2023), NormDial (Li et al., 2023c) and RENOVI (Zhan et al., 2024) aim to develop conversational experiences that better align with the expectations of a culture.

Nearly all resources are in English, sometimes Chinese, with a focus on Western, and occasionally Chinese or Indian, cultures.

3.1.5 Linguistic Form

Variations in linguistic form can be expressed 1) in the language itself, the habitual use of language, such as dialects; or 2) in style and is tightly coupled with other elements of culture (e.g., social norms) as well as sociocultural factors (e.g., pragmatics, Thomas 1983; Blum-Kulka and Olshtain 1984; Blum-Kulka 1987; Wierzbicka 2003; House and Kasper 1981).

Dialects. A dialect is a variant of a language (Haugen, 1966) at the local regional level (e.g., Hessian German), national level (e.g., Tunisian Arabic) or by other factors (e.g., African American Vernacular English vs. Standard American English).

NLP has predominantly focused on standard language, but there has been recent interest in dialects. Many works have focused on dialect identification (Salameh et al., 2018; Abdelali et al., 2021; Yusuf et al., 2022; Hämäläinen et al., 2021), but how to enable LLMs to serve dialectal communities

remains an open question.

Recently, Le and Luu (2023); Paonessa et al. (2023) identified the disparity of translation results within varieties of Vietnamese and German. Other research shows that models overestimate their performance on GLUE tasks (Ziems et al., 2022a) and masked span predictions when tested with African American (Vernacular) English (AAE/AAVE) (Deas et al., 2023). These dialect disparities correlate with linguistic, economic and social factors (Kantharuban et al., 2023).

The available dialect datasets are typically translations between dialects and standard languages or produced via dialect normalization (in text only or in audio and text). Recent works include DialectEval (Aepli et al., 2023), SDS-200 (Plüss et al., 2022), STT4SG-350 (Plüss et al., 2023) and the ones used in Kuparinen et al. (2023). Only some dialect studies exist for traditional generation tasks such as summarization or standard benchmark tasks (e.g., GLUE). DivSumm (Olabisi et al., 2022) is a dialect-diverse tweets dataset with humanwritten extractive and abstractive summaries. In Held et al. (2023), a synthetic AAE GLUE corpus is created using a rule-based framework (Ziems et al., 2023b). Recently, DIALECTBENCH (Faisal et al., 2024) provides a benchmark for over 281 dialect variants by aggregating over existing datasets.

Overall, research on German and English dialects is more advanced (marginally) than other dialect types.

3.1.6 Artifacts

NLP research on artifacts has focused on (monolingual or mono-cultural) artifacts in texts, e.g., fairy tales, fiction, poetry and songs (Xu et al., 2022; Jiang et al., 2023; Thai et al., 2022; Yang et al., 2019; Haider et al., 2020; Chakrabarty et al., 2021; Ou et al., 2023; Li et al., 2023a), or in multimodal such as movies, humour and memes (Hong et al., 2023; Sharma et al., 2020; Liu et al., 2022a; Hessel et al., 2023), to name a few. While "artifacts" is an independent cultural element, usage in adaptation typically involves tasks that align with one or more previously mentioned categories. For example, in ArtELingo (in §3.1.3), the input data is on art but cross-cultural measurement studies perceptions, which is a part of values (and Impressions, Kruk et al. 2023 is another example using art). Similarly, translations of literary novels need to account for *concept* differences such as names (Jiang et al., 2023) across cultures. Work on integrating crosscultural differences in modelling and data acquisition with artifacts is limited.

3.2 Sociocultural Elements

Although cultural variations are easily identifiable through sociocultural elements, many current work use languages or countries as the boundary for divisions. These elements are understudied in current culturally aware NLP.

3.2.1 Relationship

In many cultures, concepts and forms of communication differ depending on the relationship between the speakers. For example, Chinese has distinct terms for elder vs. younger siblings. Translations to (and from) a language without this property may result in a loss of nuances in meaning. In Korea and Japan, misused politeness level in conversation can violate cultural norms (Matsumoto, 1988; Ambady et al., 1996), especially in different social relationships. Considering relationships is important for building resources and modelling culturally appropriate methods. Zhan et al. (2023, 2024) serve as recent examples with this consideration.

3.2.2 Context

Hovy et al. (2020); Akinade et al. (2023); Stewart and Mihalcea (2024) show that machine translation systems can fail without appropriate consideration of linguistic context, revealing its importance in resource and model development. Extra-linguistic contexts (e.g., settings) have proved especially important in conversational tasks (Zhan et al., 2023), norm bank (Ziems et al., 2023a) construction, and VL applications.

3.2.3 Communicative Goals

Different cultures can have distinctive communication styles depending on communicative goals. For example, people may use indirect language for refusal (vs. direct refusal with a "no") to avoid confrontation (House, 2005). Cultures may also exhibit variations in responses to the same situation (e.g., how to make requests and when to apologize, Blum-Kulka and Olshtain 1984). Taking this type of variation into account is important for cross-cultural pragmatic-inspired tasks – an area that remains understudied.

3.2.4 Demographics

A household with a monthly income of less than 50 US dollars is likely to have different household items than that with 5000 US dollars (Rojas

et al., 2022). Névéol et al. (2022) also found that the original English CrowS-Pair dataset relied on names as proxies for a sociodemographic group ("Amy for women, Tyrone for African American men", Névéol et al. 2022), whereas the French version features direct references to sociodemographic groups. These data differences may stem not only from cultural influences but also from the demographics of the data contributors. Where and from whom one collects data matters, as it can result in dramatic differences in data and modelling.

Demographic information is also important in annotation (Sap et al., 2022; Pei and Jurgens, 2023; Santy et al., 2023), where a piece of text can be humorous to some people but offensive to others (Meaney, 2020). In such cases, culture may exist in the labels rather than in the data. Recently, Lee et al. (2023b); Frenda et al. (2023) show how to capture different cultural views of annotators using the same dataset.

4 Culturally Aware Resource Acquisition

Resources like those discussed in §3 are critical for developing culturally aware NLP. As additional resources are much needed, this section surveys methods for creating new resources.

Resources can be classified based on their acquisition methods—manual, automatic, or semi-automatic—and their source types: 1) new ones (New), or 2) culturally adapted existing ones (CA, e.g., through translation from the original data, followed by culturally appropriate changes). 1) can capture unique cultural phenomena but this is not always possible due to limited funds or access to experts or native speakers. 2) can offer an alternative, but an accurate account of cultural phenomena may prove more challenging.

4.1 Manual: Incorporating Native Speakers, Communities, and Experts

A common strategy is to employ native speakers or experts (e.g., professional translators or students) for data acquisition. This can be done via crowd-sourcing platforms such as Amazon Mechanical Turk and Prolific (Liu et al., 2021, 2023b) or in a community-driven manner, leveraging networks such as Masakhane ⁸, IndoNLP ⁹, university mailing lists, or Slack/Discord of organizations.

⁸https://www.masakhane.io/

⁹https://indonlp.github.io/

New: Most existing culture resources have been built by involving native speakers or communities for dataset acquisition. Examples include Liu et al. (2021); Koto et al. (2023); Kabra et al. (2023); Liu et al. (2023b), among others. For non-language related communities, Wino-Queer (Felkner et al., 2023) utilizes diverse channels (such as Slacks/Discord, gay Twitter) to reach the LGBTQ+ community and generates bias benchmarks based on community survey results.

CA: When starting from existing datasets, some works have involved communities (e.g., using surveys) to determine the needed modifications and supplements to datasets (Névéol et al., 2022; Jin et al., 2024; Hu et al., 2023).

In general, native speakers are typically consulted throughout the life-cycle of new data acquisition (from annotations to quality checks). However, the entire community is rarely consulted during the initiation stage (i.e., designing tasks). Involving native speakers may be costly and difficult sometimes, but the quality they bring is invaluable.

4.2 Automatic: Models and Pipelines

Since manual adaptation is slow and hard to scale, the use of automation has gained popularity in resource acquisition.

New: For instance, CANDLE (Nguyen et al., 2023) proposes a pipeline to extract cultural commonsense knowledge using various techniques like NER extraction, cultural facet classification, concepts extraction and ranking through algorithms or LMs. NormsSAGE (Fung et al., 2023) utilizes LLMs for norm discovery from conversation data, then performs model self-verification to validate and filter the data. CultureAtlas (Fung et al., 2024) extracts cultural knowledge from Wikipedia and hyperlinked document pages using LLMs for filtering and adversarial knowledge generation.

Recent works have also used sociodemographic prompting (Deshpande et al., 2023; Beck et al., 2024; Hwang et al., 2023; Santurkar et al., 2023) — extending input prompts with sociodemographic information — to generate outputs tailored to specific groups. Further research could potentially lead to a reduction of effort in data acquisition in specific areas, such as creating subcultural variations of data in WEIRD people. However, it has also been argued that LLMs do not accurately mimic individual or group behaviours (Argyle et al., 2023; Aher et al., 2023; Beck et al., 2024).

CA: Putri et al. (2024) investigate automatic adaptation of Commonsense QA in Indonesian and Sundanese, where the adaptation process includes paraphrasing and concepts/names replacements. Current GPT models show disparities in automatic cultural adaptations for different languages, indicating the need for further research.

4.3 Semi-Automatic: Structured Resources, Model-in-the-Loop

As demonstrated by Putri et al. (2024), LLMs struggle with fully automated cultural adaptations. Alternatively, semi-automatic approaches combine the quality of manual work with scalability.

New: Methods have been used to generate seed data for humans to clean and label, possibly iteratively. NormBank (Ziems et al., 2023a) uses LLMs to generate an initial set of roles and behaviours for creating detailed norm candidates grounded in situations, to be annotated by humans. Similarly, CH-Wang et al. (2023) build a dataset of cross-cultural norms for Chinese using few-shot and Chain of Thought (CoT) prompting (Wei et al., 2022) with GPT-3 (Brown et al., 2020) to generate norm candidates, followed by human verification and editing. Liu et al. (2023b) utilize LLMs to generate conversations featuring proverbs in different languages. Since many proverbs are figurative, most conversations are revised by human native speakers. SeeGULL (Bhutani et al., 2024) generates candidate multilingual associations using PaLM-2 (Anil et al., 2023) and few-shot demonstrations, then performs human annotations.

CA: For dialects, Ziems et al. (2023b, Multi-Value) create a framework based on well-documented dialect differences in English by utilizing Electronic World Atlas of Varieties of English (Kortmann et al., 2020, eWAVE), which allowed adaptation and creation of dialectic datasets that span 50 English dialects. Multi-Value was used in adapting standard corpus to dialectal variants in Held et al. (2023); Xiao et al. (2023). However, such structured resources for adaptations may not exist or be suitable for other elements of culture (e.g., for concepts, consistently replacing 'bread' with 'rice' would not be desirable).

5 Creating Culturally Adapted Models

Most work in culturally aware NLP has so far focused on resource creation and evaluation. Work on creating culturally adapted models is still emerg-



Figure 3: Categorization of the adaptation modelling methods and examples in each category.

ing. In this section, we survey the methods that have been proposed to date for creating culturally adapted models from pre-trained (L)LMs. Figure 3 shows an overview of the approaches adopted so far with example approaches. Some individual works may incorporate more than one adaptation method for comparison; we may discuss works in relation to the main method adopted.

5.1 Data Augmentation

It can be challenging to acquire large corpora for supervised training in cultural adaptation. Data augmentation has been supporting this process and increasing the robustness of models. Li and Zhang (2023) present a data augmentation technique (CultureMixup) for multilingual multicultural VL reasoning tasks. CultureMixup creates code-mixed data with concepts in English replaced through cultural concept mapping for supervised training. The cultural concept sets (for concept mapping) are constructed by querying hyponyms, synonyms, and hypernyms in the ConceptNet (Speer et al., 2017) and WordNet (Miller, 1995). However, the optimal resource will depend on the cultural element in question (§3.1). For example, a cultural knowledge base might be better for norms adaptations.

5.2 Continual Pre-training, Auxiliary Losses

Continual pre-training (CPT, including instruction tuning), intermediate task training, and utilizing multi-task or auxiliary losses have also been used for cultural adaptation. CPT involves tuning a pre-trained LM with an additional unlabeled domain or language corpus before downstream task fine-tuning. It can improve downstream task performance via full-parameter training (Xu et al., 2019; Han and Eisenstein, 2019; Gururangan et al., 2020) or by training a small set of additional parameters while keeping the original model frozen (Pfeiffer et al., 2020; Wang et al., 2021; Ke et al., 2022).

Recently, Hofmann et al. (2024) show that when combined with a geo-location prediction loss, CPT

can help to increase the awareness of dialectal variations of pre-trained LMs. Wang et al. (2024) show that mining instructions containing cultural knowledge and performing instruction tuning can improve models' ability in cultural knowledge reasoning. In VL settings, Bhatia and Shwartz (2023) use a cultural common sense knowledge graph from (Nguyen et al., 2023) for CPT, developing a geodiverse LM for downstream commonsense reasoning tasks. CultureLLM (Li et al., 2024) reformulates the World Value Survey as an intermediate training task. It shows that LLMs can enhance their downstream performance in multicultural hate speech classification where diverse value perceptions across cultures are required. However, adapting LLMs to culture may result in catastrophic forgetting (McCloskey and Cohen 1989, or termed "alignment tax" due to RLHF tuning, Ouyang et al. 2022; Askell et al. 2021), potentially worsening their performance on general tasks. Similarly, general continual fine-tuning may shift existing cultural values (Choenni et al., 2024) in pretrained models. These warrant further investigation.

5.3 Other Forms of Information Integration

Missing cultural context, knowledge or demographic information have also been integrated into models in other ways:

As representations. Cao et al. (2024a) introduce the cuDialog dataset and models for multi-turn dialogue classification and prediction through a cultural lens. Using an encoder-decoder Transformer (such as mT5, Xue et al. 2021), the cultural dimensions vectors are concatenated to the hidden states at each layer. The cultural context vectors were obtained through a regression task over the encoder hidden states with dialogue inputs to predict the corresponding cultural scores (based on Hofstede Culture Dimensions, Hofstede 1984).

As part of the prompt. AlKhamissi et al. (2024) enhance LLMs' cultural alignment by integrating anthropological reasoning with demographic-based prompts. In a game setting with multicultural players, Shaikh et al. (2023) encode sociocultural attributes into the input to enrich the model's ability for game plays. Similarly, Yao et al. (2023) propose strategies to enhance the cultural translation ability of LLMs with reference-explanation of cultural entities. LLMs effectively leverage indirect descriptions obtained from external data sources or generated by LLMs as a prior step.

Friedrich et al. (2023) propose a retrieval-based method for moral reasoning tasks. Moral contexts specific to a culture (or user) are stored in a retrieval engine. When querying with moral questions, relevant contexts are retrieved and prepend to the query as input to LLMs. LLMs can change their default answer based on the retrieved context. Using a retrieval-augmented system can be a promising future direction for adapting rich and evolving cultural information.

5.4 Parameter-Efficient Adaptations

As LMs grow larger, parameter-efficient fine-tuning methods (i.e., PEFT, by fine-tuning a small number of parameters, such as the bottle-neck adapters, Houlsby et al. 2019; LoRA, Hu et al. 2022 etc.) become increasingly important for task adaptations. Given their success in cross-lingual transfer learning (Pfeiffer et al., 2020; Ansell et al., 2021; Liu et al., 2023a,c; Üstün et al., 2020, among others), PEFT can be a natural choice for cultural adaptation of e.g., dialects.

Recently, HyperLoRA (Xiao et al., 2023) uses the Hypernetwork (Ha et al., 2017, a neural network for generating parameters) to generate LoRA adapters based on dialectal features. DADA (Liu et al., 2023d) proposes to train a pool of dialectal linguistic feature adapters and dynamically compose the adapters for dialectal tasks. Being task agnostic, PEFT methods could prove important for cultural adaptations beyond dialects.

5.5 Outlook: Reinforcement Learning from Human Feedback (RLHF), Direct Preference Optimization (DPO)

The success of LLMs has popularized RLHF (Christiano et al., 2017; Bai et al., 2022; Ouyang et al., 2022) and DPO (Rafailov et al., 2023; Ivison et al., 2023) methods. RLHF fine-tunes LMs with feedback by fitting a reward model with human preferences, and then training a reinforcement learning-based policy to maximize the learned reward. DPO avoids RL training by using a simpler supervised learning objective for an implicit reward model.

Recent work shows that RLHF can enhance the performance of multilingual instruction tuning for LLMs (Lai et al., 2023), while DPO can improve the multilingual reasoning abilities (She et al., 2024) of LLMs. The use of RLHF or DPO for multilingual multicultural adaptation is still limited, but these examples suggest that the direction

could be promising.

6 Further Discussions

As we have seen, "culture" emerges as a highly active area of NLP research. Yet significant work remains to be done on both resources and methods for various elements of culture.

An area that requires attention is the overall process of researching culturally aware NLP. First, it is important to consider whether and how the target communities could best benefit from cultural technologies at this point in time. For example, many dialects are primarily oral, and speech-to-speech or speech-to-text translations may be preferable over text-based applications (Blaschke et al., 2024). Or, if the culture in question does not have LLMs with basic conversational fluency yet, it may be more important to address this problem first, before embarking on improving pragmatic alignments of LLMs for the culture. Getting the process right requires consultation with target communities (Bird, 2020; Liu et al., 2022b; Mager et al., 2023). Ethical data collection practices are also critical and technology ownership must be considered. We refer the readers to work such as Bird (2020); Smith (2021); Cooper et al. (2024) for further details of this.

Another area to consider is the integration of insights and practices from fields beyond NLP. Incorporating cultural factors and performing cultural adaptations has a long history in application areas, such as for video games (O'hagan and Mangiron, 2013), movies (Pettit, 2009), online learning systems (Blanchard et al., 2005), and clinical psychology intervention (Bernal et al., 1995; Barrera Jr and Castro, 2006). Existing practices should be used as a foundation for adapting NLP applications to meet the needs of diverse cultural contexts.

7 Summary and Future Research Directions

Culturally aware and adapted NLP has recently emerged as a highly active research area. Significant progress has been made in the development of resources for capturing various elements of culture, but the development of NLP methods is still in its infancy. We will now summarize the main research gaps identified in this survey with respect to the categories of our new taxonomy (§2 for details): **Resources.** Currently, resources exist for all elements of culture, with considerable progress made on *values* (§3.1.3, particularly in biases) and *knowl*-

edge (§3.1.2, particularly for MMLU-style cultural knowledge benchmarks). However, research is lacking in the following areas:

<u>Cultural</u>: Multilingual data resources covering a diverse set of concepts (e.g., aesthetics, spatial relation) in both unimodal and multimodal (§3.1.1) for generation tasks are needed. Moreover, most recent developments in *norms & morals* are predominantly in English and reflect a monocultural perspective, which indicates the need for further development of multilingual and multicultural resources. Additionally, there is a significant gap in datasets that focus on different types of value perceptions (such as emotion and irony, §3.1.3), stylistic variations (§3.1.5), and artifacts (§3.1.6) across various cultural groups both in different languages and within languages.

<u>Sociocultural:</u> Resources that explicitly consider sociocultural elements of culture (§3.2) are still lacking. For example, collecting speaker relationships in dialogue datasets or distinguishing age groups in social norms datasets. These are needed to address the intricate relationship between culture and people in NLP.

Training Data and "CultureGLUE": Most existing resources focus on evaluation, but there is a pressing need for training data. In addition, there is no unified cultural equivalent of the GLUE benchmark that encompasses all elements of culture across a wide range of cultures. Creating a multicultural "CultureGLUE" may be challenging at the moment, a good first step is to focus on individual cultures with diverse tasks and element coverage.

Modelling. While modelling methods for culture are generally under-explored, continual pretraining (§5.2) and prompting (§5.3) have received marginally more attention than other approaches. Research areas needing further exploration include: PEFT-based Transfer Learning: Exploration of PEFT-based transfer learning techniques beyond dialects is limited (§5.4). Given their success in other NLP areas, these techniques warrant further investigation into other elements of culture, such as *values* or *norms & morals*.

RLHF, DPO and Other LLMs Specialties: Leveraging the success of LLMs, RLHF and DPO as a new promising avenue for cultural adaptation (§5.5). It would also be interesting to explore other unique abilities of LLMs (such as tool-use, role-playing etc.) for enabling culturally aware NLP. Evolving Culture: Culture (slowly) evolves (Boyd

and Richerson, 1988; Whiten et al., 2011), yet there have been little discussions on how to model and adapt to evolving culture. One potential approach is the use of retrieval-augmented systems to integrate evolving information (§5.3), which ensures models' relevance to cultural shifts over time.

Overall - "Surface" versus "Deep" culturally adapted NLP. Resnicow et al. (1999) devise cultural adaptations for public health research into surface and deep adaptations, where the former considers familiar languages and concepts to the target groups, and the latter considers social and historical factors that influence the behaviours of the target groups. In NLP, surface adaptations might include using the same language as a culture and recognizing explicit cultural differences (e.g., asking LLMs "what is the meaning of ..."). In contrast, deep adaptations might enable a model to "behave" (e.g., make decisions, pragmatically comply etc.) like a member of a culture without explicit inquisition (see Figure 4 in the Appendix for an illustration).

Only a few current works focus on adapting the behavioural aspect of models (which is becoming increasingly important with LLMs), and there has been no work to date measuring the depth and progress of cultural adaptations or when a model is *fully culturally aware and culturally competent*. Further research could explore these areas.

8 Conclusions

This paper proposes a new extensive taxonomy of culture that expands on earlier works in NLP and is grounded in well-established anthropology literature. The taxonomy provides a systematic framework for understanding and tracking progress in the emerging area of culturally aware and adapted NLP. We survey existing resources and methods in this area according to the taxonomy classes, identifying areas of strength as well as areas where research remains to be done. Our paper summarizes the state of the art and provides ideas for future development in this exciting and important area of research.

9 Limitations

In this work, we may only discuss research with respect to a single element of culture. Research related to culture can be multifaceted, considering multiple elements of culture simultaneously. Additionally, our survey emphasizes resources and adaptation methods due to the scope, without extensive coverage of evaluation and probing techniques.

Our taxonomy provides a systematic framework for understanding and tracking progress in the emerging area of culturally aware and adapted NLP. However, our taxonomy is not without its limitations. Future research could refine the taxonomy in areas like *values* or *communicative goals*, create subcategories and enhancing our understanding of how elements of culture interact (e.g., how changing values impact social norms over time).

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 13N15897 (MISRIK). This work has also been supported by the UK Research and Innovation (UKRI) Frontier Research Grant EP/Y031350/1 EQUATE (the UK government's funding guarantee for ERC Advanced Grants) awarded to Anna Korhonen at the University of Cambridge. Chen Cecilia Liu is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

We thank Anjali Kantharuban, Shun Shao, and Anne Lauscher for an early discussion of different aspects of this work. We thank Ji-Ung Lee for feedback on a draft of this paper.

References

- Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.
- Muhammad Farid Adilazuarda, 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.
- Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. A benchmark for evaluating machine translation metrics on dialects without standard orthography. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1045–1065, Singapore. Association for Computational Linguistics.
- Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.
- Idris Akinade, 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, Dubrovnik, Croatia. Association for Computational Linguistics.
- Badr AlKhamissi, Muhammad N. ElNokrashy, Mai AlKhamissi, and Mona T. Diab. 2024. Investigating cultural alignment of large language models. *ArXiv preprint*, abs/2402.13231.
- Nalini Ambady, Jasook Koo, Fiona Lee, and Robert Rosenthal. 1996. More than words: Linguistic and nonlinguistic politeness in two cultures. *Journal of Personality and Social Psychology*, 70(5):996.
- Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2023. SODAPOP: Open-ended discovery of social biases in social commonsense reasoning models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1573–1596, Dubrovnik, Croatia. Association for Computational Linguistics.

- Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. PaLM 2 technical report. ArXiv preprint, abs/2305.10403.
- Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavas, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: multilingual adapter generation for efficient cross-lingual transfer. In Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, pages 4762–4781. Association for Computational Linguistics.
- Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.
- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *ArXiv preprint*, abs/2112.00861.
- Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan,

- Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv preprint*, abs/2204.05862.
- Manuel Barrera Jr and Felipe González Castro. 2006. A heuristic framework for the cultural adaptation of interventions. *Clinical Psychology: Science and Practice*, 13(4):311–316.
- Tilman Beck, 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, EACL 2024 Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2589–2615. Association for Computational Linguistics.
- Guillermo Bernal, Janet Bonilla, and Carmen Bellido. 1995. Ecological validity and cultural sensitivity for outcome research: Issues for the cultural adaptation and development of psychosocial treatments with hispanics. *Journal of abnormal child psychology*, 23:67–82.
- Mehar Bhatia 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, Singapore. Association for Computational Linguistics.
- Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeG-ULL multilingual: a dataset of geo-culturally situated stereotypes. *ArXiv preprint*, abs/2403.05696.
- Cristina Bicchieri, Ryan Muldoon, and Alessandro Sontuoso. 2018. Social norms. *The Stanford encyclopedia of philosophy*.
- Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Janet Blake. 2000. On defining the cultural heritage. *International and Comparative Law Quarterly*, 49(1):61–85.
- Emmanuel Blanchard, Ryad Razaki, and Claude Frasson. 2005. Cross-cultural adaptation of e-learning contents: A methodology. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, pages 1895–1902. Association for the Advancement of Computing in Education (AACE).

- Verena Blaschke, Christoph Purschke, Hinrich Schütze, and Barbara Plank. 2024. What do dialect speakers want? A survey of attitudes towards language technology for german dialects. *ArXiv preprint*, abs/2402.11968.
- Shoshana Blum-Kulka. 1987. Indirectness and politeness in requests: Same or different? *Journal of pragmatics*, 11(2):131–146.
- Shoshana Blum-Kulka and Elite Olshtain. 1984. Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). *Applied linguistics*, 5(3):196–213.
- Robert Boyd and Peter J Richerson. 1988. *Culture* and the evolutionary process. University of Chicago press.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.
- Yong Cao, Min Chen, and Daniel Hershcovich. 2024a. Bridging cultural nuances in dialogue agents through cultural value surveys. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 929–945. Association for Computational Linguistics.
- Yong Cao, 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.
- Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of*

- the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.
- Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3548–3564, Singapore. Association for Computational Linguistics.
- Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during lm fine-tuning. *ArXiv* preprint, abs/2405.12744.
- Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 4299–4307.
- Robert B Cialdini, Carl A Kallgren, and Raymond R Reno. 1991. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology*, volume 24, pages 201–234. Elsevier.
- Ned Cooper, Courtney Heldreth, and Ben Hutchinson. 2024. "It's how you do things that matters": Attending to process to better serve indigenous communities with language technologies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 204–211, St. Julian's, Malta. Association for Computational Linguistics.
- Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: The case of Bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83, Dubrovnik, Croatia. Association for Computational Linguistics.
- Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

- Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.
- Ameet Deshpande, 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, Singapore. Association for Computational Linguistics.
- Sunipa Dev, 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, New Orleans, LA, USA, December 10 16, 2023.
- Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022.
 On measures of biases and harms in NLP. In Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 246–267, Online only. Association for Computational Linguistics.
- Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *ArXiv preprint*, abs/2306.16388.
- Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.
- AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.
- Cristina España-Bonet 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, Abu Dhabi, United

- Arab Emirates. Association for Computational Linguistics.
- EVS. 2011. EVS European Values Study 1981 integrated dataset. GESIS Datenarchiv, Köln. ZA4438 Datenfile Version 3.0.0, https://doi.org/10.4232/1.10791.
- Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: A NLP benchmark for dialects, varieties, and closely-related languages. *ArXiv preprint*, abs/2403.11009.
- Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A community-in-the-loop benchmark for anti-lgbtq+bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9126–9140. Association for Computational Linguistics.
- Maxwell Forbes, 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, Online. Association for Computational Linguistics.
- Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.
- Felix Friedrich, Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. 2023. Revision transformers: Instructing language models to change their values. In ECAI 2023 26th European Conference on Artificial Intelligence, September 30 October 4, 2023, Kraków, Poland Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023), volume 372 of Frontiers in Artificial Intelligence and Applications, pages 756–763. IOS Press.
- Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORM-SAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.
- Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & LM benchmarking. *ArXiv preprint*, abs/2402.09369.

- Bernard Gert and Joshua Gert. 2002. The definition of morality.
- Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 8342–8360. Association for Computational Linguistics.
- David Ha, Andrew M. Dai, and Quoc V. Le. 2017. Hypernetworks. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.
- Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. 2020. PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1652–1663, Marseille, France. European Language Resources Association.
- Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. Finnish dialect identification: The effect of audio and text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8777–8783, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4237–4247. Association for Computational Linguistics.
- Einar Haugen. 1966. Dialect, language, nation 1. *American anthropologist*, 68(4):922–935.
- Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle H. 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, WASSA@ACL 2023, Toronto, Canada, July 14, 2023*, pages 202–214. Association for Computational Linguistics.
- Michael Hechter and Karl-Dieter Opp. 2001. *Social norms*. Russell Sage Foundation.
- William Held, Caleb Ziems, and Diyi Yang. 2023. TADA: Task agnostic dialect adapters for English. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 813–824, Toronto, Canada. Association for Computational Linguistics.

- Dan Hendrycks, 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, Virtual Event, Austria, May 3-7, 2021. OpenReview.net.
- Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.
- Daniel Hershcovich, Stella Frank, Heather C. 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 crosscultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 6997–7013. Association for Computational Linguistics.
- Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.
- Valentin Hofmann, Goran Glavaš, Nikola Ljubešić, Janet B. Pierrehumbert, and Hinrich Schütze. 2024. Geographic Adaptation of Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 12:411–431.
- G. Hofstede. 1984. *Culture's Consequences: International Differences in Work-Related Values*. Cross Cultural Research and Methodology. SAGE Publications.
- Geert Hofstede. 2011. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):8.
- Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 2790–2799. PMLR.

- Juliane House. 2005. *Politeness in Germany: Politeness in GERMANY?* Multilingual Matters, Bristol, Blue Ridge Summit.
- Juliane House and Gabriele Kasper. 1981. *Politeness Markers in English and German*, pages 157–186. De Gruyter Mouton, Berlin, New York.
- Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "You sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.
- Dirk Hovy 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, Online. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023. Multi3WOZ: 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.
- Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP* 2023, pages 7591–7609, Singapore. Association for Computational Linguistics.
- EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.
- Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew E. Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing LM adaptation with Tulu 2. *ArXiv preprint*, abs/2311.10702.
- Ray Jackendoff. 2012. What is a concept? In *Frames*, *Fields*, *and Contrasts*, pages 191–208. Routledge.
- Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. 2024. ViSAGe: A global-scale analysis of visual stereotypes in text-to-image generation. *ArXiv preprint*, abs/2401.06310.

- Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.
- Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.
- Rebecca L. Johnson, 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.
- Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Indra Winata, Samuel Cahyawijaya, Aremu Anuoluwapo, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8269–8284. Association for Computational Linguistics.
- Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.
- Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.
- Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu, and Bing Liu. 2022. Continual training of language models for few-shot learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10205–10216. Association for Computational Linguistics.
- Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.
- Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image

- speaks a thousand words, but can everyone listen? on translating images for cultural relevance. *ArXiv preprint*, abs/2404.01247.
- Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in korean. *ArXiv preprint*, abs/2403.06412.
- Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *ArXiv preprint*, abs/2404.16019.
- Bernd Kortmann, Kerstin Lunkenheimer, and Katharina Ehret, editors. 2020. *eWAVE*.
- Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12359–12374. Association for Computational Linguistics.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda 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. *ArXiv* preprint, abs/2402.12840.
- Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024b. IndoCulture: Exploring geographically-influenced cultural commonsense reasoning across eleven indonesian provinces. *ArXiv* preprint, abs/2404.01854.
- Alfred Louis Kroeber and Clyde Kluckhohn. 1952. Culture: A critical review of concepts and definitions. Papers. Peabody Museum of Archaeology & Ethnology, Harvard University.
- Julia Kruk, Caleb Ziems, and Diyi Yang. 2023. Impressions: Visual semiotics and aesthetic impact understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12273–12291, Singapore. Association for Computational Linguistics.

- Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP* 2023, pages 13814–13828, Singapore. Association for Computational Linguistics.
- Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Okapi: Instructiontuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 System Demonstrations, Singapore, December 6-10, 2023*, pages 318–327. Association for Computational Linguistics.
- Thang Le and Anh Luu. 2023. A parallel corpus for Vietnamese central-northern dialect text transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13839–13855, Singapore. Association for Computational Linguistics.
- Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023a. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.
- Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Juho Kim, and Alice Oh. 2023b. CReHate: Cross-cultural re-annotation of english hate speech dataset. *ArXiv* preprint, abs/2308.16705.
- Nayeon Lee, Chani Jung, and Alice Oh. 2023c. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.
- Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. CultureLLM: Incorporating cultural differences into large language models. *ArXiv preprint*, abs/2402.10946.
- Chengxi Li, 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, Singapore. Association for Computational Linguistics.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023b. CMMLU: measuring massive multitask language understanding in chinese. *ArXiv preprint*, abs/2306.09212.

- Oliver Li, Mallika Subramanian, Arkadiy Saakyan, Sky CH-Wang, and Smaranda Muresan. 2023c. Norm-Dial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15732–15744, Singapore. Association for Computational Linguistics.
- Zhi Li and Yin Zhang. 2023. Cultural concept adaptation on multimodal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 262–276, Singapore. Association for Computational Linguistics.
- Chen Liu, 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Chen Liu, Jonas Pfeiffer, Anna Korhonen, Ivan Vulić, and Iryna Gurevych. 2023a. Delving deeper into cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2408–2423. Association for Computational Linguistics.
- Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023b. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *ArXiv preprint*, abs/2309.08591.
- Chen Cecilia Liu, Jonas Pfeiffer, Ivan Vulić, and Iryna Gurevych. 2023c. Improving generalization of adapter-based cross-lingual transfer with scheduled unfreezing. *ArXiv preprint*, abs/2301.05487.
- Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10467–10485. Association for Computational Linguistics.
- Yanchen Liu, William Held, and Diyi Yang. 2023d. DADA: Dialect adaptation via dynamic aggregation of linguistic rules. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13776–13793, Singapore. Association for Computational Linguistics.
- Zhixuan Liu, Youeun Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. 2023e. Towards equitable representation in text-to-image synthesis models with the cross-cultural understanding benchmark (CCUB) dataset. *ArXiv preprint*, abs/2301.12073.

- Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022b. 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, Dublin, Ireland. Association for Computational Linguistics.
- Brandon Lwowski, Paul Rad, and Anthony Rios. 2022. Measuring geographic performance disparities of offensive language classifiers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6600–6616, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.
- Olga Majewska, Evgeniia Razumovskaia, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Trans. Assoc. Comput. Linguistics*, 11:139–156.
- Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip C. 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.
- Takahiko Masuda, Richard Gonzalez, Letty Kwan, and Richard E Nisbett. 2008. Culture and aesthetic preference: Comparing the attention to context of east asians and americans. *Personality and Social Psychology Bulletin*, 34(9):1260–1275.
- David Matsumoto and Linda Juang. 1996. Culture and psychology. *Pacific Grove*, pages 266–270.
- Yoshiko Matsumoto. 1988. Reexamination of the universality of face: Politeness phenomena in japanese. *Journal of pragmatics*, 12(4):403–426.
- Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.
- Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

- J. A. Meaney. 2020. Crossing the line: Where do demographic variables fit into humor detection? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 176–181, Online. Association for Computational Linguistics.
- Batja Mesquita, Nico H Frijda, and Klaus R Scherer. 1997. Culture and emotion. *Handbook of cross-cultural psychology: Basic processes and human development*, 2:255.
- George A. Miller. 1995. WordNet: A lexical database for english. *Commun. ACM*, 38(11):39–41.
- Farhad Moghimifar, Shilin Qu, Tongtong Wu, Yuan-Fang Li, and Gholamreza Haffari. 2023. Norm-Mark: A weakly supervised Markov model for socio-cultural norm discovery. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5081–5089, Toronto, Canada. Association for Computational Linguistics.
- Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022. ArtELingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 8770–8785. Association for Computational Linguistics.
- Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global Voices, local biases: Socio-cultural prejudices across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15828–15845, Singapore. Association for Computational Linguistics.
- Tarek Naous, Michael J. Ryan, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *ArXiv preprint*, abs/2305.14456.
- Aurélie Névéol, 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), ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 8521–8531. Association for Computational Linguistics.
- Tuan-Phong Nguyen, Simon Razniewski, Aparna S. Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 4 May 2023*, pages 1907–1917. ACM.
- Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Multi-cultural commonsense knowledge distillation. *ArXiv preprint*, abs/2402.10689.

- Minako O'hagan and Carmen Mangiron. 2013. *Game Localization: Translating for the global digital entertainment industry*, volume 106. John Benjamins Publishing.
- Olubusayo Olabisi, Aaron Hudson, Antonie Jetter, and Ameeta Agrawal. 2022. Analyzing the dialect diversity in multi-document summaries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6208–6221, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Longshen Ou, Xichu Ma, Min-Yen Kan, and Ye Wang. 2023. Songs across borders: Singable and controllable neural lyric translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–467, Toronto, Canada. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 December 9, 2022.
- Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9952–9962. Association for Computational Linguistics.
- Claudio Paonessa, Yanick Schraner, Jan Deriu, Manuela Hürlimann, Manfred Vogel, and Mark Cieliebak. 2023. Dialect transfer for Swiss German speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15240–15254, Singapore. Association for Computational Linguistics.
- Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.
- Zoë Pettit. 2009. *3: Connecting Cultures: Cultural Transfer in Subtitling and Dubbing*, pages 44–57. Multilingual Matters, Bristol, Blue Ridge Summit.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),

- pages 7654–7673, Online. Association for Computational Linguistics.
- Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.
- Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.
- Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM generate culturally relevant commonsense QA data? case study in indonesian and sundanese. *ArXiv preprint*, abs/2402.17302.
- Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11253– 11271, Toronto, Canada. Association for Computational Linguistics.
- Rafael Rafailov, 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, New Orleans, LA, USA, December 10 16, 2023.
- Sunny Rai, Khushang Jilesh Zaveri, Shreya Havaldar, Soumna Nema, Lyle H. Ungar, and Sharath Chandra Guntuku. 2024. A cross-cultural analysis of social norms in bollywood and hollywood movies. *ArXiv* preprint, abs/2402.11333.
- Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.
- Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. 2023. What makes it ok to set a fire? iterative self-distillation of contexts and rationales for disambiguating defeasible social and moral situations.

- In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 12140–12159, Singapore. Association for Computational Linguistics.
- Ken Resnicow, Tom Baranowski, Jasjit S. Ahluwalia, and Ronald L. Braithwaite. 1999. Cultural sensitivity in public health: Defined and demystified. *Ethnicity & Disease*, 9(1):10–21.
- Dor Ringel, Gal Lavee, Ido Guy, and Kira Radinsky. 2019. Cross-cultural transfer learning for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3873–3883, Hong Kong, China. Association for Computational Linguistics.
- William Gaviria Rojas, 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, New Orleans, LA, USA, November 28 December 9, 2022.
- Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Sandra Sandoval, Jieyu Zhao, Marine Carpuat, and Hal Daumé III. 2023. A rose by any other name would not smell as sweet: Social bias in names mistranslation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3945, Singapore. Association for Computational Linguistics.
- Shibani Santurkar, 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, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.
- Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.
- Maarten Sap, 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, NAACL*

- 2022, Seattle, WA, United States, July 10-15, 2022, pages 5884–5906. Association for Computational Linguistics.
- Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.
- Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.
- Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569, Toronto, Canada. Association for Computational Linguistics.
- Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.
- Shuaijie She, Shujian Huang, Wei Zou, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: advancing multilingual reasoning through multilingual alignment-as-preference optimization. *ArXiv* preprint, abs/2401.06838.
- Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. *ArXiv preprint*, abs/2404.15238.
- Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.
- Linda Tuhiwai Smith. 2021. *Decolonizing methodologfies: Research and indigenous peoples*. Bloomsbury Publishing.
- Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. KMMLU: measuring massive multitask language understanding in korean. *ArXiv preprint*, abs/2402.11548.
- Robyn Speer, 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. AAAI Press.
- Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.
- Ian Stewart and Rada Mihalcea. 2024. Whose wife is it anyway? assessing bias against same-gender relationships in machine translation. *ArXiv preprint*, abs/2401.04972.
- Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. 2023. MoralDial: A framework to train and evaluate moral dialogue systems via moral discussions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2213–2230, Toronto, Canada. Association for Computational Linguistics.
- Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.
- Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9882–9902. Association for Computational Linguistics.
- Ashish V. Thapliyal, 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 715–729. Association for Computational Linguistics.
- Jenny Thomas. 1983. Cross-cultural pragmatic failure. *Applied linguistics*, 4(2):91–112.
- Edward Burnett Tylor. 1871. *Primitive culture: Researches into the development of mythology, philosophy, religion, art and custom,* volume 2. J. Murray.
- UNESCO. 1982. World conference on cultural policies, mexico city, final report.
- Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly universal dependency parsing. In *Proceed*ings of the 2020 Conference on Empirical Methods in

- Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 2302–2315. Association for Computational Linguistics.
- Mor Ventura, 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.
- Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, ethics, morals? on the use of moral concepts in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554, Singapore. Association for Computational Linguistics.
- Bin Wang, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy F. Chen. 2024. CRAFT: Extracting and tuning cultural instructions from the wild. *ArXiv* preprint, abs/2405.03138.
- Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F. Chen. 2023a. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *ArXiv* preprint, abs/2309.04766.
- Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of Findings of ACL, pages 1405–1418. Association for Computational Linguistics.
- Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R. Lyu. 2023b. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *ArXiv preprint*, abs/2310.12481.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 December 9, 2022.
- Leslie A White. 1959. The concept of culture. *American anthropologist*, 61(2):227–251.
- Andrew Whiten, Robert A Hinde, Kevin N Laland, and Christopher B Stringer. 2011. Culture evolves. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567):938–948.
- Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. COPAL-ID: indonesian language reasoning with local culture and nuances. *ArXiv preprint*, abs/2311.01012.

- Anna Wierzbicka. 1992. Semantics, culture, and cognition: Universal human concepts in culture-specific configurations. Oxford University Press.
- Anna Wierzbicka. 2003. *Cross-cultural pragmatics: The semantics of human interaction*. De Gruyter Mouton, Berlin, Boston.
- Zedian Xiao, William Held, Yanchen Liu, and Diyi Yang. 2023. Task-agnostic low-rank adapters for unseen English dialects. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7857–7870, Singapore. Association for Computational Linguistics.
- Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 2324–2335. Association for Computational Linguistics.
- Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA—an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.
- Linting Xue, 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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.
- Diyi Yang. 2019. *Computational Social Roles*. Ph.D. thesis, Carnegie Mellon University Pittsburgh, PA, USA.
- Zhichao Yang, Pengshan Cai, Yansong Feng, Fei Li, Weijiang Feng, Elena Suet-Ying Chiu, and Hong Yu. 2019. Generating classical Chinese poems from vernacular Chinese. 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 6155–6164, Hong Kong, China. Association for Computational Linguistics.
- Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Benchmarking llm-based machine translation on cultural awareness. *ArXiv preprint*, abs/2305.14328.

- Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoM-LAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2039–2055. Association for Computational Linguistics.
- Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geodiverse visual commonsense reasoning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 2115–2129. Association for Computational Linguistics.
- Mahmoud Yusuf, Marwan Torki, and Nagwa El-Makky. 2022. Arabic dialect identification with a few labeled examples using generative adversarial networks. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 196–204, Online only. Association for Computational Linguistics.
- Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, Suraj Sharma, Shilin Qu, Linhao Luo, Lay-Ki Soon, Zhaleh Semnani-Azad, Ingrid Zukerman, and Gholamreza Haffari. 2024. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. *ArXiv preprint*, abs/2402.11178.
- Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2023. SocialDial: A benchmark for socially-aware dialogue systems. SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, page 2712–2722, New York, NY, USA. Association for Computing Machinery.
- Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, and Mykola Pechenizkiy. 2023. CHBias: Bias evaluation and mitigation of chinese conversational language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 13538–13556. Association for Computational Linguistics.
- Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. World-ValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*

- Resources and Evaluation (LREC-COLING 2024), pages 17696–17706, Torino, Italia. ELRA and ICCL.
- Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3576–3591, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023a. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceed*ings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pages 8–15, Dubrovnik, Croatia. Association for Computational Linguistics.
- Li Zhou, 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, Singapore, December 6-10, 2023*, pages 12684–12702. Association for Computational Linguistics.
- Li Zhou, 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.
- Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022a. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.
- Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023a. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.
- Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023b. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.
- Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022b. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

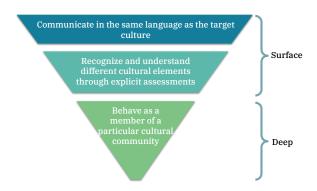


Figure 4: A proposal framework of surface versus deep culturally adapted NLP model.

A Method

We prioritize main and findings papers from ACL, EMNLP, EACL, NAACL and TACL published in 2022, 2023 and early 2024. The search terms used are: "culture", "cultural", "geo-diverse", "socio", "social", "moral", "bias", "norms" and we prioritize works that focus on datasets and modelling methods for cultural adaptation. Additional relevant papers from other sources are also discussed in this work to a lesser extent and coverage. Evaluation method and probing method papers are discussed to a much lesser extent due to the scope of our paper. We also refer readers to recent surveys like Adilazuarda et al. (2024) for focused discussions.

Element (No.)	Papers
Concepts (12)	Shwartz 2022; Cao et al. 2024b; Majewska et al. 2023; Hu et al. 2023 Kabra et al. 2023; Liu et al. 2023b, 2021; Yin et al. 2021 Thapliyal et al. 2022; Khanuja et al. 2024 Liu et al. 2023e; Ventura et al. 2023
Knowledge (15)	Kassner et al. 2021; Yin et al. 2022 Keleg and Magdy 2023; Zhou et al. 2024 Koto et al. 2024a; Li et al. 2023b; Koto et al. 2023; Wang et al. 2023a Son et al. 2024; Kim et al. 2024; Wibowo et al. 2023; Koto et al. 2024b Nguyen et al. 2023; Fung et al. 2024; Nguyen et al. 2024
Values - general (4)	Durmus et al. 2023; Santurkar et al. 2023 Kirk et al. 2024; Zhao et al. 2024
Values - bias (17)	Campolungo et al. 2022; Sandoval et al. 2023; Jin et al. 2024 Mukherjee et al. 2023; España-Bonet and Barrón-Cedeño 2022 Das et al. 2023; Naous et al. 2023; Bhutani et al. 2024 Dev et al. 2023; Zhao et al. 2023; Névéol et al. 2022 Felkner et al. 2023; Lee et al. 2023a; Zhou et al. 2022 Palta and Rudinger 2023; An et al. 2023; Jha et al. 2024
Values - hate (3)	Lwowski et al. 2022; Arango Monnar et al. 2022; Lee et al. 2023b
Values - other perceptions (3)	Rai et al. 2024; Frenda et al. 2023; Mohamed et al. 2022
Norms and Morals (17)	Forbes et al. 2020; Fung et al. 2023; Moghimifar et al. 2023 Shi et al. 2024; Pyatkin et al. 2023; CH-Wang et al. 2023 Ziems et al. 2023a, 2022b; Rai et al. 2024 Dwivedi et al. 2023; Rao et al. 2023; Huang and Yang 2023 Sun et al. 2023; Kim et al. 2022; Zhan et al. 2023 Li et al. 2023c; Zhan et al. 2024
Linguistic Form - dialects (17)	Salameh et al. 2018; Abdelali et al. 2021; Yusuf et al. 2022 Hämäläinen et al. 2021; Le and Luu 2023; Paonessa et al. 2023 Abu Farha and Magdy 2020; Ziems et al. 2022a, 2023b Aepli et al. 2023; Plüss et al. 2022, 2023 Kuparinen et al. 2023; Olabisi et al. 2022; Faisal et al. 2024 Elmadany et al. 2023; Deas et al. 2023
Artifacts (3)	Mohamed et al. 2022; Kruk et al. 2023; Jiang et al. 2023

Table 1: Recent resources considered in §3.1.