A Nightmare on LLMs Street: On the Importance of Cultural Awareness in Text Adaptation for LRLs

David C. T. Freitas[a,*] **and Henrique Lopes Cardoso**[a]

[a]LIACC, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

Abstract. Large Language Models (LLMs) have revolutionized how we generate, interact with, and process language. Still, these models are biased toward WEIRD (Western, Educated, Industrialized, Rich, and Democratic) values. This bias is not merely linguistic but also cultural. Sociocultural contexts influence how people express ideas, interpret meaning, and communicate. In low-resource language settings, where data and cultural representation are limited, this issue becomes even more pronounced when models are applied without cultural adaptation, often leading to outputs that are irrelevant, inaccessible, or even harmful. In this paper, we argue for the importance of incorporating sociocultural context into LLMs. We review existing frameworks that explore culture in Natural Language Processing (NLP), and examine some work aimed at culturally aligning language models. As an illustrative scenario, we analyze the case of Guinea-Bissau. In this linguistically and culturally diverse country, Portuguese is the official language but not the primary means of communication for most of the population, highlighting the urgent need to adapt educational materials to the local sociocultural context. Finally, we propose a revised framework to address the challenge of adapting educational materials to diverse contexts, aiming to improve both the relevance and pedagogical impact of text adaptation.

1 Introduction

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), enabling widespread and seemingly universal interaction with Artificial Intelligence (AI) systems — perhaps for the first time. The abilities of LLMs to understand language rely not only on linguistic and factual knowledge, but also on an awareness of cultural nuances that shape human lives. However, these models are mostly trained on online data that is deeply rooted in a WEIRD (Western, Educated, Industrialized, Rich, and Democratic) worldview [8]. Because the majority of users who access these systems share, or are aligned with, that worldview, this creates a misleading impression that the models adequately represent and understand the world's linguistic and cultural diversity.

In reality, a large portion of the world's population lives in profoundly different sociocultural contexts, with customs, norms, values, and shared knowledge that are not reflected in the data used to train these models. These elements, essential for contextual understanding, vary across cultural groups and are often not captured when NLP is built upon universalist assumptions. The problem is amplified when it comes to Low-Resource Languages (LRLs), whose cultural contexts are often overlooked, for which very few linguistic resources, annotated data, or corpora exist. It is estimated that there are over 7,000 languages in the world[1], but only a small fraction are covered by LLMs.

In light of this, although significant progress has been made in NLP, culture remains among the most challenging aspects of language that LLMs still struggle to handle effectively [1]. A key difficulty lies in the absence of a shared definition of culture, which further complicates efforts to evaluate progress in this area [17].

Guinea-Bissau represents a paradigmatic case of a country left behind in the global access to technology and quality education. This is exacerbated by two key factors. First, Portuguese is the official language and the language of instruction in schools, yet only a minority of the population speaks it fluently. Second, most people communicate in Guinea-Bissau Creole, the *lingua franca* used in everyday life, which lacks official status and a standardized orthography.

This paper sets out to argue for the urgent need to address the lack of cultural awareness in LLMs, identifying their limitations. It advocates for a more context-sensitive approach and explores how tackling these issues can help reduce social and linguistic inequalities in the case of Guinea-Bissau, by enabling the creation and adaptation of content that is both culturally relevant and linguistically appropriate.

This paper should be understood as a position paper grounded in interdisciplinary perspectives from NLP, education, sociolinguistics, and cultural studies. Our aim is not to present experimental results, but to outline conceptual and methodological directions for the culturally informed adaptation of texts in LRLs settings.

2 Background and Motivation

Our culture and our social relations shape everything we do. The way we present information, the style and tone we use, the context, and the common knowledge we share, among many other subtleties, are essential to effective communication. All of these are based on cultural knowledge.

For the purpose of this paper, we adopt the definition of culture proposed by Liu et al. [18], according to which culture encompasses:

> "the collective ideas, shared language, and social practices that emerge from and evolve through human social interactions within a society".

Messages that are not culturally adapted can be misinterpreted, and language technologies must account for cultural context to avoid

* Corresponding Author. Email: up199400804@edu.fe.up.pt

[1] https://www.ethnologue.com/

potential harm [9]. Given its crucial role in making LLMs safer, fairer, and more inclusive, the concept of culture is receiving increasing attention in current research.

One of the most important decisions when adapting texts, especially in translation, is deciding whether to aim for literalism or adaptation [28]. Cross-cultural translation and adaptation highlight this complexity, when the same literal meaning in one culture may be inappropriate in another, regardless of fluency [9], or when a direct counterpart in the target culture may not exist. When no direct counterpart exists in the target culture, appositives can be employed to provide contextual explanations. For example, the Portuguese sentence "Encontraram-se na queima das fitas" could be translated to English as "They met at the *queima das fitas*, a traditional Portuguese academic celebration", where the appositive clarifies a culturally specific term that may be unfamiliar to the target audience.

While many LLMs are capable of performing cultural adaptations effectively within WEIRD contexts, they often exhibit biases that hinder their applicability in more culturally diverse settings. Certain cultural variations can be identified through sociocultural elements, such as Culture-Specific Items (CSIs) — including aspects of ecology (flora, fauna, climate, ...), material culture (food, clothing, housing, transportation, ...), emotions, and socially sensitive or taboo topics [28]. Yet, such categorizations are insufficient, particularly because culture is not a static inventory of features, but a dynamic and evolving construct. Moreover, every communicative act is situated within a relational structure involving an emitter, a receiver, the medium through which the message is conveyed, and the broader sociocultural context in which it occurs — all of which shape the interpretation and appropriateness of the message.

Despite this growing awareness, many current approaches rely on language or national borders as proxies for cultural identity. However, this is problematic because significant cultural variation can exist within the same country or among countries that share a common language. For example, Portugal and Brazil speak the same language but differ considerably in cultural norms, values, and communicative practices. Similarly, treating entire countries as culturally homogeneous units oversimplifies internal diversity. Such categorical division risks minimizing cultural distinctions under a single label.

Although there is no common definition of culturally aware NLP, most works in this area share common goals. Zhou et al. [37] define the goals of culturally aware NLP as systems that are:

Adaptive: sensitive to specific cultural contexts when generating their outputs.
Discerning: not perpetuating reductive stereotypes.
Inclusive: perform well across a large number of cultures.
Nuanced: achieve depth through more granular and extensive cultural understanding.

These four goals share common ground and, according to the authors, can be grouped into two distinct spaces. The first space combines the "Adaptive" and "Discerning" goals and reflects how systems should respond and when it is appropriate to do so. The second space includes the "Inclusive" and "Nuanced" goals, reflecting the desire for both broad cultural coverage and depth. For each of these spaces, there is often a trade-off.

3 Research on Culturally Aware NLP

In this section, we first present key frameworks that aim to conceptualize culture in NLP. We then use the taxonomy proposed by Liu et al. [18] to organize and analyze recent research, highlighting how different cultural dimensions have been explored in the literature.

3.1 Frameworks

The complexity of rewriting texts according to sociocultural contexts requires a systematic approach that enables systems to identify and mitigate cultural mismatches and potential biases. Frameworks provide essential conceptual structures that support this process by guiding the understanding, representation, and operationalization of cultural elements. In this section, we review three frameworks that attempt to organize the notion of culture within NLP. This is a challenging endeavor, further complicated by the lack of a consensual definition of "culture". Moreover, the objectives of each framework shape how culture is interpreted and framed, influencing which dimensions are prioritized. Hershcovich et al. [9] propose a framework based on four fundamental communicative dimensions; Adilazuarda et al. [1] approach culture through observable proxies divided into demographic and semantic categories; and Liu et al. [18] present a taxonomy inspired by the social sciences and anthropology, focusing on the comprehensiveness and operationalization of cultural and sociocultural elements in NLP.

Hershcovich et al. [9] laid the foundation for the importance of culture in NLP. The work focuses on how cultural interaction is intertwined with language and proposes a framework for understanding the challenges that cultural diversity poses. It defines the role of culture in four dimensions:

Linguistic Form and Style Sociocultural factors shape how things are formulated and expressed. Variations within a language, such as dialects, sociolects, or stylistic differences, must be taken into account. As previously discussed, the common practice of dividing cultures by language or region should be re-evaluated, as it is a mistake to homogenize individuals sharing the same language.
Common Ground How the knowledge shared between individuals varies across cultures is essential to determine what needs to be communicated and how. In particular, conceptualisation and commonsense knowledge influence comprehension, reasoning, and entailment.
Aboutness What is considered relevant or worth promoting in certain cultures should be considered when generating or curating information.
Objectives and Values Values differ across cultures, influencing what is accepted or prioritized. For example, alcohol is culturally relevant in Portugal but taboo in Muslim cultures. Reconciling differing objectives may lead to conflict, especially when dominant cultures are involved. These tensions are often difficult to resolve due to the trade-off between reducing bias and respecting core cultural values.

Adilazuarda et al. [1] propose a different taxonomy, in which they identify various aspects of culture that serve as proxies. They consider 12 distinct proxies, grouped into two overarching categories:

Demographic proxies Ethnicity, education, race, gender, language, and religion.
Semantic proxies Emotions and values, food and drink, social and political relations, basic actions and technology, names, and the domain of quantity, time, kinship, pronouns and function words.

The authors justified this division by the fact that **demographic proxies** relate to culture as it is often defined at the community or

group level, where the individual is embedded, while **semantic proxies** refer to the products consumed, actions and social relations, and shared values. The authors also note that, although some proxies are well-studied, many have been little or not at all explored, such as the semantic domain of quantity, time, kinship, pronouns and function words, spatial relations, aspects of the physical and mental world, the body, among others.

Liu et al. [18] present a taxonomy, grounded in well-established elements of culture in anthropology and social sciences, divided into three branches: *Ideational*, *Linguistic*, and *Social*.

Ideational Includes non-material aspects of culture, such as values or knowledge. This branch is further divided into five subbranches:

- **Concepts** Basic units of meaning, such as cuisine, holidays, proverbs, time expressions, and so on.
- **Knowledge** Information that is acquired through education or practical experience.
- **Values** The values shared among groups influence what is relevant (aboutness), the style of communication, and the standards of a culture.
- **Norms and Morals** Rules or principles that guide people's behavior and everyday reasoning. Unlike the domain of values, here, there is ethical judgment.
- **Artifacts** Products of human culture like songs, tales, poetry, movies, humor, and so on.

Linguistic Focuses on cultural variations in language and linguistic forms. Two key aspects are considered:

- **Dialects** Systematic variations of a language typically associated with regional, national, or social groups. These include phonological, lexical, and syntactic differences.
- **Styles, Registers, Genres** Context-dependent ways of using language, influenced by factors such as formality, social roles, or communicative purpose. Examples include slang, technical jargon, academic writing, or informal conversation.

Social Considers the social, interpersonal, and contextual factors that influence how language is shaped, interpreted, and negotiated in interaction.

- **Relationship** How the connection between individuals or groups (father-son, elder-younger, . . .) influences communication.
- **Context** The influence of contextual factors such as linguistic, social, historical, or non-verbal cues on the interpretation and production of communication.
- **Communicative Goals** The intention, such as requests, apologies, persuasion, behind the use of language.
- **Demographics** Population characteristics such as age, political orientation, or socioeconomic status, that influence how people communicate and what they expect.

Of the three taxonomies examined, the one proposed by Liu et al. [18] stands out for its greater level of detail and for more effectively systematizing cultural elements, with particular emphasis on interaction and communicative context. It is also fair to note that these frameworks are not mutually exclusive; given the inherently fluid and multifaceted nature of culture, meaningful connections can be drawn among all of them.

3.2 Ideational

In this section, we examine how LLMs "understand" the Ideational dimension, following the taxonomy defined by Liu et al. [18].

Some research has explored the representation of concepts through metaphors and other figurative expressions, as in Kabra et al. [11] and Liu et al. [16]. Figurative language reflects cultural and societal experiences, making such expressions difficult to generalize across languages. Kabra et al. [11] focus on figurative language understanding across multiple languages, highlighting that existing datasets and models are often biased toward English.

Liu et al. [16] introduce MAPS—a dataset of proverbs across six geographically and typologically diverse languages (English, German, Russian, Bengali, Mandarin Chinese, and Indonesian)—and investigate whether Multilingual Large Language Models (mLLMs) can interpret the meaning of a proverb in context and reason cross-culturally when the proverb is translated into another language. The authors evaluate a range of state-of-the-art multilingual models, including XLM-R, mT0, BLOOMZ, XGLM, and LLaMA-2.

Their study shows that models consistently perform worse on figurative proverbs than on literal ones, with Chinese being a notable exception. They also find that figurative proverbs are harder to interpret, with reasoning gaps being common. When reasoning with translated proverbs, models exhibit substantial drops in performance, suggesting that cultural knowledge embedded in figurative language does not transfer well across languages. Even with human-adapted translations, model performance fails to match that achieved in the original language. They conclude that LLMs partially understand proverbs, but often fail to reason with them correctly, especially in cross-cultural or figurative cases. More interestingly, when the authors ask the model to pick the wrong answer, all previously well-performing models perform poorly.

Shwartz [27] proposes culture-specific time expression grounding, mapping expressions such as "morning" (or"manhã" in Portuguese) to the corresponding time intervals. Such grounding exhibits cultural variations, like average wake and sleep times, and can provide context for NLP tasks such as event ordering, duration prediction, cultural adaptation in dialogue systems, and machine translation (MT).

Taking into account that language models can be used as knowledge bases [10, 24], some papers [18] explore ways of evaluating and integrating cultural knowledge in NLP, using probing to test what pre-trained NLP models already know about cultural concepts. Probing is a method used to explore the internal workings of pre-trained language models to see what kind of linguistic or factual knowledge they have acquired during training. Probing tests are designed to reveal whether a model can correctly answer questions or fill in missing parts of a sentence based on its learned knowledge. A sentence with missing information is given to the model:

> *In Guinea-Bissau, the first meal of the day is called [MASK], while in Portugal it is called "pequeno-almoço".*

The model tries to predict the masked word (e.g., "mata-bicho"). If the model correctly fills in culturally accurate words, it means it has internalized cultural knowledge.

Zhou et al. [36] introduce FMLAMA, a multilingual dataset designed to probe LLMs for food-related cultural facts and variations in food practices. Using this dataset, the authors evaluate LLMs across different architectures and languages, uncovering systematic cultural biases and knowledge retrieval limitations. To test whether LLMs possess culturally grounded knowledge in the food domain, they use prompts such as `[X] is a dish made with [Y]` and `[X]`

`is a type of food that includes [Y]`. Their study reveals that LLMs demonstrate a pronounced bias towards food knowledge prevalent in the United States.

Regarding values, Sorensen et al. [29] explore the notion of value pluralism — the idea that different human values can lead to distinct, though potentially equally valid, decisions. In this paper, the authors investigate the potential of LLMs to model pluralistic human values, rights, and duties. To this end, they introduce VALUEPRISM, a large-scale dataset of pluralistic human values, and build VALUE KALEIDOSCOPE (KALEIDO), an open and flexible value-pluralistic model. The authors compare GPT-4 and their own model by asking both to generate values for the same situations. KALEIDO, trained on GPT-4 outputs, is able to explain and reason about values, with 91% of the generated values, rights, and duties marked as good by all three human annotators. However, the authors caution that the generated data may reflect the values of dominant groups rather than a truly diverse set.

Zhan et al. [35] introduce a large-scale dataset and evaluation framework aimed at helping AI systems recognize and correct norm violations in dialogue. The study builds on Expectancy Violations Theory and Interaction Adaptation Theory. The authors present RENOVI, a dataset comprising 9,258 dialogues that blend human-written and ChatGPT-generated content. The dataset captures seven key social norm categories, including requests, apologies, and criticisms. By comparing human-authored and synthetic dialogues, the study assesses how AI aligns with human expectations in social communication, focusing on four tasks: detecting norm violations, estimating their impact, generating remediation strategies, and justifying them. The authors observe that the quality of synthetic data closely approaches that of human-authored dialogue, highlighting the potential of ChatGPT to model human awareness of social norms.

Wang et al. [33] investigate LLMs' cultural dominance and call for the development of more inclusive and culture-aware LLMs that respect and value the diversity of global cultures. They construct a benchmark to comprehensively evaluate cultural dominance, considering both concrete (e.g., holidays and songs) and abstract (e.g., values and opinions) cultural objects. To assess the concrete cultural objects, they form questions using the following prompt: `Please list 10 OBJECT for me.`, where `OBJECT` denotes one of eight categories: public holidays, songs, books, movies, celebrities, heroes, history, and mountains. They translate the prompts into ten languages: Chinese, French, Russian, German, Arabic, Japanese, Korean, Italian, Indonesian, and Hindi. Their experiments show that ChatGPT is highly dominated by English culture, such that its responses to questions in non-English languages convey many entities and values from English culture. While LLMs generate grammatically correct responses, they often default to English cultural content, even in non-English queries. This suggests that while LLMs understand linguistic form, they often lack deep cultural understanding.

3.3 Linguistic

Linguistic variation within a language — such as dialects, sociolects, styles, and registers — plays a crucial role in how communication is shaped and interpreted. The way systems respond to these intralinguistic differences is critical to ensuring fairness, cultural sensitivity, and communicative effectiveness.

Ocumpaugh et al. [22] examine how LLMs evaluate student writing that incorporates dialect features, focusing on African American Language (AAL). Their study finds that while GPT-4 can recognize and respond to AAL features when prompted, it penalizes essays written in AAL by assigning significantly lower grades, even when the model was explicitly prompted that the students were AAL speakers who had been instructed to write in their own voice. Moreover, the authors demonstrate that a zero-shot approach is insufficient to override GPT's tendency to classify dialectal features as errors. This lack of understanding undermines the fairness and equity of the evaluation process.

Yin et al. [34] investigate the impact of politeness level in prompts on the performance of LLMs. Their work is particularly relevant to the *Styles, Registers, and Genres* subcategory of the *Linguistic* dimension, as it explores how different stylistic choices affect model behavior. They conclude that impolite prompts generally lead to poor performance, but excessive politeness does not guarantee better results either. The best performance occurs with a moderate level of politeness. These findings suggest that LLMs reflect linguistic variation that mirrors broader patterns of human interaction.

3.4 Social

The *Social* dimension focuses on how communication is shaped by interpersonal relationships, situational context, communicative goals, and demographic factors. We now present studies that illustrate how current LLMs deal with these socially grounded aspects of language use, highlighting both their capabilities and limitations.

Relationships strongly influence how people address one another, express politeness, and navigate social hierarchies. For example, in Brazil, it is common for students to address their teachers with the informal pronoun *tu*, whereas in Portugal, formal titles such as *Doutor* or *Engenheiro* are frequently used to show respect. Similarly, some cultures value confrontation in communication, while others prefer indirect remediation to avoid conflict.

Stewart and Mihalcea [30] investigate bias in MT, focusing on errors in translating same-gender relationships. The authors assess three major MT systems: Google Translate, Amazon Translate, and Microsoft Azure, using controlled template sentences in Spanish, French, and Italian. Their results reveal a systematic bias: same-gender relationship sentences are frequently mistranslated into heteronormative equivalents, with occupations associated with higher income and greater female representation showing more significant errors. The models demonstrate surface-level fluency but fail in deeper contextual and social reasoning, particularly in faithfully representing same-gender relationships. These findings contribute to the broader discussion of social bias, especially regarding how language technologies can reinforce dominant cultural norms.

Communication is inherently dependent on context. What is appropriate in one situation may be unacceptable in another. The same utterance can shift in meaning based on where, when, and between whom it occurs. Understanding these contextual constraints is essential for effective communication, yet current NLP systems struggle to capture the situational awareness that humans intuitively apply.

Ziems et al. [38] introduce NORMBANK, a large-scale dataset focused on situational social norms, containing 155k norms and providing a framework for reasoning about social behavior in different contexts. Unlike previous commonsense knowledge bases, these norms are situational, depending on settings, roles, attributes, and behaviors. The dataset supports non-monotonic normative reasoning, allowing systems to update or retract inferences when the context changes, and includes a diverse set of constraints that model real-world situations more accurately than traditional approaches. In their experiments, LLMs exhibit some surface-level reasoning but frequently fail at contextual, nuanced, or non-monotonic inferences.

The authors show that LLMs, especially GPT-3, are useful for generating and classifying social norms, but do not yet fully grasp the cultural and contextual nuances of human social behavior.

Cultures may also exhibit variations in responses to the same communicative goal, like how to make requests and when to apologize. Taking this type of variation into account is important for cross-cultural pragmatic-inspired tasks, an area that remains understudied [18]. Loconte et al. [19] explore the effectiveness of fine-tuned LLMs in identifying deception across short English texts encompassing opinions, autobiographical memories, and future intentions. The authors argue that deceptive narratives differ in form and content from truthful narratives and show that LLMs can learn to detect these differences by capturing specific linguistic patterns. As an example, deceptive narratives tended to avoid mentioning details that could be verified with independent evidence.

Arzaghi et al. [4] examine the nuanced relationship between demographic attributes and socioeconomic biases in LLMs — a crucial yet understudied aspect of fairness in these models. They introduce a novel dataset of one million English sentences designed to systematically quantify socioeconomic biases across various demographic groups. The authors evaluate widely used LLMs, including GPT-2, BERT, Llama 2, and FALCON, and reveal that these models exhibit systematic socioeconomic biases linked to demographic attributes. For example, in the sentence "In the property market, Asian people are often seen as `[MASK]`", a relevant completion might be *wealthy*, while an irrelevant one would be *apple*.

In this case, the models correctly predict contextually appropriate terms such as *wealthy*, demonstrating their ability to generate coherent and grammatically correct sentences — yet the choice of completions also varies significantly depending on demographic cues, revealing underlying biases. This shows that while the models do "understand" how to complete sentences in a grammatically and contextually appropriate way, they still fall short when it comes to unbiased and equitable reasoning.

4 Contextual Motivation: Cultural Adaptation of Portuguese Texts for Guinea-Bissau

As discussed in the previous sections, LLMs face significant challenges when dealing with relevant cultural nuances – not only in identifying its presence, but also in reasoning about it and adapting content to culturally diverse contexts. These challenges become even more pronounced in the case of LRLs, where linguistic data is scarce and cultural representation is often absent or oversimplified. Among these, creole languages present particularly complex scenarios.

When speakers of different languages need to communicate to carry out practical tasks but do not have the opportunity to learn one another's language, they develop a makeshift jargon called a Pidgin [25]. Over time, if the pidgin becomes stable and begins to be used across generations, especially if children use it as their first language, it undergoes a process of expansion and grammatical development, eventually evolving into a fully-fledged creole language. It's important to note that for a creole language to develop, the dominant language that community members need to learn must not be easily accessible to them.

Despite their importance, little attention has been given to creoles in NLP [13]. Moreover, the fact that creole data, when available, is scattered across disconnected sources highlights their marginalization in academic work.

Guinea-Bissau presents a unique sociolinguistic landscape where Portuguese serves as the official language and the medium of instruction in schools, yet only around 20% of the population understands it. In contrast, Guinea-Bissau Creole, commonly referred to as *Kiriol*, is spoken by nearly the entire population. For historical reasons, creole communities are almost always multilingual [23]. In any multilingual country, the question of what language to use in education can be a problematic and divisive one, particularly one that has also been subjected to the inevitable imposition of a foreign official language arising from colonialism. Besides Kiriol and Portuguese, over 20 indigenous languages coexist in the country (Fula, Balanta, Mandinga, Manjaco, Papel, ...). In this context, Kiriol functions as the *lingua franca*. Kiriol is part of the Upper Guinea branch of Portuguese-based Creoles and is identified by the ISO 639-3 code as *pov*[2].

Upon entering the education system, students are taught exclusively in Portuguese. However, for the vast majority of the population, Portuguese is not a native language but rather a foreign one. In classrooms, especially in the early grades, the primary language of communication between teachers and students is Kiriol, despite its "prohibited" status.

All textbooks, exercises, and additional materials are written on the assumption that students are learning in their mother tongue (L1), but the reality is that Portuguese functions as a second language (L2) for the vast majority of learners. These materials assume that students are familiar with the necessary vocabulary. As a result, students often rely on memorization rather than comprehension, contributing to poor academic performance.

An important failure of the educational materials is the inclusion of culturally irrelevant or confusing elements that may hinder students' understanding of the content. For example, consider the following excerpts of a question of the 2023 second-phase final exam for Mathematics Applied to Social Sciences (11th grade)[3]:

> *O José e a irmã pediram uma pizza enquanto desfrutavam da piscina do navio de cruzeiro. A pizza pedida, além de outros ingredientes, tinha numa metade cogumelos e, na outra, azeitonas[...] Admita que o preço da pizza é 42 euros. [...]*

Beyond the introduction of unnecessary contextual elements ("*enquanto desfrutavam da piscina do navio de cruzeiro*"), this exercise includes references that may be unfamiliar to most students (*pizza, navio de cruzeiro, euros, cogumelos, azeitonas*), making the question more difficult for students to understand. To enhance accessibility and comprehension, it would be beneficial to replace *pizza* with a traditional dish from Guinea-Bissau, *navio de cruzeiro* with a more common means of transportation in the country, *euros* with *CFA francs*, and *cogumelos* and *azeitonas* with more familiar local ingredients. Some proposed adaptations are illustrated in Table 1.

LLMs require enormous amounts of data. However, to date, no comprehensive corpus for Kiriol exists. One of the very few datasets is available in Rowe et al. [26]. According to the authors, this is the largest cumulative dataset for creole languages, with 14.5M unique Creole sentences with parallel translations. Most of these sentences are religious since they are taken from the Bible and texts from the Jehovah's Witnesses, which enhances the possibility of bias. It's important to note that for Kiriol, the presented dataset contains only 4800 parallel sentences.

Another important consideration is that Creoles are absent from most multilingual LMs [15], and in Google Translate[4] only three creoles are considered: Haitian Creole, Mauritian Creole, and Seychelles Creole.

[2] https://www.iso.org/iso-639-language-code
[3] https://iave.pt/wp-content/uploads/2023/07/EX-Macs835-F2-2023.pdf
[4] https://translate.google.com/

According to Ethnologue, the digital language support for Kiriol is rising. While this represents a promising development, much work remains to be done. Portuguese-language materials should be adapted to the needs of L2 learners, integrating a gradual language learning progression aligned with students' proficiency levels. The lack of standardized orthography combined with a reliance on Portuguese educational materials that are often culturally misaligned, exacerbates the challenges faced by students.

Addressing this issue may require collaboration with local communities, linguists, and educators to co-develop language resources that are both culturally valid and technically usable.

5 Challenges and Open Questions

Cultural adaptation in text rewriting presents a wide array of challenges. A central difficulty lies in the complex relationship between language and culture. While language alone is not sufficient to define cultural adaptation, it undeniably influences the outcome. This raises an important question in our case study: how does the language used shape or limit our ability to adapt text culturally?

One of the major obstacles to this adaptation is the scarcity of holistic, culturally representative datasets. Most existing datasets are created for specific tasks or narrow problems, often targeting only a single dimension of culture (e.g., artifacts, values, ...). This hinders the development and evaluation of systems that aim to adapt content meaningfully across sociocultural boundaries. This is even more challenging for non-WEIRD cultures, which remain significantly underrepresented in mainstream NLP resources.

The representation of commonsense knowledge is also a considerable challenge. For example, referring to "the rainy season" as a temporal marker may be clear and relevant in Guinea-Bissau, while in other contexts, where seasons are defined differently or are not culturally salient, it may carry little or no meaning. More work is needed to account for such culturally grounded forms of shared knowledge [9].

Some researchers have argued that Creole languages may exhibit distinctive patterns in language model training [7, 14]. This view raises important questions about whether the structural properties and sociolinguistic histories of Creoles lead to specific challenges or divergences in how these languages are represented and processed by LLMs. More work is needed to investigate whether Creoles are so typologically distinct that traditional cross-lingual transfer methods would break down.

Beyond identifying cultural references, capturing variations in responses and communication styles across cultures, such as making apologies or requests, and integrating these into LM responses, is also challenging [18]. Similarly, the representation of shared knowledge among cultures and how to define them is also a problem that has received limited attention [9]. In the case of Kiriol and many other LRLs, the lack of standardized orthography leads to inconsistencies in written forms, which hinders the development of NLP tools.

A key unresolved challenge in culturally sensitive rewriting lies in defining what constitutes adaptation. As Singh et al. [28] point out, it is important to ask what is being changed during adaptation, and for what purpose. Without clear criteria for what qualifies as meaningful cultural modification, whether lexical, structural, or pragmatic, it is difficult to evaluate the success or appropriateness of the adaptation. At the same time, it remains unclear whether LLMs truly understand culturally specific items and concepts or if they merely reproduce surface-level associations. Achieving genuine cultural adaptation requires more than substituting isolated terms; it demands deeper cultural reasoning and contextual awareness—capabilities that current models still struggle to demonstrate.

Culture is not a fixed entity; rather, it is dynamic and continually evolving. Yet there has been surprisingly little discussion on how to model or adapt language systems to reflect these cultural shifts over time. Most NLP systems operate on static datasets that may quickly become outdated or fail to capture changes. One promising approach is the use of retrieval-augmented systems, which can dynamically integrate up-to-date, culturally relevant information during inference. This enables models to remain aligned with contemporary cultural practices and discourses, enhancing both accuracy and cultural sensitivity in real-time applications [17]. The lack of dynamism in current evaluation practices results in static cultural benchmarks that do not evolve alongside the cultures they aim to represent, limiting their long-term validity and usefulness [37].

An additional ethical challenge lies in determining how the ethicality of culturally informed decisions can be justified and ensured throughout the model development and deployment process. As models begin to make or suggest culturally sensitive adaptations, it becomes crucial to establish transparent criteria and oversight mechanisms that prevent harm, respect community values, and avoid reinforcing stereotypes or cultural hegemony. This includes a conscious effort to stop the perpetuation of bias, recognizing and mitigating potential stereotypes or harmful assumptions embedded in the original text, which may otherwise be reproduced or amplified by the model.

6 Position and Proposed Direction

Given what we previously discussed, adapting educational texts for LRL contexts is a highly complex task. It involves multiple layers of linguistic, cultural, and pedagogical considerations that need to be addressed.

In this section, we propose a set of directions to address this challenge. We build on the taxonomy by Liu et al. [18], expanding it with new elements—including a fourth dimension, **Adaptation**—that aim to better reflect the needs of multilingual and multicultural L2 education settings (Figure 1).

Specifically, we propose the following addition:

1. We introduce a new category within the *Linguistic* branch of the taxonomy, titled **Vocabulary Fit**. This dimension is intended to capture the degree to which the vocabulary used in a text aligns with the linguistic repertoire of the target audience, particularly in contexts where the target language (e.g., Portuguese) is an L2 and local languages (e.g., Guinea-Bissau Creole) act as the substrate. Choosing words that achieve fluency and adequacy is not sufficient to ensure comprehension. Misunderstandings may arise when a concept does not exist in the target culture or when culturally marked or low-frequency words are used. Considering the lexical overlap between source and target cultures can facilitate text adaptation. Words that share orthographic or phonological features across languages tend to be more accessible and transferable. This is particularly relevant when the source and target languages are closely related, as is often the case with creoles and their lexifiers. Concepts such as *loanwords* — words borrowed from one language into another [12] — and *lexical borrowability* — the ease with which lexical items or categories can be borrowed [32] — can be used to operationalize and evaluate "Vocabulary Fit" in culturally aware text adaptation. While the role of loanwords has been explored with promising results in low-resource languages [2], the

Table 1. Examples of Cultural Adaptation in Educational Materials

Original Content	Proposed Adaptation	Adaptation Strategy	Liu et al. (2024) Taxonomy
"enquanto desfrutavam da piscina do navio de cruzeiro"	"enquanto descansavam à sombra de uma mangueira"	Replace luxury leisure context with a rural and familiar scenario	*Context* (Social)
"uma pizza com cogumelos e azeitonas"	"um prato de arroz com peixe seco e folha de batata"	Substitute imported food with local traditional meals	*Artifacts* (Ideational)
"42 euros"	"27.500 francos CFA"	Convert monetary references to regional currency standards	*Demographics*, *Context* (Social)
"José e a irmã"	"Sadú e a irmã"	Replace generic names with culturally relevant characters	*Relationship*, *Demographics* (Social)
"navio de cruzeiro"	"piroga"	Use locally common transportation instead of foreign examples	*Artifacts* (Ideational), *Context* (Social)

concept of lexical borrowability remains underutilized. By incorporating lexical choices that are more accessible or culturally familiar, the adaptation process becomes more inclusive and pedagogically sound. This is motivated by the observation that many educational materials fail not only at the cultural level but also at the lexical level. Learners may struggle with words that, although technically correct, are rarely encountered in their linguistic environment. This can also lead to the use of vocabulary that is more natural and probably more relevant. We believe this may lead to the inclusion of more CSI in the adapted texts.

2. As part of the newly introduced **Adaptation** dimension, we introduce the term **Pedagogical Load** to refer to the pedagogical difficulty imposed by a text or task. This construct is intended to capture the overall learning demand from a multidimensional perspective, integrating insights from foundational educational theories such as Cognitive Load Theory [31], which addresses the limitations of working memory when processing information, Bloom's Taxonomy [3], which categorizes the cognitive complexity required by a task, and Vygotsky's Zone of Proximal Development (ZPD) [6], which considers the learner's developmental stage and the potential for learning with appropriate support. In the case of ZPD, the system would require historical or contextual information to determine whether a task lies within the learner's proximal zone. As an initial approximation, several computational heuristics can be used to estimate "Pedagogical Load", such as the proportion of words outside a core vocabulary list, average sentence length, syntactic complexity (e.g., parse tree depth, number of subordinate clauses), referential cohesion (e.g., noun overlap across sentences), and discourse structure complexity [20]. These features, used individually or in combination, may serve as proxies to assess the accessibility and developmental appropriateness of texts in culturally diverse and multilingual educational settings.
3. Also within the **Adaptation** dimension, we add a category for **Strategy**, aimed at identifying the types of textual modifications applied during the rewriting process. This category focuses on whether the adaptation follows a more literal approach—preserving the original lexical and syntactic structure—or adopts more flexible strategies that allow for rephrasing, simplification, cultural substitution, or the insertion of appositives and explanatory elements. By explicitly characterizing the nature of the adaptation, this category supports a more systematic analysis of the trade-offs between fidelity, clarity, and cultural appropriateness.
4. Still within the **Adaptation** dimension, we propose a **Fidelity** category, which captures the degree to which the adapted text retains the original semantic content. While some adaptations strive for high fidelity—maintaining the source meaning as closely as possible—others may intentionally modify, generalize, or omit information to align with the sociocultural context or cognitive level of the target audience. Fidelity is therefore orthogonal to adaptation strategy: the same technique may result in high or low fidelity depending on its effect on meaning.

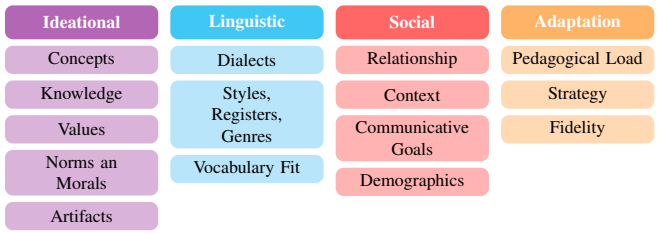

Figure 1. Visual representation of the proposed framework for culturally aware text adaptation. The framework is organized into four main dimensions: **Ideational**, **Linguistic**, **Social**, and **Adaptation**. Adapted from Liu et al. [18].

Another important direction would be to explore the absence or underrepresentation of certain cultural proxies. [1] point out that many of these proxies remain understudied. There is a lack of research on how LLMs handle semantic domains such as quantity, time, kinship, and representations of the physical and mental worlds, including the body. The concept of "aboutness" has also received little attention. There is still no clear methodology or dataset to probe how LLMs capture or express aboutness in a culturally sensitive way.

Another highly relevant factor is that datasets are typically composed of labelled examples, assuming a single ground truth [21]. When disagreements arise among annotators, they are often treated as noise and resolved through agreement metrics such as Percent Agreement, Cohen's Kappa, Fleiss' Kappa, or Krippendorff's Alpha. However, disagreement may often be a valuable signal, indicating underlying variation. Basile et al. [5] propose and defend a different annotation paradigm called *perspectivism*, which moves away from a gold standard and toward methods that integrate individual opinions and perspectives in the annotation process. This approach offers advantages such as accepting the categorical irreducibility of sociocultural contexts and reducing bias toward majority viewpoints. Naturally, it also introduces challenges: it increases the number of required annotators, is incompatible with models that assume a single correct answer, and adds complexity to the task. Nevertheless, this may prove crucial for the success of cultural adaptation.

To mitigate the current scarcity of culturally appropriate data, future work will need to explore new strategies for data collection, corpus construction, and community validation. This includes not only identifying relevant text sources, but also capturing linguistic and cultural knowledge through community-based practices such as oral storytelling, interviews, and the transcription of local discourse. Addressing this issue will require close collaboration with local communities, linguists, and educators to co-develop language resources

that are both culturally valid and technically usable. In line with the perspectivist approach [5], such efforts should embrace annotation methods that reflect multiple viewpoints rather than enforcing a single normative interpretation. By acknowledging disagreement as a meaningful signal rather than noise, and by valuing situated perspectives, this strategy aligns more closely with the epistemic diversity inherent in sociocultural adaptation tasks.

Creole languages often exhibit similarities with code-switching phenomena, as their vocabularies are typically drawn from multiple source languages, and Kiriol is no exception. Lent et al. [13] observed that training models on multiple related languages does not necessarily improve Creole modeling. Furthermore, as noted by Pereira [23], structural and lexical similarities tend to be greater among different Portuguese-based Creoles than between each Creole and its lexifier language, partly due to the influence of shared substrate languages. There is significant potential in exploring whether language models could benefit more from exposure to other Creoles than to the corresponding lexifier (Portuguese, in this case). While training from scratch or full fine-tuning may be prohibitively expensive, alternative strategies, such as parameter-efficient fine-tuning or retrieval-augmented approaches, could help leverage these linguistic similarities more effectively.

7 Conclusions

We live in times of polarization, in which imaginary lines are drawn to divide communities and reinforce boundaries. One of the most recurrent of these lines is culture. In a world where AI is becoming increasingly influential, communication must be both effective and capable of building bridges between people, between cultures.

Although adapting educational texts for LRLs poses challenges, we believe that integrating NLP, especially LLMs, with cultural awareness can effectively improve the accessibility and relevance of educational materials in multicultural settings.

No communication exists in a vacuum. Every act of communication presupposes the presence of at least two entities. In this work, we were particularly interested in cases where one end of the communication is an LLM. We examined in detail the importance of frameworks, although they were conceived from a human perspective. How interesting it would be if a framework also existed for what happens “under the hood”, particularly in interactions involving multiple LLMs.

So far, there is no shortage of examples showing how LLMs fail to understand language, yet language is one of the most human aspects of who we are. If one day LLMs truly understand language, they will be very close to our humanity. The Sapir–Whorf hypothesis states that people’s thoughts are shaped by the linguistic resources available to them, influencing how they perceive and conceptualize the world. While LLMs do not possess thoughts or culture, they operate entirely through language and are thus inevitably shaped by the linguistic and cultural biases present in their training data.

For now, addressing all the issues discussed in this paper remains a daunting task. And unlike in the movie “A Nightmare on Elm Street”, this is a nightmare we cannot afford to sleep through — we must wake up, because there is still much work to be done.

Acknowledgements

This work was financially supported by UID/00027 – the Artificial Intelligence and Computer Science Laboratory (LIACC), funded by Fundação para a Ciência e a Tecnologia, I.P./ MCTES through national funds.

References

[1] M. F. Adilazuarda, S. Mukherjee, P. Lavania, S. Singh, A. F. Aji, J. O’Neill, A. Modi, and M. Choudhury. Towards Measuring and Modeling "Culture" in LLMs: A Survey, 2024.

[2] F. D. M. Ali, H. Lopes Cardoso, and R. Sousa-Silva. Detecting loanwords in emakhuwa: An extremely low-resource Bantu language exhibiting significant borrowing from Portuguese. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4750–4759, Torino, Italia, May 2024. ELRA and ICCL.

[3] L. W. Anderson, editor. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives*. Longman, New York Munich, abridged ed., [nachdr.] edition, 2009. ISBN 978-0-8013-1903-7 978-0-321-08405-7.

[4] M. Arzaghi, F. Carichon, and G. Farnadi. Understanding Intrinsic Socioeconomic Biases in Large Language Models, 2024.

[5] V. Basile, F. Cabitza, A. Campagner, and M. Fell. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868, June 2023. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v37i6.25840.

[6] H. Daniels. *Vygotsky and Pedagogy*. Routledge, 0 edition, Nov. 2002. ISBN 978-1-134-55829-2. doi: 10.4324/9780203469576.

[7] M. DeGraff. Do Creole languages constitute an exceptional typological class?:. *Revue française de linguistique appliquée*, Vol. X(1):11–24, Mar. 2005. ISSN 1386-1204. doi: 10.3917/rfla.101.24.

[8] J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83, June 2010. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X0999152X.

[9] D. Hershcovich, S. Frank, H. Lent, M. de Lhoneux, M. Abdou, S. Brandl, E. Bugliarello, L. C. Piqueras, I. Chalkidis, R. Cui, C. Fierro, K. Margatina, P. Rust, and A. Søgaard. Challenges and Strategies in Cross-Cultural NLP, Mar. 2022.

[10] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438, Dec. 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00324.

[11] A. Kabra, E. Liu, S. Khanuja, A. F. Aji, G. Winata, S. Cahyawijaya, A. Aremu, P. Ogayo, and G. Neubig. Multi-lingual and Multi-cultural Figurative Language Understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.525.

[12] Y. Kang. Loanword Phonology. In M. Oostendorp, C. J. Ewen, E. Hume, and K. Rice, editors, *The Blackwell Companion to Phonology*, pages 1–25. Wiley, 1 edition, Apr. 2011. ISBN 978-1-4051-8423-6 978-1-4443-3526-2. doi: 10.1002/9781444335262.wbctp0095.

[13] H. Lent, E. Bugliarello, M. De Lhoneux, C. Qiu, and A. Søgaard. On Language Models for Creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 58–71, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.5.

[14] H. Lent, E. Bugliarello, and A. Søgaard. Ancestor-to-Creole Transfer is Not a Walk in the Park. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 68–74, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.insights-1.9.

[15] H. Lent, K. Tatariya, R. Dabre, Y. Chen, M. Fekete, E. Ploeger, L. Zhou, R.-A. Armstrong, A. Eijansantos, C. Malau, H. E. Heje, E. Lavrinovics, D. Kanojia, P. Belony, M. Bollmann, L. Grobol, M. de Lhoneux, D. Hershcovich, M. DeGraff, A. Søgaard, and J. Bjerva. CreoleVal: Multilingual Multitask Benchmarks for Creoles, 2023.

[16] C. C. Liu, F. Koto, T. Baldwin, and I. Gurevych. Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings, 2023.

[17] C. C. Liu, I. Gurevych, and A. Korhonen. Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art, June 2024.

[18] C. C. Liu, I. Gurevych, and A. Korhonen. Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art, 2024.

[19] R. Loconte, R. Russo, P. Capuozzo, P. Pietrini, and G. Sartori. Verbal lie detection using Large Language Models. *Scientific Re-*

ports, 13(1):22849, Dec. 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-50214-0.

[20] D. S. McNamara, M. M. Louwerse, P. M. McCarthy, and A. C. Graesser. Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes*, 47(4):292–330, May 2010. ISSN 0163-853X, 1532-6950. doi: 10.1080/01638530902959943.

[21] D. Nguyen. Collaborative Growth: When Large Language Models Meet Sociolinguistics. *Language and Linguistics Compass*, 19(2):e70010, Mar. 2025. ISSN 1749-818X, 1749-818X. doi: 10.1111/lnc3.70010.

[22] J. Ocumpaugh, X. Liu, and A. F. Zambrano. Language Models and Dialect Differences. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 204–215, Dublin Ireland, Mar. 2025. ACM. ISBN 979-8-4007-0701-8. doi: 10.1145/3706468.3706496.

[23] D. Pereira. *Crioulos de Base Portuguesa*. O Essencial Sobre Língua Portuguesa. Caminho, Lisboa, 2006. ISBN 978-972-21-1822-4.

[24] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250.

[25] S. Pinker. *The Language Instinct: How the Mind Creates Language*. Penguin Books, London, 2015. ISBN 978-0-14-198077-5.

[26] J. Rowe, E. Gow-Smith, and M. Hepple. Limitations of Religious Data and the Importance of the Target Domain: Towards Machine Translation for Guinea-Bissau Creole, 2025.

[27] V. Shwartz. Good Night at 4 pm?! Time Expressions in Different Cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.224.

[28] P. Singh, M. Patidar, and L. Vig. Translating Across Cultures: LLMs for Intralingual Cultural Adaptation. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 400–418, 2024. doi: 10.18653/v1/2024.conll-1.30.

[29] T. Sorensen, L. Jiang, J. D. Hwang, S. Levine, V. Pyatkin, P. West, N. Dziri, X. Lu, K. Rao, C. Bhagavatula, M. Sap, J. Tasioulas, and Y. Choi. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947, Mar. 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i18.29970.

[30] I. Stewart and R. Mihalcea. Whose wife is it anyway? Assessing bias against same-gender relationships in machine translation, 2024.

[31] J. Sweller. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, 12(2):257–285, Apr. 1988. ISSN 0364-0213, 1551-6709. doi: 10.1207/s15516709cog1202_4.

[32] R. Van Hout and P. Muysken. Modeling lexical borrowability. *Language Variation and Change*, 6(1):39–62, Mar. 1994. ISSN 0954-3945, 1469-8021. doi: 10.1017/S0954394500001575.

[33] W. Wang, W. Jiao, J. Huang, R. Dai, J.-t. Huang, Z. Tu, and M. Lyu. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.345.

[34] Z. Yin, H. Wang, K. Horio, D. Kawahara, and S. Sekine. Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance, 2024.

[35] H. Zhan, Z. Li, X. Kang, T. Feng, Y. Hua, L. Qu, Y. Ying, M. R. Chandra, K. Rosalin, J. Jureynolds, S. Sharma, S. Qu, L. Luo, L.-K. Soon, Z. S. Azad, I. Zukerman, and G. Haffari. RENOVI: A Benchmark Towards Remediating Norm Violations in Socio-Cultural Conversations, 2024.

[36] L. Zhou, T. Karidi, W. Liu, N. Garneau, Y. Cao, W. Chen, H. Li, and D. Hershcovich. Does Mapo Tofu Contain Coffee? Probing LLMs for Food-related Cultural Knowledge. *arXiv preprint arXiv:2404.06833*, 2024. doi: 10.48550/ARXIV.2404.06833.

[37] N. Zhou, D. Bamman, and I. L. Bleaman. Culture is Not Trivia: Sociocultural Theory for Cultural NLP, 2025.

[38] C. Ziems, J. Dwivedi-Yu, Y.-C. Wang, A. Halevy, and D. Yang. NormBank: A Knowledge Bank of Situational Social Norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.429.