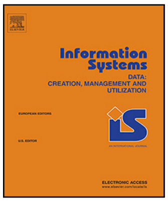
Information Systems 138 (2026) 102649

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is



Exploring cultural commonsense in multilingual large language models: A survey

Geleta Negasa Binegde, Huaping Zhang [*]

School of Computer Science and Technology Beijing Institute of Technology, Beijing, 100081, China

ARTICLE INFO

Keywords:
Multilingual LLMs
Cultural bias
Commonsense knowledge
Cultural understanding

ABSTRACT

Large language models (LLMs) have demonstrated impressive proficiency in multilingual natural language processing (NLP), yet they frequently struggle with cultural commonsense—the implicit knowledge shaped by societal norms, traditions, and shared experiences. As these models are deployed in diverse linguistic and cultural settings, their ability to understand and apply cultural commonsense becomes crucial for ensuring fairness, inclusivity, and contextual accuracy. This paper presents a systematic review and a large-scale empirical benchmark for evaluating cultural commonsense in multilingual LLMs. Through a comprehensive evaluation of 15 models on the BLEnD dataset, our analysis reveals a critical performance gap of 64.2% between high-resource and low-resource cultures. The results demonstrate significant disparities across model architectures: encoder-only models show more consistent but lower overall performance compared to decoder-based models. We identify key limitations, including data scarcity, representational bias, and inadequate cross-lingual knowledge transfer. Finally, we propose future research directions, such as culturally diverse dataset curation, hybrid knowledge graph architectures, and fairness-aware fine-tuning. The primary contributions of this work are (1) a systematic review of challenges and mitigation strategies for cultural commonsense; (2) a large-scale empirical benchmark that evaluates 15 multilingual LLMs across 13 languages and 16 countries, revealing significant performance disparities; and (3) concrete findings on the effects of model architecture and the limitations of scale in cultural understanding. This research underscores the urgent need to advance cultural commonsense in multilingual LLMs to ensure the development of fair, inclusive, and contextually accurate AI systems globally.

Contents

[*] Corresponding author.
E-mail addresses: geleta.negasa@bit.edu.cn (G.N. Binegde), kevinzhang@bit.edu.cn (H. Zhang).

https://doi.org/10.1016/j.is.2025.102649
Received 27 October 2025; Received in revised form 20 November 2025; Accepted 28 November 2025
Available online 1 December 2025

1. Introduction

Language models have demonstrated remarkable progress in natural language understanding and generation; however, their performance across multiple languages and cultures remains uneven due to variations in commonsense knowledge across societies. Cultural commonsense refers to the implicit knowledge shaped by societal norms, traditions, and shared experiences. Ensuring that multilingual large language models (LLMs) understand and incorporate cultural commonsense is crucial for improving cross-cultural communication and mitigating biases. With the rise of artificial intelligence (AI) and natural language processing (NLP) technologies, multilingual LLMs such as GPT, BERT, and XLM-R have become instrumental in global communication [1]. These models, trained on vast multilingual datasets, have the potential to bridge linguistic and cultural divides. They have demonstrated remarkable capabilities in understanding and generating human-like text, leading to widespread adoption across various industries. However, despite their successes, LLMs face significant challenges in dealing with cultural commonsense, especially when applied in multilingual contexts [2,3]. Multilingual LLMs such as GPT-4, BERT, and XLM-R have revolutionized NLP by providing pretrained representations in multiple languages, enabling applications ranging from machine translation to sentiment analysis [4]. Although these models have significantly improved cross-linguistic NLP tasks, they often exhibit biases that reflect cultural assumptions embedded in their training data. Despite their remarkable success, nearly all existing studies, as noted by [5], highlight the challenges of bias in these models. As multilingual LLMs are increasingly deployed in global applications, the importance of cultural sensitivity becomes paramount to ensure that users from diverse backgrounds feel respected and understood. Cultural harm can arise when these models fail to align with specific cultural norms, resulting in misrepresentations or violations of cultural values. On the other hand, commonsense knowledge refers to the background knowledge that humans use to make everyday decisions, interpret ambiguous situations, and understand the world around them [6]. While LLMs excel at processing explicit information, they often struggle with understanding implicit knowledge, which is essential for human-like reasoning. This survey explores cultural commonsense in multilingual LLMs, aiming to ensure that these models not only perform well linguistically but are also culturally aware and capable of reasoning about everyday situations. By reviewing existing methodologies, identifying challenges, and proposing future research directions, this survey seeks to provide valuable insights into improving the fairness, inclusivity, and commonsense capabilities of multilingual models. In terms of the impact that multilingual LLMs have on scientific research, [7] proposes categorizing language-dependent abilities into three distinct categories, which vary in the degree to which language choice affects performance: reasoning (least impact), knowledge access, and articulation (most impact). The selected set of tasks from these three categories is evaluated by assessing the multilingual abilities of an LLM using a novel prompting method called response back-translation. By comparing the generated answers, we can both measure the multilingual performance of the LLM and determine the type of multilinguality it exhibits. Tackling cultural bias requires an LLM to embrace cultural differences [8]. Although LLMs possess extensive knowledge about various cultures, prompt engineering techniques have been devised to induce LLMs to exhibit specific cultural perspectives. The intricate relationship between LLMs and cultural bias is a critical area of study. Currently, many studies analyze and address the issues associated with content generation by multilingual LLMs. Most of these studies focus on diverse cultural commonsense knowledge, primarily revolving around the detection and mitigation of cultural biases, representation issues, language inequalities, and ethical concerns. Many multilingual LLMs are trained on large datasets dominated by data from high-resource languages (such as English and French) and Western perspectives. Despite their impressive performance across various NLP tasks, LLMs face significant limitations in handling cultural commonsense in multilingual settings. Multilingual LLMs are expected to generate contextually accurate and culturally appropriate outputs across many languages, each with its own set of cultural references, historical contexts, and linguistic structures. While LLMs excel at pattern recognition, they often struggle with commonsense knowledge and understanding everyday facts or making logical inferences that humans typically know. These cultural biases arise from imbalanced data, model architectures, and inadequate debiasing mechanisms during training. As these models are deployed in critical applications like education, content moderation, and decision-making systems, addressing and mitigating bias becomes both a moral and technical imperative. While various studies have examined commonsense memorization, numerical commonsense, toxic speech, and other areas vulnerable to undermining the reliability of LLMs' commonsense knowledge capabilities [9], the lack of effective cultural commonsense significantly limits the global applicability of LLMs. In a study by [10], ensuring that models avoid reinforcing harmful stereotypes, respect cultural norms, and accurately represent diverse cultural contexts and commonsense knowledge involves understanding

Table 1
A recent survey on culturally sensitive and commonsense knowledge in multilingual LLMs.

References	Objectives	Limitations
[11]	Enhance language models with social commonsense (ATOMIC) to detect cultural biases and improve explanation generation.	ATOMIC based on Western norms; lacks non-Western coverage.
[12]	Incorporate cultural awareness into natural language inference tasks.	Dataset may not fully represent cultural differences.
[9]	Evaluate commonsense reasoning in sociocultural contexts (memorization, toxicity, proverbs).	Limited to Korean language only.
[13]	Explore cultural biases in data, algorithms, and interactions impacting diversity.	Lacks standardized bias mitigation metrics.
[14]	Assess Korean culture and language via 1995 multiple-choice questions.	Korean-specific; lacks sociolinguistic annotations.
[15]	Analyze cultural trends, language adaptation, and domain-specific LLM orientations.	Open-ended tasks underexplored due to evaluation challenges.
[16]	Evaluate cultural biases in GPT models; suggest cultural prompting to reduce biases.	Focuses on GPT; limited coverage of other transformers.
[17]	Assess cultural sensitivity and adaptiveness of LLMs.	Less suitable for natural cultural adaptability evaluation.
[18]	Investigate cultural disparities in hate speech detection across annotators' contexts.	Small dataset; may limit representativeness.
[19]	Develop AI with robust situational awareness of cultural contexts.	Relies on pre-trained models and heuristics; limited generalization.

and applying the everyday knowledge that humans use to navigate the world. Current research efforts tackle this crucial and timely issue regarding cultural biases in LLMs, providing profound insights and potential pathways for future research that could significantly improve the reliability and fairness of AI systems. The primary contributions of this survey paper are:

- We provide a comprehensive synthesis of the challenges, evaluation methodologies, and mitigation strategies for cultural commonsense in multilingual LLMs, offering a structured taxonomy of cultural biases and their solutions.
- We design and execute a large-scale evaluation of 15 state-of-the-art multilingual LLMs across 13 languages and 16 countries, quantifying a critical performance gap.
- We deliver concrete findings on the ineffectiveness of model scale alone for cultural understanding and identify a significant performance disparity between encoder-only and decoder-based model architectures, providing clear guidance for future model development.

1.1. Objective and limitation of recent surveys

Several recent surveys on cultural commonsense in multilingual LLMs are summarized in Table 1. The majority of these surveys are from 2024, focusing on multicultural knowledge acquisition to help identify cultural biases, enhance commonsense understanding, mitigate societal biases, and address the shortcomings of existing methods in capturing the diverse and rich cultures across the world. Recently, [19] introduced Candle, an end-to-end methodology for extracting high-quality cultural commonsense. It extracts cultural commonsense assertions from a large web corpus and organizes them into coherent clusters across three domains (geography, religion, and occupation) and several cultural facets (food, drinks, clothing, traditions, rituals, and behaviors). The study by [18] introduced CREHate, a cross-cultural English hate speech dataset comprising 1580 online posts annotated by individuals from five English-speaking countries. Based on cross-cultural considerations, the dataset creation procedure ensures that CREHate is more culturally comprehensive than datasets that ignore cultural differences within English-speaking countries. [17] constructed BLEnD, a hand-crafted benchmark designed to evaluate LLMs' everyday knowledge across diverse cultures and languages. It comprises question-answer pairs from different countries and languages, including low-resource ones such as Amharic, Assamese, Azerbaijani, Hausa, and Sundanese. [16] evaluated cultural biases in GPT models and proposed cultural prompting as a strategy to reduce biases in generative AI models. However, their review primarily focuses on generative models like GPT and does not cover other transformer-based architectures in detail. Additionally, it is limited to evaluating models' outputs based on survey data, which might not fully capture the nuances of cultural bias across all applications. In the study by [15], comprehensive experiments were conducted to investigate culture in mainstream LLMs from various perspectives, including the overall cultural trends of LLMs, adaptation to different language contexts, and cultural consistency within model families. [14] developed a benchmark for Korean Cultural and Linguistic Intelligence based on question-answer pairs. The data sources included official Korean exams and textbooks, with questions partitioned into eleven categories under two main categories of language and culture. Another study by [11] identified cultural biases in data, specifically causal assumptions and commonsense implications that strongly influence model decisions across a variety of tasks designed for social impact. [13] analyzed the sources of cultural bias in LLMs, categorized the origins of such biases, and proposed comprehensive strategies for mitigation. [9] introduced KoCommonGEN v2, a fine-grained benchmark dataset focused on Korean commonsense reasoning. This dataset, enriched with human annotations, comprises multiple-choice questions across seven error categories. These categories include commonsense memorization, numerical commonsense, toxic speech, and others, which are vulnerable to undermining the reliability of LLMs' commonsense reasoning capabilities. [12] presented the first culturally aware natural language inference dataset, which surfaces cultural variations by examining label disagreements between annotators from different cultural backgrounds. However, the study is limited to two cultural groups, such as the U.S. and India, which may not generalize well to other cultural contexts. Cultural sensitivity refers to a model's ability to recognize, respect, and adapt to different cultural contexts, while commonsense reasoning involves the ability to reason about everyday knowledge that is widely accepted and implicit in human understanding. To the best of our knowledge, previous surveys have primarily focused on commonsense knowledge that works well for one language or culture but may not account for cultural differences. This survey specifically addresses the need for cultural commonsense and the capability to handle commonsense knowledge across diverse cultural contexts. This paper is organized as follows: Section 2 outlines the review methodology. Sections 3 and 4 provide a systematic exploration of cultural commonsense and commonsense knowledge in multilingual LLMs, discussing key challenges, evaluation methods, and mitigation strategies. Beyond the systematic review, a key contribution of this survey is a novel empirical benchmark (Section 5) evaluating 15 state-of-the-art models across 13 languages. The results

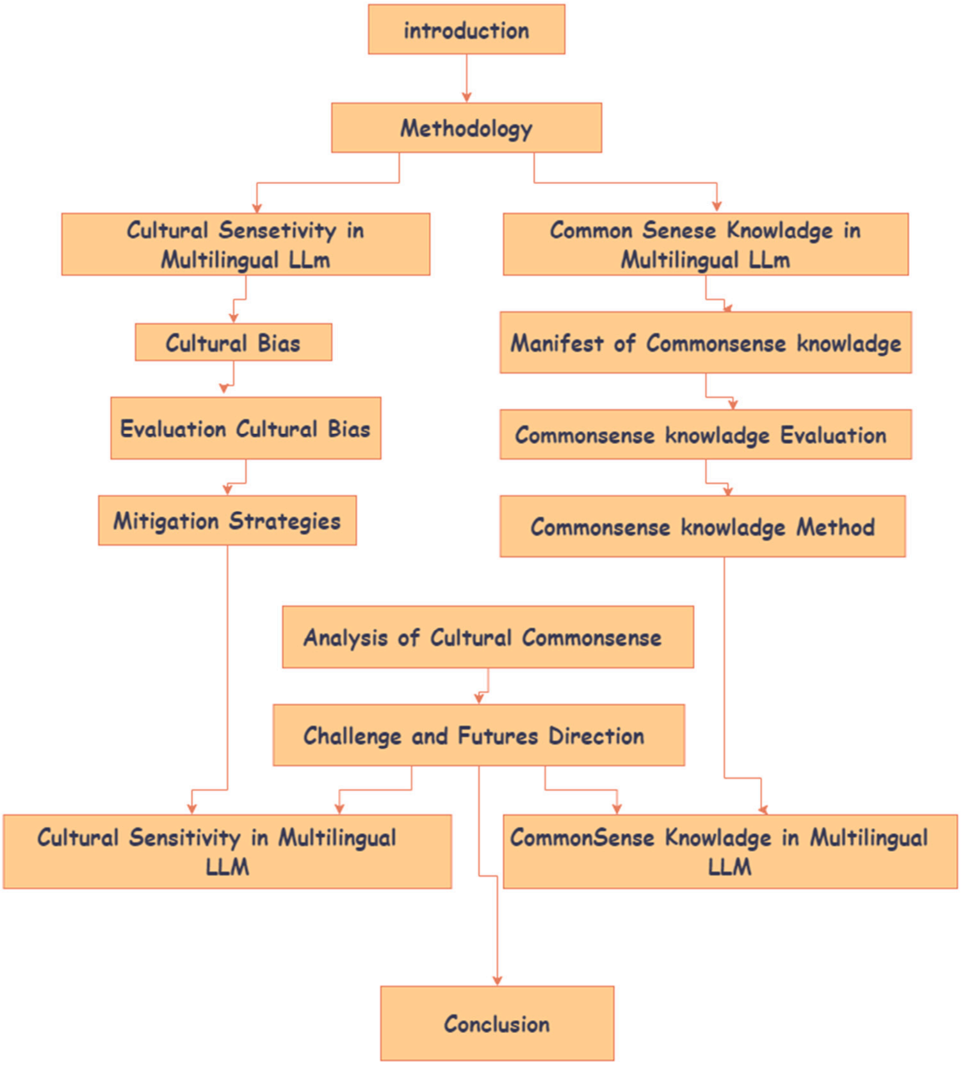

introduction
Methodology
Cultural Sensetivity in Multilingual LLm
Common Senese Knowladge in Multilingual LLm
Cultural Bias
Manifest of Commonsense knowladge
Evaluation Cultural Bias
Commonsense knowladge Evaluation
Mitigation Strategies
Commonsense knowladge Method
Analysis of Cultural Commonsense
Challenge and Futures Direction
Cultural Sensitivity in Multilingual LLM
CommonSense Knowladge in Multilingual LLM
Conclusion

Fig. 1. Flow of section paper survey.

of this benchmark directly inform our analysis of limitations and future directions. Finally, Section 6 summarizes the research trends identified and proposes concrete pathways for future work(see Fig. 1).

2. Methodology

This paper conducts a systematic search across multiple electronic databases to identify relevant articles, including Web of Science,[1] Google Scholar,[2] ACL Anthology,[3] AAAI Digital Library,[4] IEEE Xplore,[5] and Springer Link.[6] Relevant keywords such as "cultural sensitivity", "commonsense knowledge", and "multilingual LLM" were used to identify publications. The selection process involved refining the results by reviewing the titles and abstracts of the papers. Additionally, some of the gathered papers led to the inclusion of additional references through cross-referencing. Fig. 2 illustrates the number of papers published between March 2023 and October 2024 that focus on cultural sensitivity and commonsense knowledge in multilingual LLMs. The figure highlights a significant increase in research on these topics following the release of ChatGPT in November 2022, reflecting the growing academic interest in addressing these challenges. This survey methodology ensures a comprehensive understanding of the subject by incorporating both recent research and key foundational works that have shaped the field. As shown in Fig. 2, the article organizes the main concerns regarding LLM-generated content into separate sections, enabling readers to navigate the content based on their interests.

3. Cultural commonsense in multilingual LLMs

This section begins by introducing the concept of cultural commonsense and discusses its context in Section 3.1. Section 3.2 presents the commonly used datasets and evaluation metrics related to cultural bias. Finally, Section 3.3 explores various strategies and methods for incorporating cultural context.

3.1. Cultural bias

This section begins by addressing cultural bias, provides an overview of related research history, and finally explores a taxonomy of various bias categories. Cultural commonsense can be defined as the awareness of inherent cultural biases in training data and the process by which LLMs can recognize and adapt to cultural norms, social behaviors,

[1] https://www.webofscience.com/wos/alldb/basic-search
[2] https://scholar.google.com
[3] https://aclanthology.org
[4] https://www.proceedings.aaai.org/Library/library.php
[5] https://ieeexplore.ieee.org/Xplore/home.jsp
[6] https://link.springer.com

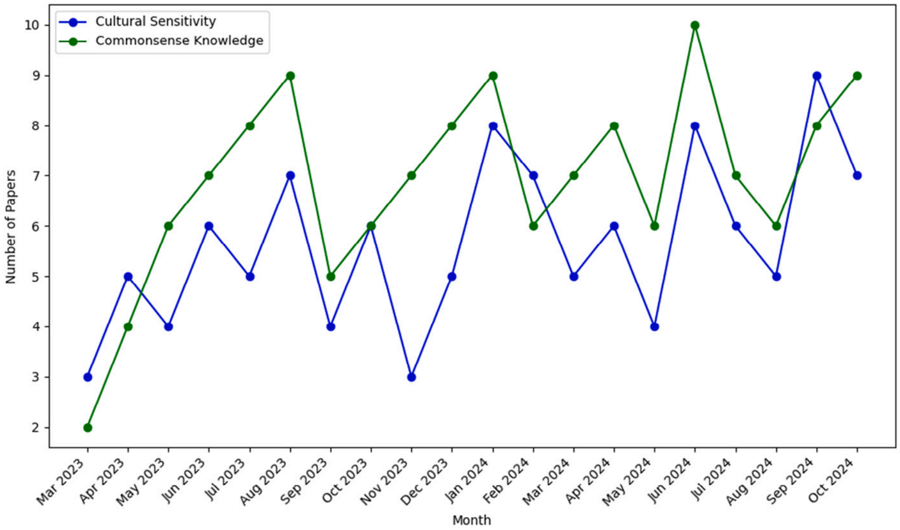


Fig. 2. The number of papers on cultural sensitive and commonsense knowledge between Mar 2023 and Oct 2024.

and linguistic diversity in multilingual settings. Notably, LLMs, particularly pre-trained models, are often trained on large-scale datasets collected from the internet, which are predominantly dominated by Western content [20]. In the context of cultural sensitivity, cultural bias refers to the tendency of a system, model, or individual to favor one culture's values, beliefs, or practices over others. This can lead to misunderstandings, misinterpretations, or unfair treatment of individuals from different cultural backgrounds. Addressing cultural bias in multilingual settings presents both theoretical and practical challenges in AI and NLP. Today, state-of-the-art generative AI systems are used globally, yet their evaluation is primarily conducted using English benchmarks [21]. Initiatives like Global-MMLU [22] have introduced subsets labeled as culturally sensitive and culturally agnostic to enable a more holistic and comprehensive evaluation of these systems. In everyday digital interactions ranging from virtual assistants to online content, AI systems have the potential to influence cultural norms, perceptions, and identities [13,23,24]. The development of culturally sensitive AI is critical for building equitable systems that respect and reflect the diverse cultures of the world [25–28]. Currently, research on cultural commonsense often focuses on specific types of cultural bias, which can be broadly categorized into three primary categories:

- **Stereotypical Bias**: A stereotype is an overgeneralized belief about a particular group of people [29]. For example, beliefs such as "Asians are good at math" or "African Americans are athletic" are common stereotypes. Such biases can harm target groups based on religion, ethnicity, race, gender, or profession.
- **Representational Bias**: This occurs when certain groups, cultures, or languages are underrepresented or overrepresented in training data, leading to skewed outputs. As highlighted by [30], representational bias in data can arise due to various factors, ranging from historical discrimination to selection and sampling biases in data acquisition and preparation methods.
- **Institutional Bias**: This type of bias stems from systemic factors within the data collection or model creation process, often reflecting dominant ideologies or narratives propagated by institutions. According to [31], understanding algorithmic bias as institutional bias (as opposed to other structural accounts) has at least two important implications. First, the existence of bias intrinsic to certain institutions—whether algorithmic or not—suggests that, in some cases, systemic change is necessary. Second, in other cases, modifying the algorithms used within institutions (rather than eliminating them entirely) is essential to addressing the underlying structural conditions of society.

3.1.1. *Intersectional biases*

Intersectional biases amplify when cultural commonsense intersects with other biases, such as gender or age. For example, gender-race biases are more pronounced in LLMs, with Black women stereotyped more negatively [32]. In cultural contexts, Asian men scored lower on "masculinity" metrics [33]. These compounded biases highlight the need for mitigation strategies that address multiple dimensions simultaneously.

3.2. *Evaluation metrics*

In some previous work, cultural bias measures are categorized into stereotype measures and fairness measures [34–36]. Stereotype measures examine stereotypes in LLM-generated content, while fairness measures assess cross-cultural fairness in LLMs. However, we note that this categorization is insufficient to evaluate cultural differences clearly. Seminal work on evaluating cultural bias in multilingual models [37] introduces CAMeL, which provides a foundation for measuring cultural biases in language models through both extrinsic and intrinsic evaluations. Similar studies have explored language models, focusing on understanding cross-cultural differences in norms and values [38] as well as cultural dimensions [39].

3.2.1. *Evaluation benchmarks*

In this section, well-known datasets such as CLIcK [14], StereoSet [29], and Global MMLU [22] are discussed. These datasets are prominent tools for measuring cultural bias in models by evaluating their tendency to prefer stereotypical content over neutral or anti-stereotypical alternatives. Furthermore, the most widely used and newly proposed benchmarks are presented in Table 2. For example, multilingual WEAT tests word embeddings across languages to assess implicit cultural biases in associations between words, providing a baseline for evaluating biases across multilingual LLMs [40]. This section also highlights some representative works in the field.

- **BLEnD** [17] is a benchmark designed to measure the performance of large language models (LLMs) on culturally specific everyday knowledge across different countries and languages, including low-resource languages such as Amharic and Sundanese. It evaluates the cultural adaptability and everyday knowledge of LLMs across diverse regions and languages. The quality of data for low-resource languages can vary significantly, which may impact the reliability of the benchmark in these contexts.

Table 2
Evaluation methods for cultural sensitivity benchmarks in multilingual LLMs.

	Type of cultural bias							
Benchmarks	Culture	Race	Language	Gender	Profession	Religions	Evaluation method	References
CREHate	✓		✓				Human-based	[42]
CAMeL			✓				Generation-based	[43]
CDEval	✓						model-based	[44]
BLEND	✓		✓				Model-based	[17]
StereoSet		✓		✓	✓	✓	MLM-based	[29]
CLIcK	✓		✓					[14]
Global-MMLU	✓		✓				Model-based	[45]
BHASA		✓		✓			Generation-based	[46]
KoBBQ		✓	✓				Generation-based	[47]
DLAMA		✓					MLM-based	[48]
ARADICE	✓						MLM-based	[49]
CultureAtlas	✓						Generation-based	[50]
SeaEval	✓		✓				Model-based	[51]
CARROT	✓						Generation-based	[52]
CulturalBench	✓						Model-based	[52]
CULTURALVQA	✓						Generation-based	[53]
LLM-GLOBE	✓						Embedded-based	[54]
TWBias	✓			✓			Embedded-based	[55]
CamelEval	✓		✓				Generation-based	[56]
M5	✓		✓					[57]
IndicGenBench			✓				Generation-based	[58]
PARIKSHA	✓		✓				Human model-based	[59]
CultureLLM	✓						Generation-based	[60]
CulturePark	✓						Model-based	[61]

- **CDEval** [15] evaluates cultural dimensions in LLMs using Hofstede's cultural dimensions theory. It measures how well LLMs align with diverse cultural orientations by assessing six cultural dimensions across seven domains (e.g., education, family, and arts). While evaluating cultural dimensions across multiple domains, CDEval offers a comprehensive understanding of how LLMs handle cultural biases in diverse contexts. However, this approach restricts the benchmark's applicability to non-Western cultures. Furthermore, its reliance on predefined cultural dimensions may fail to account for context-specific cultural biases that fall outside the scope of Hofstede's framework.

3.2.2. *Cross-lingual evaluation*

Cross-lingual evaluation methods assess how well cultural commonsense transfers across languages, often revealing gaps in low-resource settings. For example, models like GPT-4 show improved performance in Chinese but struggle with Persian or Swahili [41].

3.3. *Mitigation strategies*

Recent work has focused on fine-tuning multilingual LLMs to better handle cultural nuances. We categorize cultural bias mitigation techniques into four stages of the LLM workflow: pre-processing, in-processing, post-processing, and cross-lingual adaptation. Pre-processing techniques aim to learn more balanced associations in the training data, while adversarial debiasing focuses on identifying and eliminating biases within the model. In-processing fairness constraints involve incorporating fairness objectives into the model's loss function during training. Post-processing techniques adjust the model's outputs after training to reduce bias. Finally, cross-lingual bias mitigation techniques require adaptation to account for cultural variations across languages (see Fig. 4 for illustration).

- **Pre-processing**: This section discusses methods to balance training data by introducing anti-stereotypical examples and reducing bias propagation. For instance, if male doctors are overrepresented in the data, additional examples of female doctors are added. These strategies involve generating variations of existing data to counterbalance biases in training. To mitigate gender bias between male and female demographic groups, it is essential to ensure that gender-neutral terms exhibit consistent relationships with gender-specific terms. Consider the sentence, "He is a doctor". By employing the Counterfactual Data Augmentation (CDA) method, the gender-specific term "He" can be replaced with "She", producing an additional training sentence: "She is a doctor" [62]. Pre-processing methods aim to reduce bias in data, which is crucial because, under fixed model parameters, training data has the most significant impact on model performance [28]. Debiasing with CDA addresses bias in event coreference resolution systems caused by over-reliance on lexical matching of event triggers, which often leads to incorrect coreference predictions [63]. CDA removes this bias by introducing variations in the data, forcing the model to focus on deeper causal associations (rationales) rather than superficial lexical similarities. For example, if two event mentions have similar triggers but different arguments (e.g., participants, locations), counterfactual data augmentation alters the trigger to introduce lexical divergence while retaining the coreference status. Some methods employ strategies to analyze and compare the explanations generated for retrained machine learning models with and without fairness-aware strategies [64](see Fig. 3).

The CDA method is currently attracting significant attention, particularly the counterfactual dialog mixing approach, which generates realistic synthetic dialogs via counterfactuals to increase the amount of training data [65]. Many approaches focus on generating counterfactual data to enhance model robustness, reduce biases, and improve generalization across tasks such as classification and recommendation [66, 67]. Additionally, several methods [68,69] explore the integration of CDA with fine-tuning, prompt tuning, and adapter tuning techniques.

- **Adversarial Training**: Adversarial training has gained popularity as an effective strategy to combat cultural biases. In adversarial setups, models are jointly trained with adversaries (e.g., discriminators) whose role is to detect biased features. [70] pioneered the technique of using multiple networks with competing goals, forcing the first network to "deceive" the second network, and applied this method to the problem of generating realistic images. Furthermore, adversarial training has been used to achieve equality of opportunity in cases where the output variable is discrete [71]. Additionally, the adversary must be powerful enough to enforce fairness constraints, even when trained

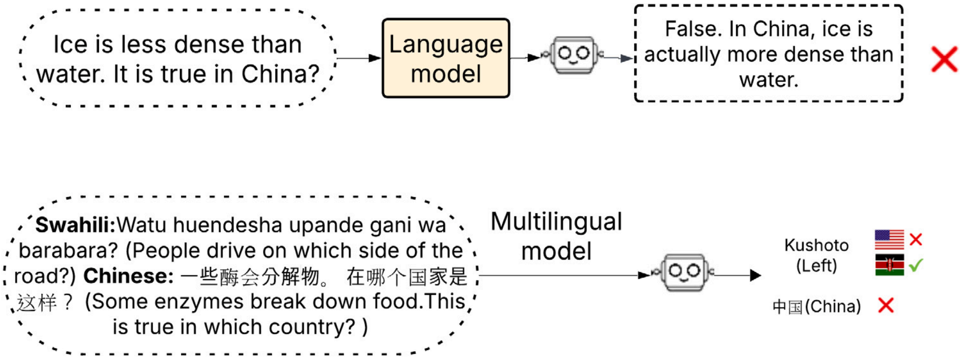

Ice is less dense than water. It is true in China?
Language model
False. In China, ice is actually more dense than water.
✗
Swahili: Watu huendesha upande gani wa barabara? (People drive on which side of the road?) **Chinese:** 一些酶会分解物。 在哪个国家是这样？ (Some enzymes break down food.This is true in which country?)
Multilingual model
Kushoto (Left)
✗
✓
中国(China) ✗

Fig. 3. Examples of cultural commonsense evaluation methods.

on a small dataset. Another approach eliminates the need for additional or task-specific debiasing. [72] proposed a novel debiasing method that employs adversarial learning during model pre-training. Without requiring hyperparameter optimization, this computationally efficient method demonstrates improved fairness in natural language generation tasks while maintaining performance. As a specialized training regimen, adversarial training seeks to minimize the maximal loss incurred by label-preserving adversarial perturbations [73]. It has also been hypothesized that adversarial training is particularly effective in cross-lingual frameworks when used with augmented data. This approach encourages the model to become more robust to divergences in similar words and word orders across languages and better adapt to modestly noisy data [74].

Adversarial training minimizes a worst-case loss function, making the model robust to small perturbations in the input embedding space, which can improve cross-lingual generalization.

- **In-Processing Fairness Constraints**: Adding fairness constraints to the model's loss function during training prevents the model from learning associations based on sensitive attributes. By incorporating fairness during the learning phase, models can avoid embedding bias in the first place, making this an effective strategy for long-term mitigation [75]. Recently, various techniques have been developed to mitigate unfairness in machine learning models. Among these, in-processing methods have gained increasing attention, as they directly incorporate fairness considerations during model design, leading to intrinsically fair models and fundamentally addressing fairness issues in outputs and representations [76]. In-processing methods integrate fairness into the model design process, which can inherently reduce bias and mitigate fairness issues in machine learning models [77–81]. For example, these techniques can address the problem of bias amplification during model training, where models tend to exacerbate biases present in the training data. [82] introduced a metric called counterfactual token fairness to measure counterfactual fairness in text classifiers. They actively optimize counterfactual token fairness during the training phase. Another approach, proposed by [83], involves directly modifying the loss function in text generation tasks. This modification aims to reduce gender bias in language models during training by ensuring an equal probability distribution for male and female words in the model's output. Meanwhile, [84] focuses on debiasing pre-trained contextualized embeddings at the token or sentence level. However, it is important to note that introducing additional loss functions relies on a rigid definition of bias. As a result, the requirements for the loss function are stringent, making this approach somewhat inflexible.
- **Post-Processing Mitigation**: Post-processing techniques involve adjusting the outputs of a model after it has been trained to reduce bias. These techniques focus on modifying model outputs to remove bias, which is particularly useful for pre-trained models that often remain black boxes. Such models provide limited information about their training data, optimization procedures, or internal mechanisms, offering only their outputs. To address this challenge, several studies have proposed post hoc methods that do not alter the original model parameters but instead mitigate bias in the generated outputs. For example, post-processing mitigation can be achieved by identifying biased tokens and replacing them through rewriting [18].

3.3.1. Cross-lingual bias mitigation

Cross-lingual mitigation adapts models to handle cultural variations across languages. For example, attention-based Cross-Lingual Commonsense Knowledge Transfer reduces performance disparities between English and non-English languages in commonsense question-answering tasks [85].

4. Commonsense knowledge

This section begins by providing a comprehensive definition of commonsense knowledge in multilingual LLMs. Next, we explore the underlying mechanisms used to represent commonsense knowledge. Finally, we evaluate the metrics and methods used for assessing commonsense knowledge in multilingual LLMs.

4.1. Manifestation of commonsense knowledge

In general, commonsense knowledge in the context of LLMs is defined as including factual knowledge (e.g., understanding physical laws), everyday knowledge (e.g., knowing how objects interact in space), social norms (e.g., politeness and communication styles), and emotional intelligence (e.g., empathy and understanding human feelings). These categories can overlap, but they also have distinct properties that require different forms of representation and reasoning. To extract high-quality cultural commonsense knowledge at scale, Candle presents assertions from a large web corpus and organizes them into coherent clusters. It covers three domains (geography, religion, and occupation) and several cultural facets (e.g., food, drinks, clothing, traditions, rituals, and behaviors) [86]. The approach leverages pre-trained models with refined techniques to enhance accuracy, utilizing three key inputs: an English text corpus (e.g., a large web crawl), a set of subjects (cultural groups), and a set of cultural facets. Additionally, [87] proposed a commonsense knowledge base completion model that learns structural and contextual representations of commonsense knowledge graph nodes and relations. This is achieved through a relational graph attention network and a pre-trained language model, respectively.

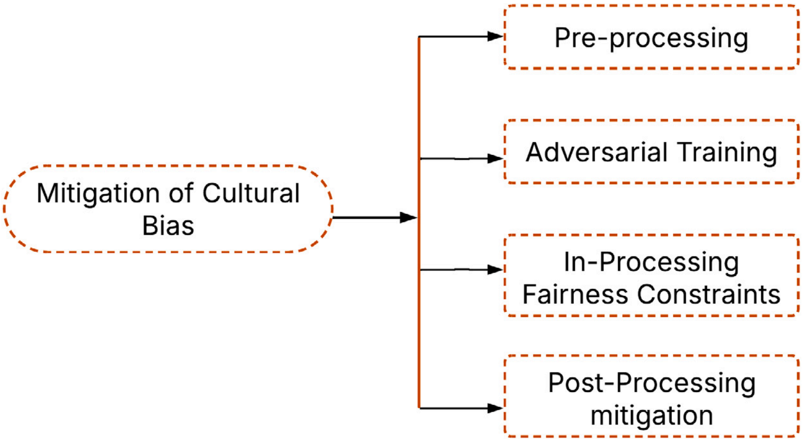

Mitigation of Cultural Bias
Pre-processing
Adversarial Training
In-Processing Fairness Constraints
Post-Processing mitigation

Fig. 4. Illustration of our taxonomy of mitigation strategies for cultural bias.

4.1.1. *Commonsense knowledge representation*

This section focuses on how to effectively represent and use commonsense knowledge across different languages and cultures within large language models. Broadly, these approaches can be categorized into three key techniques. These techniques enable models to reason, infer, and generate contextually appropriate responses based on shared world knowledge.

4.1.2. *Knowledge graphs*

Knowledge graphs are one of the most common ways to represent commonsense knowledge. These graphs capture relationships between concepts, objects, events, and actions in a structured manner [88]. Large-scale commonsense knowledge graphs are ubiquitous tools in natural language processing tasks, as access to their facts enables models to learn to reason over commonsense knowledge and make predictions. Knowledge graphs can be categorized into three types based on their coverage and precision: ConceptNet, ATOMIC, and TRANSOMCS.

- **ConceptNet**: Contains general commonsense facts about the world. This version [89] is a multilingual commonsense knowledge graph. Its nodes are primarily lexical and are connected to each other through 34 relations. Its data is largely derived from the crowdsourced Open Mind Common Sense corpus [90] and is complemented with knowledge from other resources, such as WordNet.
- **ATOMIC**: [91] introduces 23 commonsense relation types. It represents a large-scale repository of textual descriptions that encode both the social and physical aspects of common human everyday experiences. It was designed to complement the commonsense knowledge encoded in current language models.
- **TRANSOMCS**: [92] is a knowledge graph consisting of 18.48 million tuples, which were automatically generated from syntactic parses of sentences obtained from various online sources, such as Wikipedia, Yelp, and Reddit. The relationships used for mapping are sourced from ConceptNet.

4.1.3. *Embedding-based representations*

In general, embedding-based techniques use dense vector representations to encode concepts and relationships in a way that can be easily transferred across languages. [93] presents a novel ensemble model that utilizes both pre-trained word embeddings and knowledge graphs. [94] investigates a dimension reduction technique by training relations jointly with an autoencoder, which is expected to better capture compositional constraints. CrossE [95] is an embedding model that captures the interactions between entities and relations in a knowledge graph, referred to as "crossover interactions". This approach operates by exploring paths within the vector space. The authors argue that these crossover interactions enhance the similarity between entities and relations, resulting in improved explanations compared to conventional embedding models.

4.2. *Commonsense knowledge evaluation*

While surveying previous studies, the issues of commonsense evaluation and commonsense knowledge detection methods are treated as two distinct aspects [96]. Since commonsense is implicit, context-dependent, and general, evaluating it ensures that AI systems can understand, infer, and generate sensible responses in natural language tasks. This process involves assessing how effectively an AI system, particularly a language model, can use, reason with, and apply commonsense knowledge. Evaluating commonsense knowledge in LLMs includes testing the model's ability to: (1) understand common facts and assumptions about the world, (2) make reasonable inferences based on incomplete or implicit information, and (3) resolve ambiguities and contradictions in natural language. Similarly, commonsense knowledge can sometimes be culturally specific or biased. Previously, the understanding of cultural commonsense remained largely unexamined. Research has primarily focused on cultural commonsense evaluation, such as identifying inherent biases in the cultural understanding of LLMs [41,96], generative question answering (QA) [97,98], translation [99–101], and knowledge extraction from text [102]. A context-rich evaluation protocol has been designed specifically for evaluating machine commonsense. This protocol encompasses popular evaluation paradigms in machine commonsense as special cases and is suited for evaluating both discriminative and generative large language models [103]. In an extrinsic evaluation for intercultural dialogs, researchers explore augmenting dialog systems with cultural knowledge assertions [104]. This addresses the difficulty these systems face in processing the intricacies of social and cultural conventions. The following sections begin with a review of recent evaluation metrics, focusing particularly on widely used non-task-specific LLM generation models. They then proceed to discuss the taxonomies of evaluation methods and existing benchmarks (see Fig. 5).

4.2.1. *Evaluation metrics*

Previous studies on specific tasks typically use traditional metrics like BLEU [105], ROUGE [106], and METEOR [107] to evaluate the quality of generated content. However, while these metrics are still useful, they are limited in terms of reasoning quality, flexibility, and cultural sensitivity [108]. As a result, researchers have shifted their focus to evaluating reasoning depth, cultural commonsense, explainability, and multistep reasoning. These newer approaches reflect the increased complexity of modern LLMs, which are expected to handle nuanced commonsense reasoning, provide explanations, and operate effectively across multiple languages and cultures.

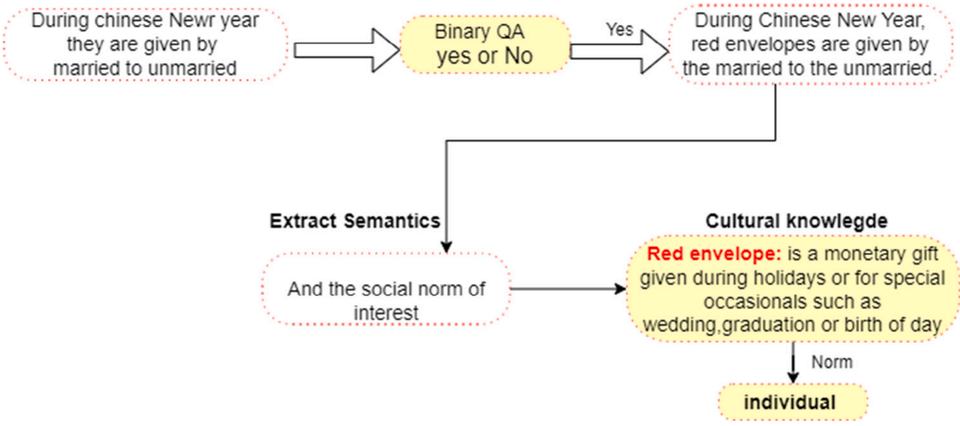

During chinese Newr year they are given by married to unmarried
Binary QA yes or No
Yes
During Chinese New Year, red envelopes are given by the married to the unmarried.
Extract Semantics
Cultural knowlegde
And the social norm of interest
Red envelope: is a monetary gift given during holidays or for special occasionals such as wedding,graduation or birth of day
Norm
individual

Fig. 5. Commonsense Knowledge construction Evaluation Method Process.

4.2.2. *Evaluation benchmarks*

In this section, we introduce widely used and newly proposed benchmarks to evaluate commonsense knowledge in multilingual LLMs, as indicated in Table 3. Evaluating commonsense knowledge has been widely regarded as a crucial task in natural language understanding systems. Commonsense knowledge benchmarks have been created for several geographical regions and countries, including Indonesia [109, 110], China [111], Korea [14,112], Taiwan [113], India [114], and Arabic regions [115]. The purpose of each benchmark is to capture the unique cultural knowledge of the target regions. Indonesia's diverse local cultures and languages have led to the creation of several benchmarks. For instance, IndoCulture [109] is designed to evaluate cultural knowledge across various cultures present in the eleven Indonesian provinces. The dataset was created by asking local people to manually develop cultural contexts and plausible options across a range of predetermined topics. Similarly, [110] develops a cultural diagnostics dataset to evaluate basic cultural knowledge in Indonesian and Tamil languages. They categorized cultural knowledge into language, literature, history, and customs. In China, FoundaBench [111] is designed to rigorously evaluate the fundamental knowledge of LLMs tailored to Chinese language and culture. While some questions in FoundaBench evaluate K-12 commonsense knowledge, others assess K-12 academic knowledge. Similar to the academic knowledge section of the dataset, questions are collected from the internet and automatically generated using GPT-4 [116]. In Korea, HAE-RAE Bench [117] is created to capture culture-specific nuances in the Korean language. It encompasses six downstream tasks across vocabulary, history, general knowledge, and reading comprehension. [14] introduces a dataset focused on linguistic and cultural knowledge in Korea. Cultural commonsense knowledge is categorized into language and culture. Additionally, questions are generated using GPT-4 [116] based on textbook content and selected from standardized Korean exams. [113] developed the Taiwanese Hakka Culture dataset, which is primarily sourced from the Hakka Culture Encyclopedia and the Taiwan Ministry of Education's Hakka Knowledge Base. The questions are designed to include cultural contexts such as history, architecture, customs, and the Hakka language. [114] created DOSA to study local cultural identities in India, based on the country's geographical states. This community-generated dataset comprises 19 distinct Indian geographic subcultures and includes 615 social artifacts. To gather significant subcultural social artifacts, researchers first conducted a survey. Subsequently, they proposed a pipeline to obtain further annotations from local residents for each artifact. [115] introduces the AraDiCE-Culture benchmark, a fine-grained dataset designed to evaluate the cultural awareness of Arabic people in the Gulf, Egypt, and the Levant. The questions address culturally significant topics such as history, geography, cuisine, and public holidays. In the European context, [118] developed RoCulturaBench to assess how effectively LLMs are grounded in Romania's historical, cultural, and social realities. A team of Romanian humanities scholars manually crafted questions covering two subtopics. The first focuses on factual data about Romania, such as its history, geography, and population. The second section addresses topics like traditions, rituals, beliefs, and stereotypes, reflecting how Romanians perceive the world and themselves. Several countries in Asia, including China, Korea, Indonesia, and others, are actively developing culturally relevant benchmarks to evaluate commonsense knowledge. In particular, benchmarks for Indonesia aim to reflect local characteristics by capturing the country's diverse cultures and languages [109,110]. Similarly, it is essential to incorporate local cultures when creating culture-specific benchmarks, rather than treating an entire country as a single homogeneous entity. This approach is especially critical in ethnographically diverse countries, where meticulous attention is required to accurately represent cultural variations.

4.3. *Mitigation strategies for commonsense knowledge*

In this section, we present methods designed to mitigate biases, improve the model's accuracy, and enhance cultural awareness. We classify these mitigation strategies into four categories: Cross-Lingual Knowledge Transfer for models, Commonsense Knowledge Graphs, Fine-Tuning with Diverse Data, and Data Augmentation for data. Each category will be discussed in Sections 4.3.1, 4.3.2, 4.3.3, and 4.3.4, respectively.

4.3.1. *Cross-lingual knowledge transfer*

Cross-Lingual Knowledge Transfer is a promising approach that leverages transfer learning, enabling models to apply knowledge acquired from high-resource languages to low-resource languages. However, several models struggle with deeper cross-lingual knowledge transfer, both in general contexts (e.g., the MMLU benchmark) and domain-specific scenarios [124]. For humans, being multilingual naturally facilitates an understanding of how words and phrases with the same meaning function across different languages (cross-lingual). Building on this premise, our proposed solution is attention-based Cross-Lingual Commonsense Knowledge Transfer, which reduces performance disparities between English and non-English languages in commonsense question-answering tasks [85]. Demonstrations on public benchmarks show that this approach achieves significant improvements in cross-lingual commonsense reasoning tasks for languages other than English. Additionally, [125] introduced methods such as commonsense text infilling and commonsense relation prediction to enhance the commonsense reasoning abilities of pre-trained language models. Furthermore, [126] introduced 29 datasets, including 7 new ones designed for cultural reasoning and cross-lingual consistency, to address gaps in existing benchmarks. These datasets help identify how models handle semantically equivalent queries in different languages, revealing the lack of robust cross-lingual alignment.

Table 3
Evaluation methods in commonsense knowledge benchmarks.

Benchmarks	Descriptions	Tasks	Evaluation methods	References
COMMONSENSEQA	A new QA dataset to capture commonsense beyond associations, extracted from ConceptNet.	Generation	Multiple-choice QA	[119]
CultureAtlas	a dataset, which spans a broad range of geographical regions and ethnolinguistic groupings at the sub-country level.	Human model based	Binary QA	[120]
Global-Liar	Global-Liar is dataset demonstrates innovation in tackling geographic and temporal biases in LLMs, which is crucial for balanced AI evaluation.	model-based	Binary QA	[44]
CAMeL	CAMeL, a resource with 628 naturally-occurring prompts and over 20,000 entities. Its focus on eight entity types for cross-cultural evaluation offers a structured and comprehensive framework for identifying cultural biases	Generation	Long Form Generation	[37]
OMGEval	Emphasizing that each of the 804 open-ended questions is rigorously verified by human annotators adds credibility to the dataset and ensures quality control such as general knowledge and logical reasoning.	Human model based	Long Form Generation	[121]
CULTURE-GEN	a dataset of generations on different culture-related topics on many countries and regions, using gpt-4, llama2-13b, mistral-7b.	model based	Long Form Generation	[122]
MultiNativQA	A multilingual QA dataset with 6̃4k manually annotated pairs across seven languages,culturally and regionally aligned QA datasets in native languages for LLM evaluation and tuning.	model based	Short term Generation	[45]
GD-COMET	a comprehensive human evaluation across 5 diverse cultures, as well as extrinsic evaluation on a geo-diverse tasks.	Human model based	Short term Generation	[123]
DLAMA	curating culturally diverse facts for probing the factual knowledge of PLMs, and building 3 sets of facts from pairs of contrasting cultures representing the culture.	Human model based	Mask Filling	[48]
FoundaBench	a pioneering benchmark designed to rigorously evaluate the fundamental knowledge capabilities of Chinese LLMs.	model based	Multiple-choice QA	[48]

4.3.2. *Commonsense knowledge graph*

The commonsense knowledge graph structures common sense knowledge from resources such as ConceptNet and ATOMIC [127]. This is particularly useful in tasks involving social interactions, where understanding implicit human behavior and social norms is crucial. However, the study by [128] highlights the need to structure common sense knowledge to provide high-quality, simple test cases for tasks like Natural Language Inference (NLI). Unfortunately, commonsense knowledge graphs often contain triples that are not universally accepted. Since the validity of these triples is subjective, they are not ideal for generating test cases and are therefore filtered out. The study presents several templates that define the locations of the triple's head and tail in the text, transforming triples into appropriately formatted inputs for every relation and task. Modern LLMs contain a significant amount of world knowledge, enabling strong performance in commonsense reasoning and knowledge-intensive tasks when utilized effectively. While several works propose mitigation strategies for LLM safety, it remains uncertain how well these strategies address social biases. To tackle this, [129] proposes an approach to exploit social bias in LLMs using a structured BiasKG, transforming free-form stereotypes from the Social Bias Inference Corpus into a structured knowledge graph. Additionally, [130] explores the use of language-specific adapters to integrate external knowledge from ConceptNet and Wikipedia into multilingual LLMs, enhancing their performance on low-resource languages.

4.3.3. *Fine-tuning with diverse data*

Fine-tuning is an effective approach for improving commonsense knowledge across multiple languages. [131] proposes a simple method that diversifies the outputs of LLMs while preserving their quality. Numerous widely used commonsense benchmarks exist, each focusing on a different aspect. For instance, mCSQA [132] evaluates the cross-lingual transfer capabilities of multilingual language models, emphasizing language-specific commonsense and knowledge. These benchmarks provide high-quality, simple test cases for specific tasks. [133] explores a novel application of LLMs for generating consensus statements that align with diverse human opinions, utilizing techniques such as fine-tuning, supervised learning, reward modeling, and social welfare functions. Deep learning has extensively leveraged pre-training to enhance model performance, especially in scenarios with limited task-specific training data. [134] investigates how the diversity and properties of pre-training data influence the robustness of fine-tuned models, particularly under real-world distribution shifts. One of the primary challenges in machine learning is acquiring and utilizing commonsense knowledge. Previous research has shown that supervised algorithms for commonsense knowledge mining often perform poorly on novel data due to the sparsity of training data. To address this, [135] introduces a method for generating commonsense knowledge using a large, pre-trained bidirectional language model. This model ranks the validity of a triple based on the estimated point wise mutual information between the two entities by converting relational triples into masked sentences. The coverage of cultural contexts in multilingual LLMs includes some attention to low-resource languages from regions like Africa, South America, and Asia. Despite the growing interest in multilingual LLMs, most research and datasets remain centered on high-resource languages, particularly those from Western countries. This imbalance not only skews model performance but also limits the ability of LLMs to effectively represent cultural commonsense from low resource regions such as Africa, South America, and parts of Asia. Low resource languages often suffer from a lack of publicly available datasets, significantly hindering the ability of LLMs to learn their cultural norms and implicit knowledge. Since LLMs are primarily trained on internet-based corpora, the limited online presence of low-resource languages results in models that inherit biases from dominant cultures. This leads to misinterpretations, stereotyping, or the complete erasure of certain cultural perspectives. For example, many African languages encode communal values that differ significantly from the individualistic norms of Western cultures, yet most LLMs fail to reflect these distinctions. While some studies have made initial progress in incorporating low-resource languages, more extensive and culturally diverse datasets are needed, particularly for African and Latin American languages. For instance, the BLEnD benchmark [17] is constructed from datasets directly collected from native speakers rather than relying on translated corpora. It includes low-resource languages like Amharic, Assamese, and Sundanese, highlighting the difficulties in ensuring cultural sensitivity and accuracy. Another study by [136] evaluates large multimodal models across 100 languages including many low-resource languages from Africa, South America, and Asia. The benchmark covers 73 countries and 24 scripts, with a focus on cultural aspects such as heritage, customs, architecture, and traditions. The diverse local cultures and languages of Africa, South America, and Asia have led to the creation of several benchmarks. For example, CulturalVQA [137] is designed to evaluate vision-language models on their ability to understand culturally diverse visual and textual data. The dataset includes images and questions from diverse cultural contexts, such as African tribal ceremonies, South American festivals, and Asian religious practices. It highlights the challenges faced by models in understanding culturally specific visual cues. Similarly, [30] explores visually grounded reasoning across languages and cultures, particularly for low-resource languages. It includes images and questions from diverse cultural contexts, including Africa and South America. Additionally, GeoMLAMA [138] designs a benchmark for evaluating the geo-diverse commonsense knowledge of multilingual pre-trained language models. It focuses on 16 geo-diverse concepts (e.g., traffic rules, date formats) across five countries, including low resource regions.

4.3.4. Data augmentation

Data augmentation involves generating synthetic data to enrich training sets, particularly for low-resource languages. For example, counterfactual data augmentation (CDA) creates variations to counter biases [62].

4.4. Multimodal cultural commonsense

Cultural commonsense is not solely textual; it is often conveyed through images, gestures, and visual context. Multimodal LLMs (MLLMs) that process both text and images must therefore be evaluated for cultural biases in visual understanding. Multimodal benchmarks like CulturalVQA [139] evaluate vision-language models' ability to understand culturally diverse visual and textual data across diverse cultural contexts, and M5 [10] assesses multimodal reasoning and cultural understanding across multiple modalities.

5. Empirical benchmark evaluation

To complement our systematic review with concrete, reproducible evidence, we conducted a large-scale empirical evaluation of cultural commonsense capabilities in multilingual LLMs. This benchmark moves beyond theoretical discussion to provide a standardized assessment of how current models handle culturally grounded everyday knowledge across diverse linguistic and geographic contexts.

5.1. Benchmark design

We evaluated 15 multilingual models spanning diverse architectures—including closed-source, open-source, encoder-only, decoder-only, and region-specific models using the BLEnD dataset [17]. We selected this dataset because it is specifically designed for evaluating cultural commonsense with native annotations across diverse languages. Although we considered creating a custom benchmark, we found BLEnD to be sufficiently comprehensive and validated in prior work. This hand-crafted benchmark comprises 52,557 question-answer pairs across 13 languages and 16 countries, with native annotations ensuring cultural authenticity. It focuses on evaluating cultural adaptability and everyday knowledge across categories including food, clothing, holidays, rituals, and social norms, with particular coverage of low-resource languages such as Amharic (Ethiopia) and Sudanese Arabic. Our experimental setup employed standardized evaluation protocols to ensure reproducibility. Testing was conducted on a Linux-based system (Ubuntu 20.04 LTS) with CUDA 12.1 and cuDNN 8.6, utilizing 8 NVIDIA A100 GPUs (80 GB memory each). We maintained consistent parameters across all models: temperature of 0 (deterministic outputs), maximum tokens configured to match BLEnD response requirements, identical prompt structures, and uniform batch processing. For API-based models, we controlled for software-level consistency through standardized prompt engineering and output processing protocols.

5.2. Results and analysis

We present performance analysis across six representative countries from the BLEnD dataset: the United States (US) and Spain (ES) representing Western cultures; China (CN) and Korea (KR) for Eastern cultures; and Ethiopia (ET) and Sudan (SD) for low-resource African cultures. All performance metrics report 95% confidence intervals derived from three independent runs with different random seeds. Standard deviations are visualized in Fig. 6(b) with error bars representing ±1 standard deviation. Pairwise significance testing between major model families using paired t-tests with Bonferroni correction ($\alpha = 0.05$) revealed statistically significant differences: (1) between encoder-only and decoder-based architectures (t = 8.42, $p < 0.001$), and (2) between models trained on high-resource versus low-resource languages when evaluated on low-resource cultures (t = 12.31, $p < 0.001$). These results confirm that observed performance gaps reflect systematic differences in model design and training data rather than random variation.

The Ineffectiveness of Scale Alone: Larger parameter counts did not reliably correlate with improved cultural commonsense. While GPT-4 demonstrated robust performance across high-resource cultures, its accuracy substantially decreased for Ethiopia (25.85%) and Sudan (19.42%). Notably, the 176B-parameter BLOOM model was consistently outperformed by much smaller architectures like the 7.5B-parameter XGLM, indicating that model architecture and training data diversity are more critical than scale alone for cultural understanding.

Encoder vs. Decoder Disparity: Encoder-only models (mBERT, XLM-R) exhibited more consistent but overall lower performance across all cultures, with a smaller maximum performance gap of 58.00% for mBERT. This pattern suggests that while generative decoder models achieve higher peak performance in favorable conditions, they also demonstrate greater cultural bias and instability across diverse contexts. This empirical benchmark provides concrete, quantifiable evidence for the theoretical challenges discussed throughout this survey. The significant performance disparities – particularly the severe underperformance on low-resource cultures – demonstrate that without fundamental shifts toward more equitable data curation and culturally aware model design, the global deployment of LLMs will continue to perpetuate and potentially amplify existing cultural biases. These findings underscore the urgent need for the mitigation strategies and future directions outlined in subsequent sections.

5.3. Distinct contributions

Cultural awareness plays a crucial role in ensuring that LLMs can handle the diversity of cultural contexts across the globe, particularly in maintaining ethical correctness and avoiding harm in multilingual environments. It helps models align with local norms and expectations, while commonsense knowledge ensures that LLMs can reason effectively in ways grounded in universal human experiences. Commonsense knowledge is essential for making everyday inferences shared across cultures, enabling practical and accurate decision-making. Although cultural and commonsense knowledge share the broad objective of improving model understanding and output, they often diverge in their application. For instance, some universal commonsense rules may conflict with cultural norms, such as differing perceptions of privacy or personal space. Understanding when cultural awareness should override commonsense reasoning or vice versa is crucial for the effective deployment of multilingual LLMs. Communication is often challenging due to language and cultural differences. In their study, [140] introduced GlobalMind, a framework that provides commonsense databases for various countries and languages, along with two inference modules to analyze and compute cultural differences and similarities from these databases. As NLP systems reach users from diverse cultural backgrounds, achieving cultural awareness in language understanding is becoming increasingly important. However, current research has

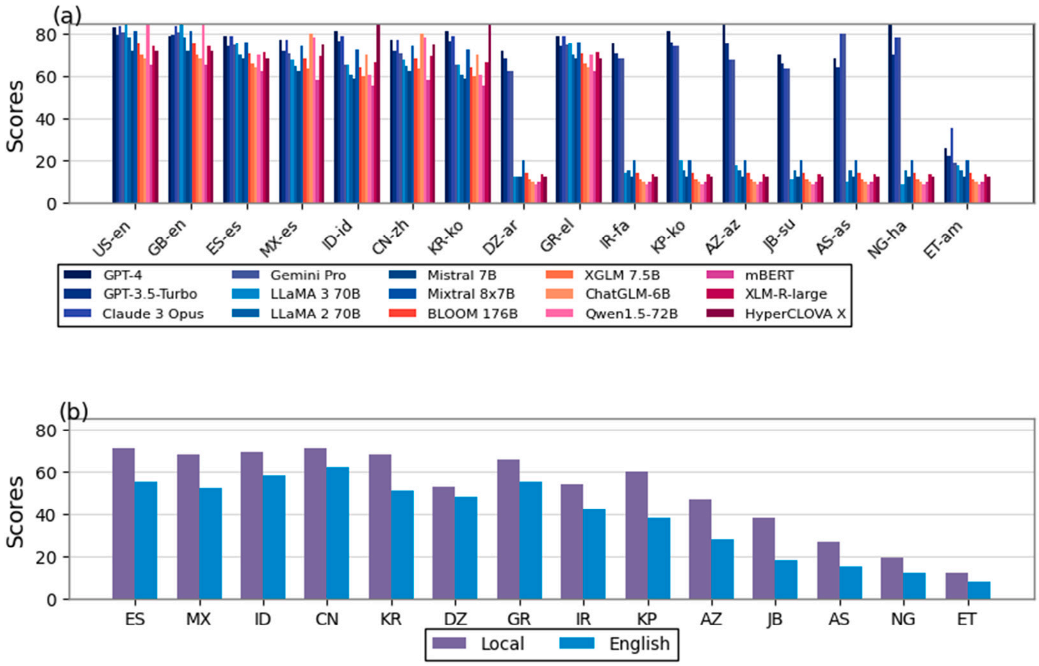


Fig. 6. (a)Performance of multilingual LLMs on short-answer questions for each country or region in its local language. Models originating from Western countries are shown in shades of blue, while those from non-Western countries are shown in shades of red. (b) Average performance of all multilingual LLMs on short-answer questions in both local languages and English. The gray error bars indicate standard deviations across all models.

primarily focused on developing cultural knowledge bases without thoroughly examining how such knowledge influences contextualized text interpretations. In their study, [12] operationalized cultural differences in language understanding through a natural language inference task, revealing cultural differences as label disagreements among annotators from various cultural groups. This paper has significant practical implications for improving AI applications in real-world settings, particularly in content moderation, education, and customer service. By incorporating cultural commonsense into foundational models, AI systems can achieve greater cultural sensitivity and fairness, leading to more inclusive and effective outcomes across diverse applications [19]. For example, the integration of culturally diverse datasets enables content moderation systems to better understand how hate speech and harmful content manifest differently across cultures [17]. This reduces false positives (e.g., incorrectly flagging culturally specific expressions as harmful) and false negatives (e.g., failing to detect harmful content rooted in cultural contexts). Culturally aware models can more accurately identify and mitigate biased or harmful content, ensuring that digital platforms are inclusive and respectful of diverse cultural norms [42]. By employing debiasing techniques and diverse training data, AI systems can minimize the risk of reinforcing stereotypes or misrepresenting minority groups, leading to fairer and more ethical moderation practices [24]. In education, culturally aligned LLMs can inform the development of AI-driven tools that are sensitive to students' cultural backgrounds. For instance, personalized learning experiences can incorporate culturally relevant examples and contexts, improving student engagement and learning outcomes. Such tools can also help educators address cultural biases in curricula, fostering a more inclusive learning environment. Similarly, in customer service, AI chatbots and virtual assistants can be fine-tuned to understand and respond to cultural nuances. Culturally sensitive models can adapt their communication style to align with users' cultural expectations, enhancing user satisfaction and trust. By leveraging cross-lingual knowledge transfer and culturally diverse training data, these applications can provide more accurate and empathetic interactions, fostering better communication across diverse customer bases.

5.4. Mitigation methods

Following the discussions in Sections 3.3 and 4.3, approaches to addressing cultural commonsense are represented by data augmentation [139], which incorporates representations from diverse cultural and linguistic backgrounds. This involves curating corpora that cover a broad range of cultures and dialects, enabling models to handle cultural and linguistic diversity in a more balanced way. Similarly, commonsense knowledge mitigation methods tackle the transfer of knowledge between languages to enhance reasoning capabilities in multilingual settings. For instance, ConceptNet enables models to reason more effectively across different languages. [141] highlights that, although related, these approaches focus on distinct aspects. Cultural sensitivity emphasizes ensuring fairness, inclusivity, and respect for cultural diversity through methods such as bias detection, culturally diverse datasets, and ethical guidelines. On the other hand, commonsense knowledge aims to enhance a model's reasoning abilities by injecting knowledge, using transfer learning, and employing reasoning benchmarks to ensure effective application of commonsense reasoning across multiple languages. While both areas benefit from techniques like data augmentation and human feedback, their strategies differ in their ultimate goals: cultural sensitivity focuses on reducing harmful stereotypes and ensuring fairness, while commonsense knowledge aims to improve reasoning accuracy across diverse linguistic contexts.

5.5. Case study

To ground our theoretical discussion in a specific sociocultural context, we present a case study on African-American Culture (AAC). This case illustrates the limitations of current benchmarks and the perils of treating cultural varieties without their sociohistorical context. Work analyzing cultural commonsense in the context of AAC has demonstrated that understanding cultural context, commonsense reasoning, and the reinforcement of stereotypes often perform poorly when applied to the nuanced nature of cultural commonsense associated with AAC [142,143], and that diverse cultural contexts associated with AAC are addressed as more offensive than culture without them [17,42].

These papers have been critical for highlighting AAC as a diverse cultural context for which existing NLP systems may not work, illustrating their limitations. However, they do not all address cultural bias in the same way. The first four of these papers focus on system performance differences between representing diverse cultural contexts associated with AAC and culture without these contexts. In contrast, the last two papers not only examine these performance differences but motivate this focus with the following additional reasoning: if diverse cultural contexts associated with AAC are scored as more offensive than culture without these contexts, this could (1) yield negative perceptions of AAC, (2) result in the erasure or misrepresentation of AAC contributions in digital spaces, and (3) cause AAC users to feel excluded or marginalized on online platforms. More importantly, none of these studies engage with the broader literature on AAC, cultural hierarchies, or cultural-linguistic ideologies, which are essential for understanding how large language models handle cultural and linguistic diversity, particularly in the context of systemic anti-Black racism. By treating AAC as merely another linguistic variety rather than situating it within its sociohistorical and cultural context, these papers miss an opportunity to address how LLMs perpetuate cultural biases and fail to accurately represent marginalized communities like African Americans. Engaging with the broader literature on cultural commonsense in systems offers a cultural basis for normative analysis. Researchers and practitioners should be concerned about cultural bias in the representation of underrepresented cultures and low-resource languages not only because performance differences in LLMs trained on predominantly Western cultural datasets struggle to generate culturally appropriate responses when prompted with questions about AAC, but also because these differences highlight the need for more inclusive training data that incorporates AAC knowledge. From our perspective, which understands cultural hierarchies and cultural-linguistic ideologies as structural conditions that govern the development and deployment of technology, techniques for measuring or mitigating cultural commonsense in NLP systems will necessarily be incomplete unless they interrogate and dismantle these structural conditions, including the power relations between technologists and culturalized communities. Finally, researchers and practitioners aiming to develop more equitable and effective systems can draw on a growing body of work focused on anti-stereotype cultural diversity. This emerging field challenges the deficit perspective often applied to AAC and other cultural practices [14,22]. This case underscores that mitigating cultural bias requires more than technical fixes; it necessitates a deep engagement with the structural and historical dimensions of culture, especially for marginalized communities.

6. Challenges and future directions

This survey, complemented by our empirical benchmark, reveals significant challenges in achieving cultural commonsense in multilingual LLMs while highlighting promising research directions.

6.1. Cultural commonsense in multilingual LLMs

Current approaches to cultural commonsense face several fundamental challenges that require innovative solutions:

- **Culturally Equitable Data Curation**: The predominant reliance on internet-scraped corpora introduces inherent cultural biases, as these sources disproportionately represent Western perspectives and high-resource languages. Future efforts should prioritize the development of balanced, culturally diverse datasets that authentically represent global traditions and viewpoints. This necessitates creating automated tools for systematic bias detection and mitigation that can evaluate cultural appropriateness at multiple levels from embeddings and probability distributions to generated text outputs [144].
- **Benchmark-Driven Development**: Our findings demonstrate that model development must be guided by rigorous, culturally nuanced benchmarks. While studies like [145] have explored LLM capabilities in low-resource settings through various learning paradigms, there remains an urgent need for standardized evaluation protocols. Future work should focus on developing effective data augmentation strategies, including synthetic data generation, to enhance cultural sensitivity in low-resource languages. Additionally, advancing cross-lingual transfer learning techniques will enable more efficient adaptation of high-resource models to diverse cultural contexts without extensive retraining.

6.2. Commonsense knowledge in multilingual LLMs

The integration of commonsense knowledge presents distinct but related challenges:

- **Multilingual Knowledge Graphs**: Current commonsense knowledge systems often depend on language-specific resources that fail to capture cross-cultural variations. While embedding-based methods leverage knowledge graph structures [146] and text-based approaches utilize encoder-only models [147], both struggle with semantic and contextual nuances across languages. Future research should develop improved cross-lingual transfer techniques that enable effective adaptation of reasoning models from high-resource to low-resource languages, ensuring comprehensive coverage of cultural commonsense.
- **Multilingual Commonsense Benchmarks**: The evaluation landscape suffers from insufficient comprehensive benchmarks that account for cultural and linguistic diversity. Existing datasets like GeoMLAMA and adapted CAMeL [148] represent initial steps but lack the granularity needed for thorough assessment. There is a critical need for benchmarks that evaluate reasoning capabilities across multiple languages and cultural contexts. Incorporating human evaluation from diverse cultural backgrounds will be essential for validating the contextual accuracy of commonsense reasoning in multilingual models.

6.3. Emerging research priorities

Based on our analysis, we identify three key priorities for future research:

Integrated Evaluation Frameworks: Developing unified benchmarks that simultaneously assess cultural sensitivity and commonsense reasoning across linguistic and cultural dimensions.

Architectural Innovation: Designing model architectures that explicitly incorporate cultural reasoning capabilities, potentially through specialized modules or cross-cultural attention mechanisms.

Community Driven Resource Development: Establishing collaborative frameworks for native speakers and cultural experts to contribute to dataset creation and evaluation, particularly for low-resource languages and marginalized cultural contexts.

Addressing these challenges requires coordinated efforts across academia, industry, and cultural communities to develop multilingual LLMs that are truly culturally competent and commonsense-aware.

7. Conclusions

This survey has systematically investigated the critical challenge of cultural commonsense in multilingual large language models (LLMs), combining a comprehensive review with a large-scale empirical benchmark. Our findings underscore that achieving robust cultural commonsense is not a byproduct of scaling but requires sustained, coordinated efforts across data curation, model architecture, evaluation methodologies, and deployment frameworks. Our empirical analysis, evaluating 15 models across 13 languages and 16 countries, reveals two primary limitations. First, it quantifies the severe impact of persistent cultural biases, exposing a performance gap of 64.2% between high-resource and low-resource cultures, a direct threat to the fairness and efficacy of global AI systems. Second, it demonstrates significant limitations in cross-lingual commonsense reasoning, a fundamental barrier to creating inclusive and locally effective digital services. Furthermore, our benchmark confirms that current evaluation standards are often inadequate, failing to capture nuanced cultural contexts and highlighting an urgent need for more culturally sensitive metrics. Moving forward, this research dictates clear priorities: the development of comprehensive, culturally aware benchmarks, the advancement of cross-lingual transfer learning techniques, and the establishment of robust frameworks for ethical AI deployment. To ensure multilingual LLMs are both safe and effective, the research community must address these challenges through coordinated, interdisciplinary efforts that actively prioritize cultural representation and linguistic diversity. Only then can we build AI systems that are truly equitable and contextually accurate on a global scale.

CRediT authorship contribution statement

Geleta Negasa Binegde: Writing – original draft, Conceptualization. **Huaping Zhang:** Supervision, Resources.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge financial support under the National Key Research and Development Program of China grant number SQ2024YFC3300144.

Data availability

No data was used for the research described in the article.

References

[1] J. Vinothkumar, A. Karunamurthy, Recent advancements in artificial intelligence technology: Trends and implications, Quing: Int. J. Multidiscip. Sci. Res. Dev. 2 (1) (2023) 1–11.

[2] Y. Xu, L. Hu, J. Zhao, Z. Qiu, Y. Ye, H. Gu, A survey on multilingual large language models: Corpora, alignment, and bias, 2024, arXiv preprint arXiv:2404.00929.

[3] S.L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of" bias" in nlp, 2020, arXiv preprint arXiv:2005.14050.

[4] M.A.K. Raiaan, M.S.H. Mukta, K. Fatema, N.M. Fahad, S. Sakib, M.M.J. Mim, J. Ahmad, M.E. Ali, S. Azam, A review on large language models: Architectures, applications, taxonomies, open issues and challenges, IEEE Access (2024).

[5] S. Banerjee, S. Layek, H. Shrawgi, R. Mandal, A. Halder, S. Kumar, S. Basu, P. Agrawal, R. Hazra, A. Mukherjee, Navigating the cultural kaleidoscope: A hitchhiker's guide to sensitivity in large language models, 2024, arXiv preprint arXiv:2410.12880.

[6] S. Pawar, J. Park, J. Jin, A. Arora, J. Myung, S. Yadav, F.G. Haznitrama, I. Song, A. Oh, I. Augenstein, Survey of cultural awareness in language models: Text and beyond, 2024, arXiv preprint arXiv:2411.00860.

[7] X. Zhang, S. Li, B. Hauer, N. Shi, G. Kondrak, Don't trust chatGPT when your question is not in english: A study of multilingual abilities and types of LLMs, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 7915–7927, http://dx.doi.org/10.18653/v1/2023.emnlp-main.491, URL https://aclanthology.org/2023.emnlp-main.491.

[8] G. Hofstede, M.H. Bond, Hofstede's culture dimensions: An independent validation using rokeach's value survey, J. Cross-Cultural Psychol. 15 (4) (1984) 417–433.

[9] J. Seo, J. Lee, C. Park, S. Hong, S. Lee, H.-S. Lim, Kocommongen v2: A benchmark for navigating korean commonsense reasoning challenges in large language models, in: Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 2390–2415.

[10] S. Shen, L. Logeswaran, M. Lee, H. Lee, S. Poria, R. Mihalcea, Understanding the capabilities and limitations of large language models for cultural commonsense, 2024, arXiv preprint arXiv:2405.04655.

[11] L. Bauer, H. Tischer, M. Bansal, Social commonsense for explanation and cultural bias discovery, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3745–3760.

[12] J. Huang, D. Yang, Culturally aware natural language inference, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 7591–7609.

[13] Z. Liu, Cultural bias in large language models: A comprehensive analysis and mitigation strategies, J. Transcult. Commun. (2024).

[14] E. Kim, J. Suk, P. Oh, H. Yoo, J. Thorne, A. Oh, CLIck: A benchmark dataset of cultural and linguistic intelligence in Korean, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 3335–3346, URL https://aclanthology.org/2024.lrec-main.296.

[15] Y. Wang, Y. Zhu, C. Kong, S. Wei, X. Yi, X. Xie, J. Sang, CDEval: A benchmark for measuring the cultural dimensions of large language models, 2023, arXiv preprint arXiv:2311.16421.

[16] Y. Tao, O. Viberg, R.S. Baker, R.F. Kizilcec, Cultural bias and cultural alignment of large language models, PNAS Nexus 3 (9) (2024) pgae346.

[17] J. Myung, N. Lee, Y. Zhou, J. Jin, R.A. Putri, D. Antypas, H. Borkakoty, E. Kim, C. Pérez-Almendros, A.A. Ayele, V. Guti'errez-Basulto, Y. Ib'anez-Garc'ia, H. Lee, S.H. Muhammad, K. Park, A. Rzayev, N. White, S.M. Yimam, M.T. Pilehvar, N.D. Ousidhoum, J. Camacho-Collados, A. Oh, Blend: A benchmark for LLMs on everyday knowledge in diverse cultures and languages, 2024, arXiv abs/2406.09948, URL https://api.semanticscholar.org/CorpusID:270521296.

[18] I.O. Gallegos, R.A. Rossi, J. Barrow, M.M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N.K. Ahmed, Bias and fairness in large language models: A survey, Comput. Linguist. (2024) 1–79.

[19] T.-P. Nguyen, S. Razniewski, A. Varde, G. Weikum, Extracting cultural commonsense knowledge at scale, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 1907–1917.

[20] T. Bolukbasi, K.-W. Chang, J.Y. Zou, V. Saligrama, A.T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Adv. Neural Inf. Process. Syst. 29 (2016).

[21] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, 2020, arXiv preprint arXiv:2009.03300.

[22] S. Singh, A. Romanou, C. Fourrier, D.I. Adelani, J.G. Ngui, D. Vila-Suero, P. Limkonchotiwat, K. Marchisio, W.Q. Leong, Y. Susanto, et al., Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2024, arXiv preprint arXiv:2412.03304.

[23] Y. Luo, J.S. Du, L. Qiu, On the cultural sensitivity of large language models: Gpt's ability to simulate human self-concept, in: 2024 11th International Conference on Behavioural and Social Computing, BESC, IEEE, 2024, pp. 1–8.

[24] S. Pawar, J. Park, J. Jin, A. Arora, J. Myung, S. Yadav, F.G. Haznitrama, I. Song, A. Oh, I. Augenstein, Survey of cultural awareness in language models: Text and beyond, 2024, arXiv abs/2411.00860, URL https://api.semanticscholar.org/CorpusID:273811670.

[25] J. Liu, B. Fu, Responsible multilingual large language models: A survey of development, applications, and societal impact, 2024, arXiv preprint arXiv:2410.17532.

[26] Y. Luo, J.S. Du, L. Qiu, On the cultural sensitivity of large language models: Gpt's ability to simulate human self-concept, in: 2024 11th International Conference on Behavioural and Social Computing, BESC, 2024, pp. 1–8.

[27] N. Lee, C. Jung, A. Oh, Hate speech classifiers are culturally insensitive, in: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), 2023, pp. 35–46.

[28] Z. Lin, S. Guan, W. Zhang, H. Zhang, Y. Li, H. Zhang, Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models, Artif. Intell. Rev. 57 (9) (2024) 1–50.

[29] M. Nadeem, A. Bethke, S. Reddy, StereoSet: Measuring stereotypical bias in pretrained language models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 5356–5371, http://dx.doi.org/10.18653/v1/2021.acl-long.416, URL https://aclanthology.org/2021.acl-long.416.

[30] N. Shahbazi, Y. Lin, A. Asudeh, H.V. Jagadish, Representation bias in data: A survey on identification and resolution techniques, 55 (13s) (2023) http://dx.doi.org/10.1145/3588433.

[31] C.H. Flowerman, (Some) algorithmic bias as institutional bias, Ethics Inf. Technol. 25 (2) (2023) 24.

[32] J. Zhang, S. Jiang, S. Guo, S. Chen, Y. Xiao, H. Feng, J. Liang, M. He, S. Tao, H. Ma, CultureScope: A dimensional lens for probing cultural understanding in LLMs, 2025, arXiv preprint arXiv:2509.16188.

[33] D. Zhong, C. Wu, Y. Jiang, Y. Yuan, M.-g. Kim, Y. Nishio, C.-C. Shih, W. Wang, J.-C. Lai, X. Ji, et al., High-speed and large-scale intrinsically stretchable integrated circuits, Nature 627 (8003) (2024) 313–320.

[34] S. Nie, M. Fromm, C. Welch, R. Görge, A. Karimi, J. Plepi, N.A. Mowmita, N. Flores-Herr, M. Ali, L. Flek, Do multilingual large language models mitigate stereotype bias?, 2024, arXiv abs/2407.05740URL https://api.semanticscholar.org/CorpusID:271050875.

[35] A. Stafanovičs, T. Bergmanis, M. Pinnis, Mitigating gender bias in machine translation with target gender annotations, in: Proceedings of the Fifth Conference on Machine Translation, 2020, pp. 629–638.

[36] M.Z. Khan, AI revolutionizing content diversity and cultural sensitivity in India, Int. J. Cult. Stud. Soc. Sci. (2024) 124–130.

[37] T. Naous, M.J. Ryan, W. Xu, Having beer after prayer? Measuring cultural bias in large language models, in: Annual Meeting of the Association for Computational Linguistics, 2023, URL https://api.semanticscholar.org/CorpusID:258865272.

[38] V. Pham, S. Qu, F. Moghimifar, S. Sharma, Y.-F. Li, W. Wang, R. Haf, Multi-cultural norm base: Frame-based norm discovery in multi-cultural settings, in: Proceedings of the 28th Conference on Computational Natural Language Learning, 2024, pp. 24–35.

[39] Y. Wang, Y. Zhu, C. Kong, S. Wei, X. Yi, X. Xie, J. Sang, CDEval: A benchmark for measuring the cultural dimensions of large language models, in: Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 1–16, http://dx.doi.org/10.18653/v1/2024.c3nlp-1.1, URL https://aclanthology.org/2024.c3nlp-1.1.

[40] A. Caliskan, J.J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (6334) (2017) 183–186.

[41] S. Shen, L. Logeswaran, M. Lee, H. Lee, S. Poria, R. Mihalcea, Understanding the capabilities and limitations of large language models for cultural commonsense, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 5668–5680, http://dx.doi.org/10.18653/v1/2024.naacl-long.316, URL https://aclanthology.org/2024.naacl-long.316.

[42] N. Lee, C. Jung, J. Myung, J. Jin, J. Camacho-Collados, J. Kim, A. Oh, Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 4205–4224.

[43] T. Naous, M. Ryan, A. Ritter, W. Xu, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Having Beer after Prayer? Measuring Cultural Bias in Large Language Models, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 16366–16393, http://dx.doi.org/10.18653/v1/2024.acl-long.862, URL https://aclanthology.org/2024.acl-long.862.

[44] S. Mirza, B. Coelho, Y. Cui, C. Pöpper, D. McCoy, Global-liar: Factuality of LLMs over time and geographic regions, 2024, ArXiv abs/2401.17839, URL https://api.semanticscholar.org/CorpusID:267334746.

[45] M.A. Hasan, M. Hasanain, F. Ahmad, S.R. Laskar, S. Upadhyay, V.N. Sukhadia, M. Kutlu, S.A. Chowdhury, F. Alam, Nativqa: Multilingual culturally-aligned natural query for LLMs, 2024, arXiv abs/2407.09823, URL https://api.semanticscholar.org/CorpusID:271212597.

[46] W.Q. Leong, J.G. Ngui, Y. Susanto, H. Rengarajan, K. Sarveswaran, W.C. Tjhi, Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models, 2023, arXiv preprint arXiv:2309.06085.

[47] J. Jin, J. Kim, N. Lee, H. Yoo, A. Oh, H. Lee, KoBBQ: Korean bias benchmark for question answering, Trans. Assoc. Comput. Linguist. 12 (2023) 507–524, URL https://api.semanticscholar.org/CorpusID:260334475.

[48] A. Keleg, W. Magdy, 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, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 6245–6266, http://dx.doi.org/10.18653/v1/2023.findings-acl.389, URL https://aclanthology.org/2023.findings-acl.389/.

[49] B. Mousi, N. Durrani, F. Ahmad, M.A. Hasan, M. Hasanain, T. Kabbani, F. Dalvi, S.A. Chowdhury, F. Alam, AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs, 2024, ArXiv abs/2409.11404, URL https://api.semanticscholar.org/CorpusID:272693942.

[50] Y. Mohamed, R. Li, I.S. Ahmad, K. Haydarov, P. Torr, K. Church, M. Elhoseiny, No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 20939–20962, http://dx.doi.org/10.18653/v1/2024.emnlp-main.1165, URL https://aclanthology.org/2024.emnlp-main.1165/.

[51] B. Wang, Z. Liu, X. Huang, F. Jiao, Y. Ding, A. Aw, N.F. Chen, SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning, 2023, arXiv preprint arXiv:2309.04766.

[52] T. Hu, M. Maistro, D. Hershcovich, Bridging cultures in the kitchen: A framework and benchmark for cross-cultural recipe retrieval, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 1068–1080.

[53] S. Nayak, K. Jain, R. Awal, S. Reddy, S. van Steenkiste, L.A. Hendricks, K. Stańczak, A. Agrawal, Benchmarking vision language models for cultural understanding, 2024, arXiv abs/2407.10920, URL https://api.semanticscholar.org/CorpusID:271212228.

[54] E. Karinshak, A. Hu, K. Kong, V. Rao, J. Wang, J. Wang, Y. Zeng, LLM-GLOBE: A benchmark evaluating the cultural values embedded in llm output, 2024, arXiv preprint arXiv:2411.06032.

[55] H.-Y. Hsieh, S.-C. Huang, R.T.-H. Tsai, Twbias: A benchmark for assessing social bias in traditional Chinese large language models through a Taiwan cultural lens, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 8688–8704.

[56] Z. Qian, F. Altam, M. Alqurishi, R. Souissi, CamelEval: Advancing culturally aligned arabic language models and benchmarks, 2024, arXiv preprint arXiv:2409.12623.

[57] F. Schneider, S. Sitaram, M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 4309–4345, http://dx.doi.org/10.18653/v1/2024.findings-emnlp.250, URL https://aclanthology.org/2024.findings-emnlp.250/.

[58] H. Singh, N. Gupta, S. Bharadwaj, D. Tewari, P. Talukdar, IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on indic languages, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 11047–11073, http://dx.doi.org/10.18653/v1/2024.acl-long.595, URL https://aclanthology.org/2024.acl-long.595/.

[59] I. Watts, V. Gumma, A. Yadavalli, V. Seshadri, M. Swaminathan, S. Sitaram, Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data, 2024, arXiv preprint arXiv:2406.15053.

[60] C. Li, M. Chen, J. Wang, S. Sitaram, X. Xie, Culturellm: Incorporating cultural differences into large language models, 2024, arXiv preprint arXiv:2402.10946.

[61] C. Li, D. Teney, L. Yang, Q. Wen, X. Xie, J. Wang, CulturePark: Boosting cross-cultural understanding in large language models, 2024, arXiv preprint arXiv:2405.15145.

[62] K. Lu, P. Mardziel, F. Wu, P. Amancharla, A. Datta, Gender bias in neural natural language processing, in: Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday, Springer, 2020, pp. 189–202.

[63] B. Ding, Q. Min, S. Ma, Y. Li, L. Yang, Y. Zhang, A rationale-centric counterfactual data augmentation method for cross-document event coreference resolution, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1112–1140, http://dx.doi.org/10.18653/v1/2024.naacl-long.63, URL https://aclanthology.org/2024.naacl-long.63.

[64] M. Suffian, A. Bogliolo, Investigation and mitigation of bias in explainable AI, in: CEUR Workshop Proceedings, vol. 3319, 2022, pp. 89–94.

[65] S. Steindl, U. Schäfer, B. Ludwig, Counterfactual dialog mixing as data augmentation for task-oriented dialog systems, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 4078–4087, URL https://aclanthology.org/2024.lrec-main.363.

[66] S. Pitis, E. Creager, A. Mandlekar, A. Garg, Mocoda: Model-based counterfactual data augmentation, Adv. Neural Inf. Process. Syst. 35 (2022) 18143–18156.

[67] A. Balashankar, X. Wang, Y. Qin, B. Packer, N. Thain, E. Chi, J. Chen, A. Beutel, Improving classifier robustness through active generative counterfactual data augmentation, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 127–139, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.10, URL https://aclanthology.org/2023.findings-emnlp.10.

[68] Y. Li, M. Du, X. Wang, Y. Wang, Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 14254–14267, http://dx.doi.org/10.18653/v1/2023.acl-long.797, URL https://aclanthology.org/2023.acl-long.797.

[69] T. Lukasiewicz, Z. Xie, An empirical analysis of parameter- efficient methods for debiasing pre- trained language models, 2023.

[70] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, in: International Conference on Machine Learning, PMLR, 2013, pp. 1319–1327.

[71] B.H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340.

[72] J.S. Ernst, S. Marton, J. Brinkmann, E. Vellasques, D. Foucard, M. Kraemer, M. Lambert, Bias mitigation for large language models using adversarial learning, in: CEUR Workshop Proceedings, vol. 3523, RWTH Aachen, 2023, pp. 1–14.

[73] C. Szegedy, Intriguing properties of neural networks, 2013, arXiv preprint arXiv:1312.6199.

[74] X.L. Dong, Y. Zhu, Z. Fu, D. Xu, G. De Melo, Data augmentation with adversarial training for cross-lingual nli, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5158–5167.

[75] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, ITCS '12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 214–226, http://dx.doi.org/10.1145/2090236.2090255.

[76] M. Wan, D. Zha, N. Liu, N. Zou, In-Processing Modeling Techniques for Machine Learning Fairness: A Survey, vol. 17, (3) Association for Computing Machinery, New York, NY, USA, 2023, http://dx.doi.org/10.1145/3551390.

[77] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012, pp. 214–226.

[78] H. Edwards, A. Storkey, Censoring representations with an adversary, 2015, arXiv preprint arXiv:1511.05897.

[79] T. Hashimoto, M. Srivastava, H. Namkoong, P. Liang, Fairness without demographics in repeated loss minimization, in: International Conference on Machine Learning, PMLR, 2018, pp. 1929–1938.

[80] X. Shen, S. Diamond, Y. Gu, S. Boyd, Disciplined convex-concave programming, in: 2016 IEEE 55th Conference on Decision and Control, CDC, IEEE, 2016, pp. 1009–1014.

[81] M. Kearns, S. Neel, A. Roth, Z.S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: International Conference on Machine Learning, PMLR, 2018, pp. 2564–2572.

[82] S. Garg, V. Perot, N. Limtiaco, A. Taly, E.H. Chi, A. Beutel, Counterfactual fairness in text classification through robustness, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 219–226.

[83] Y. Qian, U. Muaz, B. Zhang, J.W. Hyun, Reducing gender bias in word-level language models with a gender-equalizing loss function, 2019, arXiv preprint arXiv:1905.12801.

[84] M. Kaneko, D. Bollegala, Debiasing pre-trained contextualised embeddings, 2021, arXiv, URL https://api.semanticscholar.org/CorpusID:231698657.

[85] R. Su, Z. Sun, S. Lu, C. Ma, C. Guo, Clicker: Attention-based cross-lingual commonsense knowledge transfer, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.

[86] T.-P. Nguyen, S. Razniewski, A. Varde, G. Weikum, Extracting cultural commonsense knowledge at scale, in: Proceedings of the ACM Web Conference 2023, WWW '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1907–1917, http://dx.doi.org/10.1145/3543507.3583535.

[87] J. Ju, D. Yang, J. Liu, Commonsense knowledge base completion with relational graph attention network and pre-trained language model, in: Proceedings of the 31st ACM International Con- ference on Information & Knowledge Management, 2022, URL https://api.semanticscholar.org/CorpusID:252904784.

[88] J.D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, Y. Choi, On symbolic and neural commonsense knowledge graphs, 2021.

[89] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, (1) 2017.

[90] P. Singh, T. Lin, E.T. Mueller, G. Lim, T. Perkins, W. Li Zhu, Open mind common sense: Knowledge acquisition from the general public, in: On the Move To Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE: Confederated International Conferences CoopIS, DOA, and ODBASE 2002 Proceedings, Springer, 2002, pp. 1223–1237.

[91] J.D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, Y. Choi, COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs, in: AAAI Conference on Artificial Intelligence, 2020, URL https://api.semanticscholar.org/CorpusID:222310337.

[92] H. Zhang, D. Khashabi, Y. Song, D. Roth, Transomcs: From linguistic graphs to commonsense knowledge, 2020, arXiv preprint arXiv:2005.00206.

[93] L. Fang, Y. Luo, K. Feng, K. Zhao, A. Hu, A knowledge-enriched ensemble method for word embedding and multi-sense embedding, IEEE Trans. Knowl. Data Eng. 35 (6) (2023) 5534–5549, http://dx.doi.org/10.1109/TKDE.2022.3159539.

[94] R. Takahashi, R. Tian, K. Inui, Interpretable and compositional relation learning by joint training with an autoencoder, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2148–2159, http://dx.doi.org/10.18653/v1/P18-1200, URL https://aclanthology.org/P18-1200.

[95] W. Zhang, B. Paudel, W. Zhang, A. Bernstein, H. Chen, Interaction embeddings for prediction and explanation in knowledge graphs, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 96–104.

[96] B.Y. Lin, S. Lee, X. Qiao, X. Ren, Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1274–1287, http://dx.doi.org/10.18653/v1/2021.acl-long.102, URL https://aclanthology.org/2021.acl-long.102.

[97] S. Aggarwal, D. Mandowara, V. Agrawal, D. Khandelwal, P. Singla, D. Garg, Explanations for commonsenseqa: New dataset and models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 3050–3065.

[98] Y.R. Fung, T. Chakraborty, H. Guo, O. Rambow, S. Muresan, H. Ji, Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly, 2022, arXiv preprint arXiv:2210.08604.

[99] V. Ondrejová, M. Šuppa, Can LLMs handle low-resource dialects? A case study on translation and common sense reasoning in šariš, in: Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), 2024, pp. 130–139.

[100] W. Zhang, M. Aljunied, C. Gao, Y.K. Chia, L. Bing, M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, Adv. Neural Inf. Process. Syst. 36 (2023) 5484–5505.

[101] N. Moghe, A. Fazla, C. Amrhein, T. Kocmi, M. Steedman, A. Birch, R. Sennrich, L. Guillou, Machine translation meta evaluation through translation accuracy challenge sets, Comput. Linguist. (2024) 1–65.

[102] T.-P. Nguyen, S. Razniewski, G. Weikum, Advanced semantics for commonsense knowledge extraction, in: Proceedings of the Web Conference 2021, 2021, pp. 2636–2647.

[103] M. Kejriwal, H. Santos, K. Shen, A.M. Mulvehill, D.L. McGuinness, Context-rich evaluation of machine common sense, in: International Conference on Artificial General Intelligence, Springer, 2023, pp. 167–176.

[104] L.Q.T. Nguyen, L.D.T. Huynh, Bank culture and bank liquidity creation, Corp. Gov.: An Int. Rev. 32 (6) (2024) 1087–1109.

[105] N. Ibrahim, S. Aboulela, A. Ibrahim, R. Kashef, A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): models, evaluation metrics, benchmarks, and challenges, Discov. Artif. Intell. 4 (1) (2024) 76.

[106] E. Hwang, V. Thost, V. Shwartz, T. Ma, Knowledge graph compression enhances diverse commonsense generation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 558–572.

[107] J. Omeliyanenko, A. Zehe, L. Hettinger, A. Hotho, Lm4kg: Improving common sense knowledge graphs with language models, in: The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part I 19, Springer, 2020, pp. 456–473.

[108] X. Zhang, S. Li, B. Hauer, N. Shi, G. Kondrak, Don't trust ChatGPT when your question is not in english: A study of multilingual abilities and types of LLMs, in: Conference on Empirical Methods in Natural Language Processing, 2023, URL https://api.semanticscholar.org/CorpusID:258947405.

[109] F. Koto, R. Mahendra, N. Aisyah, T. Baldwin, IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces, Trans. Assoc. Comput. Linguist. 12 (2024) 1703–1719.

[110] P. Lahoti, N. Blumm, X. Ma, R. Kotikalapudi, S. Potluri, Q. Tan, H. Srinivasan, B. Packer, A. Beirami, A. Beutel, et al., Improving diversity of demographic representation in large language models via collective-critiques and self-voting, 2023, arXiv preprint arXiv:2310.16523.

[111] W. Li, R. Ma, J. Wu, C. Gu, J. Peng, J. Len, S. Zhang, H. Yan, D. Lin, C. He, FoundaBench: Evaluating Chinese fundamental knowledge capabilities of large language models, 2024, arXiv preprint arXiv:2404.18359.

[112] H. Kim, J. Jung, D. Jeong, J. Nam, K-pop lyric translation: Dataset, analysis, and neural-modelling, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 9974–9987, URL https://aclanthology.org/2024.lrec-main.872/.

[113] C.-C. Chang, C.-Y. Chen, H.-S. Lee, C.-C. Lee, Benchmarking cognitive domains for LLMs: Insights from Taiwanese hakka culture, in: 2024 27th Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE, 2024, pp. 1–6.

[114] A. Seth, S. Ahuja, K. Bali, S. Sitaram, DOSA: A dataset of social artifacts from different Indian geographical subcultures, 2024, arXiv abs/2403.14651, URL https://api.semanticscholar.org/CorpusID:268667574.

[115] B. Mousi, N. Durrani, F. Ahmad, M.A. Hasan, M. Hasanain, T. Kabbani, F. Dalvi, S.A. Chowdhury, F. Alam, Aradice: Benchmarks for dialectal and cultural capabilities in llms, 2024, arXiv preprint arXiv:2409.11404.

[116] O.J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L.A.B. Zoph, GPT-4 technical report, (2023) URL https://api.semanticscholar.org/CorpusID:257532815.

[117] G. Son, H. Lee, S. Kim, H. Kim, J. Lee, J.W. Yeom, J. Jung, J.W. Kim, S. Kim, Hae-rae bench: Evaluation of korean knowledge in language models, 2023, arXiv preprint arXiv:2309.02706.

[118] M. Masala, D.C. Ilie-Ablachim, A. Dima, D. Corlatescu, M. Zavelca, O. Olaru, S. Terian, A. Terian, M. Leordeanu, H. Velicu, et al., “ vorbe\c {s} ti rom\^ ane\c {s} te?" a recipe to train powerful Romanian LLMs with english instructions, 2024, arXiv preprint arXiv:2406.18266.

[119] A. Talmor, J. Herzig, N. Lourie, J. Berant, CommonsenseQA: A question answering challenge targeting commonsense knowledge, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4149–4158, http://dx.doi.org/10.18653/v1/N19-1421, URL https://aclanthology.org/N19-1421/.

[120] Y. Fung, R. Zhao, J. Doo, C. Sun, H. Ji, Massively multi-cultural knowledge acquisition & lm benchmarking, 2024, arXiv preprint arXiv:2402.09369.

[121] Y. Liu, M. Xu, S. Wang, L. Yang, H. Wang, Z. Liu, C. Kong, Y. Chen, M. Sun, E. Yang, Omgeval: An open multilingual generative evaluation benchmark for large language models, 2024, URL https://api.semanticscholar.org/CorpusID:267770422.

[122] H. Li, L. Jiang, N. Dziri, X. Ren, Y. Choi, CULTURE-GEN: Revealing global cultural perception in language models through natural language prompting, 2024, arXiv abs/2404.10199, URL https://api.semanticscholar.org/CorpusID:269157402.

[123] M. Bhatia, V. Shwartz, GD-COMET: A geo-diverse commonsense inference model, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 7993–8001, http://dx.doi.org/10.18653/v1/2023.emnlp-main.496, URL https://aclanthology.org/2023.emnlp-main.496/.

[124] L. Chua, B. Ghazi, Y. Huang, P. Kamath, R. Kumar, P. Manurangsi, A. Sinha, C. Xie, C. Zhang, Crosslingual capabilities and knowledge barriers in multilingual large language models, 2024, arXiv abs/2406.16135, URL https://api.semanticscholar.org/CorpusID:270703607.

[125] J. Zheng, Q. Ma, S. Qiu, Y. Wu, P. Ma, J. Liu, H. Feng, X. Shang, H. Chen, Preserving commonsense knowledge from pre-trained language models via causal inference, in: Annual Meeting of the Association for Computational Linguistics, 2023, URL https://api.semanticscholar.org/CorpusID:259203213.

[126] B. Wang, Z. Liu, X. Huang, F. Jiao, Y. Ding, A.T. Aw, N.F. Chen, SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning, 2023, arXiv abs/2309.04766, URL https://api.semanticscholar.org/CorpusID:261682140.

[127] T.-Y. Chang, Y. Liu, K. Gopalakrishnan, B. Hedayatnia, P. Zhou, D. Hakkani-Tur, Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks, in: Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Association for Computational Linguistics, Online, 2020, pp. 74–79, http://dx.doi.org/10.18653/v1/2020.deelio-1.9, URL https://aclanthology.org/2020.deelio-1.9/.

[128] Y. Razeghi, R.L.L. I.V., S. Singh, Deriving behavioral tests from common sense knowledge graphs, 2020, URL https://api.semanticscholar.org/CorpusID:229381747.

[129] C. Luo, A. Ghawanmeh, X. Zhu, F.K. Khattak, BiasKG: Adversarial knowledge graphs to induce bias in large language models, 2024, arXiv abs/2405.04756, URL https://api.semanticscholar.org/CorpusID:269626396.

[130] D. Gurgurov, M. Hartmann, S. Ostermann, Adapting multilingual LLMs to low-resource languages with knowledge graphs via adapters, 2024, arXiv abs/2407.01406, URL https://api.semanticscholar.org/CorpusID:270869492.

[131] T. Zhang, B. Peng, D. Bollegala, Improving diversity of commonsense generation by large language models via in-context learning, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 9226–9242, http://dx.doi.org/10.18653/v1/2024.findings-emnlp.540, URL https://aclanthology.org/2024.findings-emnlp.540/.

[132] Y. Sakai, H. Kamigaito, T. Watanabe, mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans, 2024, arXiv preprint arXiv:2406.04215.

[133] M. Bakker, M. Chadwick, H. Sheahan, M. Tessler, L. Campbell-Gillingham, J. Balaguer, N. McAleese, A. Glaese, J. Aslanides, M. Botvinick, et al., Fine-tuning language models to find agreement among humans with diverse preferences, Adv. Neural Inf. Process. Syst. 35 (2022) 38176–38189.

[134] V. Ramanujan, T. Nguyen, S. Oh, L. Schmidt, A. Farhadi, On the connection between pre-training data diversity and fine-tuning robustness, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23, Curran Associates Inc., Red Hook, NY, USA, 2024.

[135] J. Davison, J. Feldman, A. Rush, Commonsense knowledge mining from pretrained models, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1173–1178, http://dx.doi.org/10.18653/v1/D19-1109, URL https://aclanthology.org/D19-1109/.

[136] A. Vayani, D. Dissanayake, H. Watawana, N. Ahsan, N. Sasikumar, O. Thawakar, H.B. Ademtew, Y. Hmaiti, A. Kumar, K. Kuckreja, et al., All languages matter: Evaluating lmms on culturally diverse 100 languages, 2024, arXiv preprint arXiv:2411.16508.

[137] S. Nayak, K. Jain, R. Awal, S. Reddy, S. Van Steenkiste, L.A. Hendricks, A. Agrawal, et al., Benchmarking vision language models for cultural understanding, 2024, arXiv preprint arXiv:2407.10920.

[138] D. Yin, H. Bansal, M. Monajatipoor, L.H. Li, K.-W. Chang, Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models, 2022, arXiv preprint arXiv:2205.12247.

[139] B.Y. Lin, S. Lee, X. Qiao, X. Ren, Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning, 2021, arXiv preprint arXiv:2106.06937.

[140] H. Chung, GlobalMind: bridging the gap between different cultures and languages with common-sense computing (Ph.D. thesis), Massachusetts Institute of Technology, 2006.

[141] S. Nie, M. Fromm, C. Welch, R. Görge, A. Karimi, J. Plepi, N.A. Mowmita, N. Flores-Herr, M. Ali, L. Flek, Do multilingual large language models mitigate stereotype bias?, 2024, arXiv preprint arXiv:2407.05740.

[142] S.L. Blodgett, S. Barocas, H. Daum’e, H.M. Wallach, Language (technology) is power: A critical survey of “bias” in NLP, 2020, arXiv abs/2005.14050, URL https://api.semanticscholar.org/CorpusID:218971825.

[143] S. Shen, L. Logeswaran, M. Lee, H. Lee, S. Poria, R. Mihalcea, Understanding the capabilities and limitations of large language models for cultural commonsense, in: North American Chapter of the Association for Computational Linguistics, 2024, URL https://api.semanticscholar.org/CorpusID:269626686.

[144] I.O. Gallegos, R.A. Rossi, J. Barrow, M.M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N.K. Ahmed, Bias and fairness in large language models: A survey, Comput. Linguist. 50 (3) (2024) 1097–1179.

[145] F. Alam, S.A. Chowdhury, S. Boughorbel, M. Hasanain, LLMs for low resource languages in multilingual, multimodal and dialectal settings, in: Conference of the European Chapter of the Association for Computational Linguistics, 2024, URL https://api.semanticscholar.org/CorpusID:268417133.

[146] Z. Zhou, S. Conia, D. Lee, M. Li, S. Huang, U.F. Minhas, S. Potdar, H. Xiao, Y. Li, KG-TRICK: Unifying textual and relational information completion of knowledge for multilingual knowledge graphs, 2025, URL https://api.semanticscholar.org/CorpusID:275342539.

[147] L. Wang, W. Zhao, Z. Wei, J. Liu, SimKGC: Simple contrastive knowledge graph completion with pre-trained language models, 2022, ArXiv abs/2203.02167, URL https://api.semanticscholar.org/CorpusID:247244896.

[148] A. Mukherjee, A. Caliskan, Z. Zhu, A. Anastasopoulos, Global gallery: The fine art of painting culture portraits through multilingual instruction tuning, in: North American Chapter of the Association for Computational Linguistics, 2024, URL https://api.semanticscholar.org/CorpusID:270514431.